Rôles et enjeux des flux dans les problématiques d'information

Christian Cote

UNIVERSITÉ LYON III
JEAN MOULIN

Rôles et enjeux des flux dans les problématiques d'information

Mémoire pour l'habilitation à diriger des recherches.

Présenté par Christian Cote

07/11/2014

elico
Équipe de recherche de Lyon en sciences de l'information et de la communication

Jury :
Mme Sylvie Lainé Cruzel, Professeur, Université Jean Moulin-Lyon3 (garant)
Mme Maria Caterina Manès Gallo, Professeur, Université Bordeaux 3 (rapporteur)
M. Khaldoun Zreik, Professeur, Université Paris 8 (rapporteur)
M. Bruno Bachimont, Professeur, UTCompiègne (rapporteur)
M. Laurent Romary, Directeur de Recherches, INRIA (examinateur)
M. Aldo Gangemi, Professeur, Université Paris Nord & CNR (examinateur)

Non pas l'œuvre tendue, sourde, monotone autant que la mer qu'on sculpte sans fin – mais des éclats, accordés à l'effervescence de la terre - et qui ouvrent au cœur, par-dessus le souci et les affres, une stridence de plages – toujours démis, toujours repris, et hors d'achèvement - non des œuvres mais la matière elle-même dans quoi l'ouvrage chemine –tous, liés à quelque projet qui bientôt les rejeta- premiers cris, rumeurs naïves, formes lassées- témoins, incommodes pourtant, de ce projet –qui, de se rencontrer imparfaits se trouvent solidaires parfaitement – et peuvent ici convaincre de s'arrêter à l'incertain – cela qui tremble, vacille et sans cesse devient – comme une terre qu'on ravage – épars.

Edouard Glissant, *le sel noir*, 1960

A partir de là vont simultanément se dérouler plusieurs événements qui, en dépit ou peut-être en raison de leur apparente incohérence, constituent un tout pratiquement homogène et cohérent.

Claude Simon, *le Jardin des Plantes*, 1997 (Pléiade, p.1048)

A ma femme, mes enfants, mes parents, ma sœur et son compagnon, mes neveux et nièces, mes beaux-parents…

Mes premiers remerciements vont bien évidemment à Sylvie Lainé-Cruzel, qui suit ce travail depuis de longues années,

Que tous les membres du jury soient également remerciés d'avoir accepté de lire et discuter ce travail,

Mes remerciements vont ensuite à l'ensemble des membres du département Information et Communication, avec qui je travaille depuis de longues années, et principalement Sylvie Lainé-Cruzel, Mabrouka El-Hachani, Catherine Dessinges, Alain Van Cuyck, Jean Pierre Esquenazi, Marida Di Crosta, Martine Vila-Raimondi, Gérald Lachaud, Lucien Perticoz…

Mes remerciements vont également à l'équipe « METEORIST », en premier lieu Richard Dapoigny, Caroline Wintergerst, Guilaine Talens, Patrice Bellot, Laurent Romary…

Je tiens également à remercier Pascal Maire et l'équipe réunie autour de lui à la pharmacie de l'hôpital Antoine Charrial pour leur ouverture d'esprit scientifique, leur curiosité…

N'oublions pas les amis de l'ENSSIB Omar Larouk et Mohamed Hassoun, ni ceux de l'IUT Valérie Larroche, Olivier Dupont…

Et surtout pas l'équipe administrative hors pair qui gère le département (Claire Morel qui a pris la suite d'Esther Hacquin, rejointe par Abdesalem Laïd) et la recherche (Sophie Rey-Mekhloufi).

TABLE DES MATIERES.

TABLE DES FIGURES.

INTRODUCTION

Dans ce travail, nous formulons un certain nombre de propositions relatives à l'amélioration des flux d'information au sein des infrastructures organisant les relations entre les bibliothèques numériques. Cette question des flux, liée aux possibilités à la fois techniques et sociales de structuration des relations entre les collections des différentes bibliothèques, ne trouve pas une réponse seulement en matière d'ingénierie. Nous voulons montrer en quoi une démarche pluridisciplinaire amène à proposer des concepts et un modèle formulés de façon satisfaisante par rapport aux usages.

Notre travail se présente à la fois comme un travail théorique à propos de la définition de l'information, méthodologique dans la démarche de prise en compte de phénomènes existants et observables, pour la formulation de propositions modélisées, et enfin ouvrant la voie à des améliorations concernant l'accès à l'information. In fine, de par la mise en évidence de nouvelles relations entre les documents, nos propositions visent à augmenter la précision des réponses dans le cadre de la recherche d'information par une amélioration de description des documents. Pour cela, nous présentons en fin de travail un projet que nous menons actuellement et qui constitue une première application, parmi d'autres possibles, du modèle présenté.

Très globalement, et cet aspect sera encore accentué dans la partie 6 où l'on positionnera le projet, les possibilités offertes par les standards du web permettent l'émergence d'outils et de services extrêmement divers, interopérables et libres d'accès. L'émergence et le développement d'outils, d'infrastructures ou de collections tout simplement apparait incontrôlé, multiple et contradictoire. Ainsi, l'innovation échappe au contrôle des institutions, qu'elles soient politiques ou professionnelles. Par ailleurs, il n'existe pas de logique ou d'acteur : on voit autant fleurir des projets de domaine (géographie, biologie, etc.) que des projets généralistes. Cette absence de direction apparente est paradoxale, eu égard au fait que l'ensemble des réalisations utilise globalement les mêmes standards fondamentaux de représentation et d'échange : les recommandations du W3C.

La légèreté des standards, la disponibilité des outils et leur gratuité rendent possibles des réalisations rapides et peu onéreuses. La structuration du web, notamment dans la dimension qui nous concerne (RDF et les Linked Data), est une force extrêmement attractive, et qui oblige les institutions (bibliothèques, universités) à réécrire leurs services et leurs offres, mais également à penser un nouveau statut et un nouveau rôle. La publication des offres sur le web constitue un impératif de survie ou de développement pour les acteurs concernés. Le raisonnement serait d'abord de saisie d'une opportunité.

Mais en même temps, cette évolution est nettement plus importante que cela : les limites techniques à la circulation des informations sont très largement dépassées par les promesses des grilles (ou « grid », qui désignent des technologies permettant d'accroître considérablement les quantités l'information numérique pouvant circuler de façon simultanée sur les réseaux). Dans un autre cadre, l'évolution des terminaux, notamment la connexion entre ordinateur et téléphone, entraîne un accroissement des outils, des applications, et des usages qu'il est difficile de mesurer. Par ailleurs, les capacités de traitement des moteurs, surtout si on utilise les capacités de SPARQL, permettent d'envisager des interrogations de descriptions longues et diverses : la pluralité des métadonnées peut être prise en charge par SPARQL.

Le modèle des Linked Data constitue un mode de partage, de fructification et d'échange de services qui devient de plus en plus prégnant dans le cadre de l'évolution des outils et des fonctionnalités du web. Quel que soit le point de vue que l'on choisit d'adopter, on s'aperçoit

que l'accroissement est sensible sur l'ensemble des dimensions de la technologie que l'on considère. Les LINKED DATA constituent une plate-forme sur laquelle il est possible de déposer toutes sortes de données (selon les règles prescrites) et de les relier à d'autres. Néanmoins, les Linked Data ne proposent pas de structuration de ces données et outils. Cette limite est d'importance pour l'exploitation des réseaux de ressources.
Tous ces arguments ne sont pas suffisants pour rendre compte de cette extension illimitée des outils et des services. On fera l'hypothèse que les standards du web, qui constituent les fondements de ces outils et services, constituent des langages et en ont donc les mêmes propriétés. Celle de l'infinitude des réalisations en est une. Cette hypothèse est validée entre autre par les propos de Berners-Lee concernant la dimension prédicative de RDF. Ces capacités expressives sont liées à la fois aux règles formelles des inférences prédicatives et à la sémantique propre des unités symboliques (capacité de référer à des unités du monde). Une autre qualité fondatrice de RDF (et de ses corrélats comme OWL et RDFS) est sa capacité à représenter des inférences, et donc, à partir de certaines données, d'en déduire d'autres.
Comme RDF est structuré comme un langage, qu'il en possède les mêmes capacités d'expression, on peut donc considérer que les langages du web sont promis à la fabrication d'un nombre de texte infini, chacun d'eux étant lié à l'ensemble des autres grâce à l'interopérabilité. La particularité de ce langage est qu'il relie des descriptions, des outils et des données : c'est donc un langage ayant aussi une dimension pragmatique.

Il serait intéressant de développer la dimension anthropologique et culturelle de cette hypothèse. En effet, nous préférons pour notre part explorer deux aspects essentiels : la capacité inférentielle et la distribution/hétérogénéité des objets liés sur le web. Le travail proposé est encore accentué par l'intuition que nous avons que les descriptions documentaires, qui n'entrent pas dans le contenu des textes, peuvent être corrélées à des procédures d'extraction.
Nous ne chercherons pas à fournir une explication approfondie de cette situation. Sur un processus aussi multiforme et évolutif, il est en effet relativement difficile de poser une analyse définitive. Nous nous contenterons de la décrire et de prendre position par le biais d'un projet.
Par contre, cette hypothèse de RDF comme langage permet de caractériser le fait que le niveau de l'information est le plus approprié pour caractériser les phénomènes d'échange de données, de renseignement de données par d'autres.
Dans ce cadre, la caractérisation de l'information constitue une part essentielle de la mise en place du cadre de travail.

Formuler la sémantique du web de données.

Comme nous le verrons, les formats du web offrent des possibilités d'interopérabilité généralisée. Néanmoins, cette interopérabilité n'est généralement considérée que du point de vue syntaxique. Or, l'absence d'une prise en compte de la dimension sémantique entraîne des aberrations, des confusions, et donc une perte de fluidité et de pertinence des outils du web.

Nous ferons donc un certain nombre de propositions d'outils dont l'objectif consiste à structurer les dimensions sémantiques du web, entendant cette question de sémantique comme à l'interface des questions linguistiques, logiques et enfin d'études cognitives.

En effet, on ne peut complètement considérer les objets du web comme des objets linguistiques, ni non plus comme les objets purement formels de la logique, ni encore de simples questions informatiques.

En prenant comme point de départ le document, et en considérant que ce qui circule est d'abord constitué de descriptions de documents, nous nous considérons la discipline la mieux à même de répondre à des questions relevant de la description de documents.

Nous partons du principe que les outils que l'on propose au web de données ne doivent pas seulement être formellement satisfaisants, mais être fondés sur des pratiques humaines attestées. En effet, les outils que l'on propose sont destinés à être utilisés par le plus grand nombre, et par conséquent, en dehors des questions d'ergonomie et de convivialité, il est nécessaire que les opérations réalisées par ces outils puissent être considérées comme des modèles de processus que l'on met en œuvre de façon quotidienne dans la vie de tous les jours.

Globalement, notre point de vue est celui de l'interprétation de l'information. Cette position est issue du cadre de la sémantique formelle, par laquelle on cherche à comprendre comment les expressions peuvent être interprétées, considérant la plus grande généralité possible. Ainsi, on ne s'intéresse pas à tel usager ou catégorie d'utilisateur, mais bien à la façon dont n'importe quel utilisateur peut être amené à interpréter l'information qu'il perçoit.
Au vu des disparités disciplinaires (entre les SHS et la médecine par exemple), géographiques (en vertu de politiques nationales différentes) comme également des spécificités de travail collaboratif des différents laboratoires, il apparaît malaisé de prendre le point de vue des utilisateurs comme êtres sociaux.
La définition de l'interprétation comme activité générique permet d'aborder l'ensemble des domaines de recherche que l'on vient de présenter. Cette position a d'abord un intérêt analytique, parce qu'elle permet d'observer et de relier des phénomènes hétérogènes comme notamment les expressions linguistiques, le cours de l'activité dans lequel ces expressions sont interprétées et enfin le raisonnement mis en œuvre dans le cadre de cette activité.
Ce point de vue permet de développer une analyse mais son principal intérêt réside dans le fait qu'il permet de fédérer différentes perspectives pour l'élaboration de projets de recherche pluridisciplinaires. La seule dimension de l'étude qui ne soit pas couverte par l'interprétation concerne le cadre des bibliothèques numériques et la structuration de leur offre.

Problématique de la description documentaire.

La description documentaire définit l'ensemble des outils, procédures et langages utilisés pour publier des renseignements normés à propos d'objets documentaires, et plus particulièrement de publications. La description documentaire comprend la description bibliographique et l'étend à des contextes autres que celui du contrôle bibliographique. On entend par là notamment le contexte des bibliothèques numériques, qui nous concerne plus particulièrement.
On considère généralement la description documentaire en lien à la recherche d'information dans la mesure où la description sert à répondre aux questions posées lors de requêtes utilisant des systèmes d'interrogation ou des moteurs de recherche. Or aborder ces deux questions de façon simultanée pose un problème de faisabilité du fait de l'ampleur du domaine. On aurait pu considérer que les problèmes de recherche d'information déterminent la description de l'information. Néanmoins, la complexité même des questions de recherche, y compris du seul point de vue des outils (sans parler des usages ou des usagers), constitue en soi un thème de

recherche à part entière. En effet, doit-on s'intéresser à l'évolution des langages d'interrogation, des moteurs, des usages avant d'aborder la description documentaire ? Non, parce que la description documentaire sert aussi à fixer l'identité de chaque document et à spécifier sa place dans l'ensemble de la documentation disponible. Elle a donc un rôle structurant pour l'organisation des collections.
On en restera donc à une schématisation restreinte : la description documentaire sert à constituer la représentation la plus précise et la plus proche des contenus des documents de façon à ce qu'elle offre à un utilisateur la satisfaction de sa requête.

Fondements prédicatifs de la description documentaire. Choix d'une perspective sémantique.

On peut caractériser la description documentaire comme le fait de représenter par un vocabulaire contraint des aspects ou des traits d'un document donné.
Fondamentalement, il s'agit d'un acte de prédication particulier. Il est particulier parce qu'il requiert des outils spécifiques, comme notamment les métadonnées. Il l'est également parce que les vocabulaires utilisés sont contraints, et enfin parce que cette prédication a comme usage la recherche d'information.

Nous proposerons un certain nombre d'outils permettant d'analyser cette forme de prédication particulière et d'accompagner son évolution par des propositions intégrées dans les projets du web de données. Notre objectif n'est pas tant de produire un discours sur la description documentaire, mais de proposer une analyse et des méthodes permettant d'appréhender la production de descriptions documentaires, en considérant les questions de signification. Si les dimensions techniques sont abordées, c'est parce que la sémantique des outils de description documentaires reposent sur une syntaxe partagée (garantie de l'interopérabilité), XML. L'interopérabilité syntaxique n'est pas une garantie de la cohérence sémantique des produits, et justement, nous situons notre apport dans la cadre de l'élaboration d'outils sémantiques pour la description documentaire.

Ainsi, nous nous appuyons sur des théories et des modèles qui ont en commun de traiter la question de la prédication dans des termes différents. Qu'il s'agisse de propositions logiques, linguistiques ou d'anthropologie cognitive, elles ont toutes en commun la capacité à traiter de la question de la prédication en considérant des contraintes, notamment informationnelles.
Nous développerons donc plusieurs propositions théoriques en les mettant en perspective dans le cadre de la description documentaire.

Notre objectif consiste à montrer en quoi les flux d'information permettent de spécifier les relations crées dans le cadre du développement du web de données. Les flux caractérisent des contraintes interprétatives sur des relations entre ressources hétérogènes. Cet apport se fait à l'aide de structures de données reliées par une inférence. Notre travail se double d'un intérêt pour la structure d'information, que l'on considère comme une forme de réalisation des flux.

Comme nous le verrons, les langages du web et les cadres de travail proposés par le W3C constituent des outils que l'on peut mettre en valeur de façon originale. Le premier objectif consiste donc à spécifier par les flux des relations entre ressources hétérogènes, comme par exemple la description d'un document par un autre auquel il est lié.
Comme il est couramment admis, le web ne peut entrer dans le contenu des documents. Or, des travaux se poursuivent, à propos notamment de la structure d'information. A travers l'annotation et l'extraction d'information, il existe aujourd'hui tout un travail centré sur le

traitement du document en vue de la reconnaissance et de l'analyse de son contenu par une machine. Notre second objectif consiste à caractériser à l'intérieur des documents des relations entre données hétérogènes de façon à élaborer un modèle de la structure d'information fondé sur des flux d'information.
Nous montrerons qu'il est possible de lier des questions de relations entre données et l'exploration de ces structures pour l'élaboration d'un modèle complet d'exploration et de mise en relation de documents.

Notre travail, et en premier lieu les flux, concernent les représentations de l'information à différents niveaux d'abstraction. S'il s'agit d'abord de propositions logiques et sémantiques, dans le cadre du web de données, on voit apparaître des outils qui permettraient de relier des contenus à différents niveaux d'abstraction, comme par exemple SKOS. Généralement il s'agit d'un niveau lexical et d'un niveau conceptuel. Les flux étant caractérisés par deux niveaux différents d'abstraction, ils peuvent fournir des réponses à ces questions.

Conformément à la nécessité d'un fondement humain (social, psychologique) de nos propositions, nous proposons de fonder notre modèle sur des pratiques existantes. La pharmacie constitue un exemple d'usage des flux d'information, « usage » au sens des méthodes mises en œuvre par une communauté pour résoudre certains problèmes d'information liés à son activité. La pharmacie constitue un objet d'observation, mais pas le domaine d'application de notre modèle. Le domaine d'application est celui des bibliothèques numériques et des infrastructures du web dans lesquelles ces dernières sont mises en relation.
Plus globalement, les flux d'information constituent un usage de l'information, « usage » au sens de fonctionnalités agrégées à un concept, fonctionnalités permettant au concept d'apparaître pertinent dans certaines situations d'ingénierie ou d'analyse. Ainsi, l'information constitue un concept générique subsumant les flux. (Nous définirons donc l'information puis nous caractériserons les flux dans ce cadre).

Les questions de flux, et plus généralement d'information, impliquent des données symboliques. Par conséquent, on ne pourra passer outre les dimensions linguistiques et cognitives. Et ce d'autant plus que ce qui nous intéresse en termes d'information, ce sont les contenus, et par conséquent la sémantique interprétative. La sémantique interprétative s'intéresse à l'ensemble des raisonnements par lesquels une expression donnée signifie pour un sujet récepteur dans un certain contexte.

Il n'existe ni théorie, ni modèle, qui permette de rendre compte de façon satisfaisante des rapports entre les dimensions cognitives, linguistiques et sociales des flux. Par conséquent, on utilisera des approches du langage et de l'activité compatibles avec les flux.
On pourra même considérer que ces approches anticipent les flux parce qu'elles sont fondées sur une définition plus ou moins explicite de l'information : l'information n'est pas une problématique centrale de la linguistique et des sciences cognitives. L'information est explicitement définie dans le cas de la théorie des situations qui constitue une théorie de la signification fondée sur les principes d'information ; elle est nettement moins explicitement caractérisée dans le cadre de la cognition située et distribuée, alors même que les phénomènes décrits sont explicitement des flux d'information.

Plan général.

Dans la partie 1, nous présenterons le cadre des bibliothèques numériques et leur évolution. Cette problématique est intégrée à l'intérieur du développement des ressources numériques et

de leur structuration. Nous présenterons ainsi un cadre factuel. Nous n'entrerons pas dans les détails des outils et des méthodes utilisées mais nous présenterons un panorama d'une évolution dans laquelle l'ensemble de notre étude s'inscrit et trouve sa pertinence.

Les parties 2 et 3, concerneront le cadre définitionnel dans lequel on se situe, à travers deux concepts essentiels, l'usage et l'information. Dans la partie 2, il s'agira essentiellement de caractériser un cadre méthodologique qui permettra de caractériser globalement l'approche choisie pour ce travail. La partie 3 concernera plus particulièrement la définition de l'information la plus pertinente pour notre travail.
Les parties 4, et 5 constituent le centre de la démonstration concernant les flux et les structures d'information. Les structures d'information étant dans notre cas déduites des flux, l'ordre des parties est cohérent.

Dans la partie 6, nous entrerons très précisément dans l'application de l'ensemble de nos propos à l'intérieur d'un projet inscrit dans le cadre du développement des bibliothèques numériques. Nous montrerons l'opportunité de notre projet par rapport aux problématiques actuelles des bibliothèques numériques et de leur contexte. Ainsi, nous développerons une présentation des lexiques sémantiques, des ontologies et des métadonnées qui nous préoccupent, et donc par rapport auxquels nous situons notre projet.

GLOSSAIRE.

Nous présentons quelques définitions de termes utilisés fréquemment afin d'éviter des ambigüités ultérieures. Ces définitions ne concernent que les termes qui ne constituent pas (mis-à-part la notion d'usage) le cadre de notre problématique.

DOCUMENT : tout objet circonscrit du monde pouvant entrer dans une collection et donc pouvant être soumis à une approche scientifique.
USAGE : Dans notre travail, nous nous référons à trois définitions différentes de l'usage :

- La description de ce à quoi sert ou peut servir un dispositif, un outil.
- La façon dont une communauté s'approprie et utilise un objet, un outil ou encore un concept.
- La mise en œuvre d'entités dans des contextes fonctionnels définis. Cette définition est plus centrée sur le langage.

Nous développerons dans les parties opportunes chacune de ces définitions. Elles doivent néanmoins être posées maintenant pour éviter toute confusion.
PORTAIL : dans un site internet, le portail est la partie visible pour l'utilisateur et qui lui permet d'accéder aux ressources. Le portail présente à l'utilisateur une offre de services.
SERVEUR : un serveur web est une ordinateur relié à internet et qui contient des pages web, disponibles pour un utilisateur à distance. Il peut s'agir d'un groupe d'ordinateurs ayant la même adresse http. Il contient un interpréteur dynamique (PHP) et un gestionnaire de base de données (MySqL).
INFRASTRUCTURE : une infrastructure caractérise un ensemble de relations établies entre des serveurs de façon à pouvoir faire circuler des informations entre eux. Une infrastructure repose sur la mise en relation de l'offre de plusieurs bases de données.
PLATEFORME : une plateforme constitue un espace dans lequel sous certaines conditions chacun peut déposer des données. Il s'agit donc d'une base de données et d'un serveur. La plateforme peut intégrer des liens à des services et des bases extérieurs mais ne met pas en relation les différentes bases par, par exemple, un moteur de recherches commun. Une plateforme est donc une structure plus légère qu'une infrastructure mais qui ne permet pas

d'offrir des services aussi complets qu'une infrastructure.
GRILLES : Les grilles ou « grid » constituent des technologies de transfert de quantités d'information fondées sur une dynamique.
LINKED DATA : les Linked Data ou données liées constituent une préconisation du W3C destinée à mettre en réseau les outils élaborés selon certains principes d'interopérabilité, de représentation (notamment l'usage des préconisations RDF) et d'ouverture des données. Ces outils (qui peuvent être des ontologies, des représentations sémantiques lexicales, des thésaurus ou des jeux de métadonnées sont ainsi disponibles et combinables selon les besoins de l'utilisateur)
JEU DE METADONNEES : un jeu de métadonnées désigne un ensemble cohérent de descripteurs susceptible de cerner un objet sur l'ensemble de ses dimensions en lien à une certaine finalité. Un jeu se caractérise soit par une DTD soit par un schéma.
WEB DE DONNEES : le web de données (ou web sémantique) caractérise l'ensemble des langages, des outils et objets du web (dont des documents) inscrits dans un processus de description et de mise en relation permettant à des « machines d'interpréter des données ». « Le web de données (Linked Data en anglais) combine les technologies du web sémantique avec les principes fondamentaux du web (protocole HTTP, identifiants URI), avec pour objectif la construction d'un réseau d'informations structurées, disponibles en ligne et facilement réutilisables dans de nombreux contextes[1]. » Il faut être relativement précis sur certaines distinctions à l'intérieur même du web de données : le web de données caractérise à la fois les outils qui vont traiter et mettre en relation des données et ceux qui vont décrire des documents. L'appellation web de données ne permet pas de rendre compte de cette distinction, le document étant une donnée. L'appellation web sémantique permettait de la faire. Le web de données proprement dit caractérise l'ensemble des opérations de balisage et de relations entre balises qui se réalisent à l'aide de données. Par exemple, un balisage TEI d'un poème numérisé n'aura pas du tout la même signification que la description de ce poème par des métadonnées Dublin Core[2]. Nous nous intéresserons essentiellement à la description de documents, et plus particulièrement à la mise en relation de descriptions.

[11] http://www.bnf.fr/fr/professionnels/web_semantique_donnees/s.web_semantique_intro.html
[2] http://dublincore.org/

PARTIE 1. Cadre de travail : bibliothèques numériques et services d'information.

Nous présentons ici le cadre dans lequel notre travail trouve sa pertinence. Il s'agit d'un contexte social, économique et politique : l'évolution des bibliothèques numériques et la mise en place d'infrastructures permettant de fédérer ces ressources. Notre objectif est de contribuer au développement des services au sein des infrastructures reliant les bibliothèques numériques en intégrant les questions d'e-science et les avancées en termes de description, de traitement et de représentation que cette transformation de l'activité scientifique porte. Or, les bibliothèques numériques ne constituent qu'une partie des problématiques de l'e-science, qui concernent globalement l'activité scientifique. Les moyens mis en œuvre pour le développement de l'e-sciences, et qui reprennent les questions d'ontologies et de terminologie, ne peuvent que transformer fortement l'univers en cours de structuration des bibliothèques numériques. Des membres du réseau ISIDORA aux responsables du Dublin Core, les activités de veille, de recherche et de proposition de solutions concernant l'e-science sont prioritaires. Ainsi, si l'on envisage les bibliothèques numériques, c'est avant tout pour en proposer un état de fait. Le développement de notre propos montrera comment ces bibliothèques numériques sont intégrées à l'intérieur du cadre de l'e-science.

L'objectif de cette partie consiste d'abord à élaborer le cadre dans lequel nos travaux prennent place. Nous nous limiterons ici à décrire le cadre de l'évolution des bibliothèques numériques, sans développer des éléments contextuels fondamentaux comme la stratégie du W3C et l'évolution des outils d'organisation des connaissances. Les dimensions techniques et les propriétés des outils et techniques utilisés seront développées dans la partie 6. Néanmoins, nous introduirons l'impact de ces évolutions sur les domaines scientifiques directement liées à la représentation de l'information : lexicologie, ontologies.

Notre objectif est ici de présenter un tableau de l'évolution des bibliothèques numériques et des outils qu'elles utilisent. Il s'agit donc de caractériser l'environnement dans lequel ces bibliothèques se situent ; nous dépassons donc le cadre initial pour nous intéresser aux principales dynamiques technologiques du web, notamment les langages de représentation et leurs usages.
Par ailleurs, les initiatives dont nous rendons compte se situent d'abord dans le domaine des Sciences Humaines et Sociales. Nous ne traiterons pas du domaine biologique et médical ainsi que les spécificités de la géographie : de par l'antériorité et l'ampleur des recherches menées dans ces domaines, celles-ci peuvent être considérées comme pionnières et connaissent un développement propre.
Cette partie du travail est contrée sur la description du domaine des bibliothèques numériques. Nous l'abordons en considérant que la technique et les dimensions politiques jouent un rôle prépondérant pour expliquer l'explosion actuelle. Les paramètres sociaux et économiques, mais également juridiques, sont importants, mais ils ne nous semblent pas aujourd'hui structurant du développement et de l'amorce d'organisation du domaine. Si l'on choisit ces deux paramètres, c'est parce que les langages de représentation ont des propriétés logiques et expressives très proches des langues naturelles et un maniement aisé, rapide et peu coûteux. La dimension politique, ou plutôt institutionnelle, est également essentielle parce que derrière les services et outils se trouvent des enjeux fondamentaux en matière d'organisation et de performance de la recherche, des enjeux d'information et d'enseignement.

Nous aimerions montrer deux tendances essentielles :

- d'une part aux questions d'interopérabilité, de numérisation et de représentation des documents s'ajoutent des questions de modalités d'échange et d'aide effective au travail scientifique en ligne,
- d'autre part la montée en puissance des infrastructures permettant de lier ces bibliothèques numériques et d'offrir des services nouveaux, qui bouleversent les modalités de travail des usagers. Cette tendance s'inscrit dans un autre projet, l'e-science, qui se caractérise comme la montée en puissance à la fois de la transparence scientifique, du travail collaboratif et du partage des données.

Ce dernier cadre est celui dans lequel nos propositions s'inscrivent. Si ces deux tendances sont liées, c'est parce qu'elles constituent deux dimensions (l'une liée aux usages, l'autre aux politiques publiques) de la structuration du web. Cette question de la structuration du web est d'abord institutionnelle, mais elle se traduit aussi par des questions relatives aux contenus et à leur mise en valeur. La perspective qui se dessine est celle d'un web dynamique dont nous essayons de formuler les contours.

Nous ne considérons pas les bibliothèques numériques et les infrastructures d'un point de vue statique, mais en mettant en évidence les logiques qui fondent ces réalisations. En ce sens, on ne reviendra guère sur les questions techniques. On peut tout de même remarquer qu'il existe depuis le début de ces initiatives un certain consensus entre les acteurs concernant l'évolution des bibliothèques sur le web, qui s'intègrent de plus en plus dans le cadre de services et amorcent la mise en place d'outils pour le travail (plus précisément scientifique) sur le web. L'enjeu dépasse alors très largement les questions des bibliothèques et s'inscrit dans celui de la politique de la recherche et de l'innovation.
Les premières questions qui ont fondé les pratiques de création et de développement des bibliothèques numériques ont été celles de la numérisation, du renseignement des données (via les métadonnées) et de l'interopérabilité des formats. Elles sont aujourd'hui liées à la mise en relation et à l'exploitation de l'ensemble de ces données, de façon à élaborer des outils. C'est d'ailleurs maintenant que la question des usages se pose, tout simplement parce que l'on invente de nouveaux outils.

Conformément à notre propos de départ, les bibliothèques numériques et plus généralement les offres documentaires, se développent sans une logique (politique, économique) intrinsèque apparente. Ce qui fait la cohésion de l'ensemble est le langage utilisé, à savoir les standards du web et leurs possibilités expressives. L'impression d'absence de cadre unique de développement, entre autre au niveau national, est due en grande partie à la multiplicité des acteurs engagés, à des statuts et des tutelles différents. Leur seul point commun est la visibilité de leur offre sur le web et celle-ci est vue comme un corrélat de leur développement récent.

1.1. La montée en puissance des bibliothèques numériques.

Nous définissons en premier lieu les bibliothèques numériques, en les distinguant entre autre des archives ouvertes. On peut reprendre la définition usuelle de Lorcan Dempsey [3] reprise entre autre par Lionel Maurel[4]: "Any collection of digital resources managed with the primary goal of maximizing the collection's utility to a defined user community".
Cette définition a comme intérêt d'associer des documents numériques, des services et un ou des publics cibles. Ensuite, y compris en observant les bibliothèques numériques d'information scientifique et technique, on s'aperçoit aisément de la diversité des objectifs,

[3] http://orweblog.oclc.org/archives/000349.html
[4] http://www.bnf.fr/documents/definition_bibnum.pdf

des acteurs et des outils.
Par contre, cette définition ne prend en compte la dimension technique des bibliothèques numériques. En effet, les bibliothèques numériques supposent une base de données, un moteur de recherches et un portail permettant d'accéder à l'ensemble de cette information. Néanmoins, ce n'est pas cette structure technique qui permet à nos yeux de définir de façon précise la bibliothèque numérique, parce qu'une telle structure peut être partagée par des objets totalement différents comme les sites commerciaux notamment ou les entrepôts de données. Nous préférons donc utiliser des critères d'usage pour définir les bibliothèques numériques.
Ce qui spécifie la bibliothèque numérique, ce sont trois critères :

- La mise à disponibilité de la connaissance. (accessibilité et structuration de la connaissance (notamment par une catégorisation ou une classification) modularité permettant d'actualisation des collections). Cette mise à disponibilité requiert une identification précise des objets (s'ils ne sont pas stockés dans la bibliothèque, ils doivent être accessibles par des liens) et une actualisation (notamment si les objets ont des versions) fréquente.
- L'adaptation aux besoins de l'utilisateur (à savoir être dotée d'un moteur de recherche structuré par au minimum une liste de mots-clés et de possibilités de navigation). Elle doit être dotée d'une capacité d'édition, à savoir d'intégration d'objets par l'utilisateur, de façon directe par une contribution, ou indirecte l'actualisation des collections.
- L'adoption de standards permettant l'intégration de la connaissance dans le cadre d'outils de recherche promus par les infrastructures. Ces standards permettent l'interopérabilité, donc facilitent l'accessibilité de l'information. Ces standards concernent à la fois les descriptions documentaires, l'indexation des objets (notamment pour les moteurs de recherche généralistes) et les outils de recherche intégrés.

On organise la partie qui suit en montrant à la fois la diversité des projets, de leurs initiateurs et du contenu et, dans un second temps, en montrant en quoi les infrastructures du web constituent des outils permettant de fédérer cette offre multiple.
Cette montée en puissance des infrastructures s'inscrit dans le cadre de l'augmentation de la puissance et de l'offre des services web. Par exemple, si le TGE-ADONIS accueille le moteur de recherche ISIDORE, il accueille également des espaces de stockage de données et des espaces de travail pour les chercheurs. Ainsi, les bibliothèques numériques s'inscrivent dans une offre plus globale d'outils et de services qui amène une transformation radicale des pratiques scientifiques.
Nous aimerions, dans cette contextualisation de notre travail, présenter les grandes lignes de cette évolution, sans avoir la prétention de l'exhaustivité. Nous nous limiterons aux projets en IST, et plus précisément en SHS.

Les objets dont on va parler peuvent à priori sembler très divers, sans véritable cohérence interne. Or, on ne pourra manquer d'observer une constante : tous ces projets sont issus et soutenus par le CNRS, pour la France, et directement par le ministère de l'Education et de la Recherche en Allemagne http://www.d-grid-gmbh.de/index.php?id=1&L=1[5]. En Grande Bretagne, les financements sont également publics [6] et associent les « départements » d'enseignement et de recherche et ceux centrés sur l'innovation et l'industrie (Department for BusinessInnovation & Skills). La particularité de la Grande Bretagne consiste à associer les questions d'enseignement avec celles de la valorisation de la recherche. Cette mise en relation

[5] http://www.bnf.fr/documents/definition_bibnum.pdf
[6] http://www.jisc.ac.uk/aboutus/howjiscworks/finance.aspx

permet de réduire les coûts. Les bibliothèques sont absentes de ces consortiums. Ainsi, les bibliothèques numériques et les infrastructures d'e-science émanent fondamentalement des institutions de la recherche, ce qui leur permet de maîtriser le développement de la diffusion et de la valorisation de leurs propres travaux. En France, la particularité est la présence du CNRS. La conséquence de cette disposition est le fait que ces consortiums vont utiliser et développer des outils et des méthodes propres, distinctes de celles des bibliothèques. Ainsi, les personnels chargés de l'information au CNRS sont des documentalistes et non des bibliothécaires.

1.1.1. Des projets ciblés.

En France, les projets qui voient le jour autour de 2005. On pourrait citer d'abord GALLICA, qui constitue le projet initiateur de la BNF. On doit tout de suite distinguer les bibliothèques numériques à vocation d'abord patrimoniale des bibliothèques d'information scientifique et technique. Par ailleurs, dans ces dernières, nous distinguons d'une part les bibliothèques proprement dites proposant des documents en accès libres et d'autre part les sites des éditeurs. Le recensement des bibliothèques numériques publiques est proposé par le Ministère de la Culture (et concerne la France)[7].

On distinguera également les bibliothèques numériques des portails de ressources, généralement disciplinaires, qui se mettent en place. Ils n'ont pas vocation à mettre en ligne des publications, mais à fournir des services (agendas, ressources, outils, etc.). Le CLEO propose une offre mixte, à la fois de revues numériques et de services.

Nous nous arrêterons plus précisément sur trois projets en information Scientifique et Technique, mais dont les lignes éditoriales sont très distinctes et complémentaires. Ces trois projets sont emblématiques des bibliothèques numériques entre autre de par leurs différences d'objectif et de contribution à la disponibilité de l'information scientifique sur le web : il s'agit de PERSEE, du CLEO et de HAL.
Par ailleurs, GERICCO a présenté une étude exhaustive des bibliothèques numériques[1]. Nous reprendrons leur historique (p. 15) de façon à clarifier le cadre des trois initiatives dont nous parlons :
« Avec un peu de recul, on pourrait diviser cette courte histoire des archives ouvertes en quatre périodes :
1. 1995-2000 : Les précurseurs. Création des premiers sites à caractère institutionnel ou scientifique comme CITHER, CNUM-CNAM et LACITO Archive.
2. 2001-2004 : Le développement. Une première prise de conscience politique avec une accélération du mouvement, autour d'archives à caractère national comme HAL, TEL, NUMDAM, Cyberthèses, EduTICE et ArchiveSIC.
3. 2005-2007 : L'essor. Mise en place de plus en plus d'archives ouvertes, surtout institutionnelles. Parmi ces sites figurent Archimer (IFREMER), PERSEE (SHS), LARA (rapports) et plusieurs des archives institutionnelles de HAL (INSERM, IRD, Pasteur).
4. 2008-... : La consolidation. Après une forte croissance, on constate un certain ralentissement de la mise en place de nouveaux sites qui n'est pas nécessairement définitif. On repère plusieurs projets de taille, objectif, architecture et contenu assez différents : SPIRE (Sciences Po Paris), OATAO (Toulouse) ou encore DUMAS (mémoires) ».

Nous insisterons ici un peu plus sur des projets relativement autonomes par rapport aux

[7] http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f_02.htm

bibliothèques universitaires et aux bibliothèques municipales. A la différence des projets de bibliothèque, qui reposent sur la mise en valeur de collections existantes, les bibliothèques numériques dont nous parlons construisent progressivement leurs collections, entre autre par le recueil et la numérisation de documents disponibles par ailleurs.
Ce sont aussi des projets de très grande ampleur qui ont pour origine des appels d'offres publics [8], ou l'évolution propre des institutions du CNRS (cas de HAL) ou encore une unité mixte de services (UMS) associant le CNRS, et des universités de Marseille et Avignon pour le CLEO. Ce sont les acteurs de la recherche (Ministère de l'Enseignement Supérieur, CNRS, Universités) qui mettent sur pied des outils répondant à des besoins propres : répertoire de publications, actualités et archives de la recherche. Le consortium ERUDIT, au Canada, possède les mêmes caractéristiques fondamentales[9]. On peut trouver dans la présentation d'ERUDIT un résumé de l'ancrage de ces bibliothèques numériques dans les préoccupations des organismes de recherche plus que des bibliothèques : « Érudit a pour mission de promouvoir et de diffuser les résultats de la recherche et de la création ».
Via les portails, l'objectif de communication guide la création et le développement de ces bibliothèques numériques.

PERSEE est un projet fondé sur la duplication des données : des revues papier sont désormais dupliquées sous forme de fac-similé. L'objectif est à la fois une numérisation de qualité permettant de disposer sur le web de documents identiques à la version papier et de rendre disponibles des productions scientifiques relativement difficiles d'accès dans les bibliothèques universitaires (et nécessitant entre autre le prêt interbibliothèques). C'est donc un outil très orienté vers le public, y compris à l'international : le respect de la pagination et donc la possibilité de citer constitue un exemple de cette forme de passage au numérique « de substitution ». En dehors de la numérisation, PERSEE s'est focalisé sur les questions de structuration des documents, au travers notamment de l'adoption de le TEI. L'adoption de ce standard permet de rendre le document numérique utilisable comme un document papier. En ce sens, PERSEE ne constitue pas un outil novateur en termes d'usage : il vise simplement à rendre le document numérique d'une utilisation aussi facile que le document papier.

Le CLEO se présente comme un outil permettant l'édition de publications scientifiques, et notamment de revues en ligne. Par le biais de son portail revues.org ([10]fondé en 1999), le CLEO participe à la mise en valeur de la production scientifique dans le domaine des SHS. Ce principe communautaire a permis le développement d'autres outils comme calenda.org en 2000 (pour diffuser les annonces de manifestations scientifiques) et hypothese.org (en 2008), pour mettre à disposition des travaux en cours. Dans sa formulation initiale, le CLEO est d'abord un outil d'édition. La diffusion numérique permet à des revues jusque-là confidentielles de devenir accessibles dans le monde entier. Comme PERSEE, il utilise la TEI. Son évolution actuelle le rend proche de la mise en place d'un service pour l'émergence d'un soutien à l'activité de recherche. Le CLEO s'inscrit dans le projet de TGIR (Très Grande Infrastructure de Recherche), initié en 2008, et participe au volet BSN (Bibliothèque Scientifique Numérique). Conformément à cet engagement, le CLEO construit actuellement une offre de service sous forme d'abonnement de façon à se doter d'un modèle économique viable. (Cette initiative d'édition ouverte est essentielle parce qu'elle permet à des bibliothèques de disposer à la fois du service CLEO et d'un ensemble d'autres services d'édition (en libre accès)[11]. Cette offre est constituée à la fois de l'accès à l'ensemble des

[8] http://www.persee.fr/web/support/history
[99] http://www.erudit.org/apropos/info.html
[10] http://www.revues.org/
[11] http://cleo.cnrs.fr/936

documents proposés par le CLEO, mais également à des outils de référencement et d'édition. Il s'agit donc d'une offre aux bibliothèques, afin qu'elles disposent d'un outil d'édition, leur permettant de réaliser plus facilement leurs bibliothèques numériques : « OpenEdition *freemium* constitue une opportunité historique pour associer libre accès et développement durable de l'édition. *Cette offre* est tout simplement le moyen concret que propose le Cléo aux bibliothèques et aux éditeurs pour créer une alliance durable en faveur du libre accès en sciences humaines et sociales ».
Avec ses différentes offres, et notamment CALENDA et HYPOTHESES.ORG, le CLEO a créé une offre d'animation de la recherche qui va bien au-delà de la bibliothèque numérique vers le soutien à des réseaux sociaux.

HAL, crée et maintenu par le CCSD, fonctionne différemment des deux premiers. Il s'agit d'un outil permettant le dépôt par les auteurs de leurs publications. Les questions de droit imposent quelquefois que le texte intégral soit absent, ou que l'on présente une version de travail. Il s'agit ainsi d'une mise en commun de publications fondée sur la volonté des auteurs. Ce dépôt est fondé sur le volontariat et le contrôle par les pairs. Il s'agit aussi d'un outil de référence, notamment dans le cadre de l'évaluation de la production des chercheurs. Enfin, il permet de mettre en valeur la production scientifique nationale. Cela dit, HAL ne constitue maintenant que l'un des services du CCSD[12]. Le CCSD comprend maintenant, notamment après la mise en place de HELOISE en mai 2012 [13]une panoplie de services orientés vers le management de la recherche.

L'ensemble des initiatives présentées s'inscrit dans le cadre du libre accès. Comme l'exemple du CLEO le montre, le soubassement économique de ce mouvement est fragile, et en France au moins, réside dans la persistance d'un financement public. C'est la raison pour laquelle l'alliance avec les bibliothèques universitaires, sous forme d'abonnement à l'offre de services, constitue pour le CLEO une nécessité.
La dépendance aux financements publics de l'enseignement et de la recherche constitue un problème pour les projets qui ne sont pas directement, comme PERSEE, des soutiens directs à l'activité de recherche et de publication.

Il serait donc prématuré de considérer que les bibliothèques numériques constituent des outils ayant atteint leur maturité. Selon l'étude de J. Schöpfel & H. Prost (op.cit.), « l'espace d'un an, le nombre des archives ouvertes en France a pratiquement triplé, passant de 56 à 150 sites. » (p.27) En même temps, les rapports de cette étude montrent la réelle difficulté qu'il y a à appréhender les usages de ces ressources. En effet, il n'y a que peu de traces des lecteurs dans ces bibliothèques.

Si l'on observe par rapport à cette étude, on remarquera que les projets portés par les bibliothèques, et notamment les bibliothèques de recherche, sont relativement absents (on pourra tout de même mentionner le SUDOC, qui constitue un référentiel structuré, mais pas vraiment une bibliothèque, puisqu'il ne contient pas de collections. Relativement à la quantité de données traitées par les projets précédents et des enjeux sociaux et professionnels qui les sous-tendent, ils peuvent sembler relativement secondaires. Ils sont en réalité nettement plus spécialisés et à vocation de mise en valeur d'un patrimoine existant. Par conséquent, ils ne peuvent apparaître comme fédérateurs du mouvement social et économique auquel les trois projets que nous avons présenté contribuent fortement. Néanmoins, comme nous le verrons plus loin, des projets fédérateurs émergent au sein des associations de bibliothécaires

[12] http://ccsd.cnrs.fr/realisations.html
[13] http://heloise.ccsd.cnrs.fr/)

scientifiques, ce qui fait que c'est par le biais de la structuration de l'offre aux équipes et au suivi des projets scientifiques que les bibliothécaires scientifiques s'inscrivent dans l'évolution dont nous parlons.

Enfin, cette relative discrétion des bibliothèques ne s'explique pas simplement parce que les initiatives qui ont émergé dans le cadre des bibliothèques numériques répondaient à des problèmes propres au monde de la recherche et de l'enseignement supérieur (référencement des publications, diffusion, accès) qui ne concernaient pas directement les bibliothèques.
Ces dernières ont concentré leur effort sur des questions de numérisation, de mise en valeur et d'accès au patrimoine pour ce qui concerne leur fond, et la montée en charge des métadonnées et des protocoles normalisés pour ce qui concerne les échanges d'information.

En définitive, on aura pu constater que les bibliothèques numériques montrent une diversité de fondation, de choix techniques, d'objectifs et d'usage qui rend nécessaire l'existence d'infrastructures. Celles-ci permettent d'unifier, au travers d'un moteur de recherche, l'offre disparate des bibliothèques numériques.
On s'aperçoit aussi que cette description introductive ne permet pas de proposer des conclusions sur les logiques à l'œuvre. Ce niveau d'observation est trop fin pour que l'on puisse vraiment observer les phénomènes à l'œuvre. C'est au niveau plus international des outils partagés, des infrastructures liant différentes bases de données que l'on pourra vraiment observer les logiques à l'œuvre.

1.1.2. Les archives ouvertes.

L'initiative de l'OAI (Open Archive Initiative) s'inscrit dans le mouvement de préservation d'un web gratuit, utilisant les formats ouverts et fournissant des services qui ne sont pas commerciaux. L'OAI est à lier à une seconde initiative internationale majeure, datant de 2003, qui est la déclaration de Berlin sur le libre accès à la connaissance en sciences et sciences humaines. Elle est construite et basée sur la définition issue de l'Open Access Initiative de Budapest (de 2001). Cette déclaration est fondatrice du mouvement libre accès.
Les produits proposés par l'OAI sont des formats et un outil de moissonnage de métadonnées. PERSEE et le CLEO utilisent les standards de l'OAI.

Le but de l'OAI consiste, au travers de l'agrégation de ressources et l'échange de métadonnées, à rendre visible des collections de documents, quelles que soient les métadonnées utilisées (Dublin Core, IMDI[14], EAD[15], etc.). Le second projet de l'OAI concerne l'agrégation de sites[16]. Il en résulte la mise en commun, le référencement et l'accès simultané à des ressources hétérogènes.
Cette agrégation permet de lier des documents différents mais également de spécifier la nature de ces liens : « décrit par » ou « décrit », « agrège », « est agrégé par » et « est similaire à ». Ces descripteurs sont interopérables avec le DC et FOAF[17], et s'écrivent en utilisant les recommandations de RDFS.

Au-delà de ces aspects techniques, le projet des archives ouvertes s'intègre dans un mouvement collectif, celui des « humanités numériques ». Regroupés autour d'une charte, ces

[14] http://www.mpi.nl/imdi/
[15] http://www.loc.gov/ead/
[16] http://www.openarchives.org/ore/1.0/datamodel.html
[17] http://www.foaf-project.org/

acteurs s'engagent à développer une forme d'accès au savoir spécifique (voir le manifeste[18]).
Si l'on regarde plus précisément, le mouvement des archives ouvertes est fondé sur les programmes du NSF (National Science Foundation, agence américaine chargée du management de la recherche), et est soutenu par le mouvement des universités américaines (la « Coalition for Networked Information »,[19] la fédération américaine des bibliothèques numériques (« Digital Library Federation »,[20]) et la fondation Andrew W. Mellon.
Les domaines d'intervention de ces différents acteurs sont complémentaires : la CNI s'occupe autant d'e-science [21] que la DLF (Digital Libraries Foundation), qui se penche notamment sur l'ensemble des opportunités offertes par les LINKED DATA [22]pour les bibliothèques.
Ainsi, on trouve au fondement des archives ouvertes les principales institutions universitaires américaines et leurs bibliothèques. C'est donc un mouvement institutionnellement très puissant qui est à la base du web libre et gratuit.

Si l'on veut élargir la question, on peut utiliser les ressources des Digital Humanities[23]. Les Digital Humanities constituent un projet et un réseau militant pour que les SHS ne soient pas exclues des développements du web, et surtout, que des réponses spécifiques soient apportées aux problèmes posés par les SHS. Dans leur texte de présentation, les DH s'inscrivent dans un mouvement plus ancien, qui est représenté par la revue "Computers and the Humanities »[24] "" (1966-2004). (Aujourd'hui, la revue "Language Resources and Evaluation[25] " ", qui en est la continuité, est centrée sur la question des ressources linguistiques).

Si l'on suit l'historique des Digital Humanities [26], on s'aperçoit que le projet s'intègre de plus en plus dans celui des cyber-infrastructures, et en premier lieu l'e-science. Les DH constituent ainsi un réseau intellectuel pouvant servir d'interlocuteur aux institutions et entreprises (notamment GOOGLE). Comme on le verra, les acteurs privés s'inscrivent dans le mouvement initié par ces consortiums.
En définitive, ce mouvement comprend plusieurs dimensions, liées à la diversité des acteurs mobilisés : les institutions de la recherche et de l'enseignement, les pouvoirs publics et les usagers. Il est le précurseur de la dynamique actuelle : il mêle les dimensions collaboratives, internationales, de disponibilité des données. Si l'on regarde plus précisément, on identifie des communautés académiques (les réseaux de philosophes, de linguistes, etc.), des institutions (comme par exemple la JISC (Joint Information Systems Committee, en Grande Bretagne)) et des équipes relativement restreintes de chercheurs. L'adoption de standards permet à ce genre de projets peu coûteux d'avoir potentiellement une audience internationale très importante.

En définitive, lorsque l'on observe les initiatives précédentes, on s'aperçoit d'un éparpillement des projets, que la perspective des archives ouverte permet néanmoins de fédérer. En effet, ces projets sont nés à un moment où la perspective d'une mise en relation généralisée de l'offre globale était encore lointaine.

[18] http://tcp.hypotheses.org/318
[19] http://www.cni.org/about-cni/membership/members/)),
[20] http://www.diglib.org/members/
[21] http://www.cni.org/program/current-program-plan/2011-2012/developing-managing-networked-information-content/institutional-disciplinary-implications-of-e-research/
[2222] http://www.diglib.org/archives/3167/
[23] (Site en français : http://dhi.intd.cnam.fr/).
[24] http://www.springerlink.com/content/0010-4817/" Computers and the Humanities
[25] http://www.springerlink.com/content/1574-020x/
[26] http://dhi.intd.cnam.fr/digital-humanities/

Code de champ modifié

1.2. Émergence des infrastructures de services.

Les services web constituent l'une des innovations essentielles qui aient requis le cadre du RDF. En effet, on ne se situe plus dans le cadre de l'accès à l'information mais dans celui d'opérations réalisées directement en ligne. Le commerce en ligne est l'une des premières applications de cette nouvelle caractérisation de l'usage du web. Nous évoquons ici des initiatives qui se développent en dehors des bibliothèques numériques, mais qui vont peu à peu rejoindre les préoccupations de ces dernières, et former la convergence à laquelle nous assistons et dans laquelle nous intégrons notre projet.

Un intérêt fondamental de ces travaux est qu'ils intègrent la dimension collaborative (notamment via ZOTERO), et donc l'émergence de communautés. Néanmoins, l'objectif va bien au-delà du seul développement d'un laboratoire virtuel. Il s'agit de rendre réutilisables les données et les résultats de la science. Les conséquences peuvent être économiques ; elles sont aussi épistémologiques parce que les conclusions d'un travail ne sont pas fondées sur les seules données collectées par un chercheur, mais bien sur l'ensemble de celles qui, dans un champ de recherche, ont été collectées et mises en commun.

Les propositions dont nous venons de parler s'inscrivent dans un projet beaucoup plus vaste, qui vise à donner les moyens d'un travail scientifique en ligne, et qui ne soit pas limité à l'accès à l'information. Ce projet à long terme est en cours de définition, et par conséquent nous l'aborderons en fin de présentation. Nous nous limiterons ici aux services offerts aux communautés scientifiques.
Cette émergence des services au travers des infrastructures constitue à l'heure actuelle un fait essentiel de l'évolution des technologies de l'Information et de la Communication. Le point de départ est un rapport de la NSF : ATKINS Daniel E. : « Revolutionizing Science and Engineering through Cyberinfrastructure : Report of the National Science Foundation » Blue-Ribbon Advisory Panel on Cyberinfrastructure [en ligne] New York, NSF, janvier 2003, 84p. [27]

Si nous présentons le versant institutionnel et politique de ce projet, il faut de suite mentionner qu'il s'appuie très fortement sur le réseau social des « humanités digitales »[28]. Cette dimension est essentielle parce qu'elle permet de montrer l'implication des usagers, en l'occurrence les chercheurs, à l'intérieur du développement des outils. A l'image de la Text Encoding Initiative (TEI), l'un des outils des Humanités Digitales, un travail de diffusion est systématiquement réalisé, ce qui permet d'instaurer un dialogue permanent entre les usagers et les concepteurs.

« Il s'agit donc de créer de vastes répertoires (des« cyber infrastructures » dans la terminologie de l'e-Science) avec les capacités managériales des bibliothèques numériques traditionnelles en y ajoutant une couche supérieure de service, notamment des outils spécifiques d'exploitation de ces données (chercher, déplacer, manipuler, personnaliser …), d'indexation, et de communication synchrone et asynchrone. »[2]p. 2. La bibliothèque ne constitue qu'une part de l'offre ; vue du côté de l'usager final, celle-ci constitue un outil parmi d'autres.

L'importance de cette problématique des services est telle qu'un certain nombre de nouveaux

[27] http://www.nsf.gov/od/oci/reports/atkins.pdf
[28] Pour une présentation en français : "http://dhi.intd.cnam.fr/digital-humanities/" http://dhi.intd.cnam.fr/digital-humanities/.

Code de champ modifié

champs d'expertise apparaissent ou sont reconfigurés par la problématique des services en ligne.
Ainsi en est-il des SOAS (Service-oriented Architectures). A la différence des accès aux services fondés des applications bureau et des ressources à télécharger, les architectures orientées services proposent des applications web standards pouvant être adaptées individuellement. Ce type d'architecture peut être utilisé directement pour concevoir les applications et les services. Par exemple, on peut envisager une annotation interactive de documents.

Ces services en ligne requièrent des infrastructures. En effet, ils ne peuvent se développer de façon rapide que dans le cadre d'un regroupement de l'offre. Ce sera alors le rôle des infrastructures.

On peut définir une infrastructure de recherche de la façon suivante (précisant la définition générale des infrastructures donnée plus haut) : « RIs help to create a research environment in which all researchers - whether working in the context of their home institutions or in national or multinational scientific, initiatives - have shared access to unique or distributed scientific facilities (instruments and services, also including data and their management), regardless of their type and location in the world. RIs foster knowledge and skills development by enabling research. These can, in turn, be disseminated to the research, education and enterprise communities, and thereby contribute to innovation.
RIs include knowledge-based and enabling resources, research facilities, equipment, materials and services. Human resources are strongly needed to develop and maintain them and to ensure their sustainability. These RIs support basic or applied research, and maintain and develop research capacity. RIs may be single-sited, or distributed, or "virtual"; they could also have a national, regional, pan-european or global dimension ».[3].5.

Mis en forme : Anglais (États Unis)

Nous présentons d'abord les infrastructures françaises, puis celles qui se mettent en place au niveau européen. Globalement, ces infrastructures ont comme objectif de fédérer les initiatives que nous avons précédemment évoquées. Le poids institutionnel y est bien plus fort ; elles constituent donc les outils d'application des politiques nationales ou communautaire de développement des infrastructures numériques, notamment en ce qui concerne le travail scientifique.

1.2.1. Survol d'infrastructures nationales.

Les infrastructures n'ont pas comme objectif de construire des bibliothèques, mais en fournissant une offre de services aux chercheurs, la question de l'information devient rapidement centrale. Les questions de recherche d'information s'intègrent donc rapidement à l'offre des infrastructures. Nous commencerons par présenter cet aspect avant de montrer que d'autres enjeux viennent s'inscrire dans cette construction d'infrastructures et en premier lieu les questions d'e-science. Cela s'explique assez facilement : le développement de l'activité scientifique et sa modernisation sont des impératifs fondamentaux des politiques nationales de recherche, bien plus que les seules dimensions de l'information scientifique.
En France, la première réalisation d'envergure permettant de lier les questions d'infrastructure et celles de bibliothèques est le méta-moteur ISIDORE mis au point par le TGE-ADONIS en2011. Il repose sur le principe du moissonnage de métadonnées. ISIDORE ne prend pas en charge que les bibliothèques numériques. D'autres ressources telles que les pages web des principales institutions de l'enseignement et de la recherche sont intégrées à l'intérieur des

1604 sources moissonnées par ISIDORE. « Les données qui sont à l'origine des archives constituées ou des corpus sont dans le périmètre "primaires", les publications scientifiques sont dans le périmètre "secondaires" et les actualités de la recherche dans le périmètre "Données évènementielles". Bien évidemment, ISIDORE vous permet d'interroger l'ensemble des périmètres en même temps, mais il vous est ainsi possible de classer les données »[29] [. Néanmoins, ISIDORE intègre les outils et services du CLEO, qui constituent au moins en partie des services d'e-science. Par ailleurs, une même institution peut proposer plusieurs sources. Ainsi, on parle de collections : l'offre d'une institution constitue alors une collection.
En ce sens, ISIDORE a un rôle de fédération de l'offre en matière de bibliothèques numériques, d'indépendance par rapport aux moteurs de recherches les plus usuels et de normalisation des pratiques. Il faut en effet que les métadonnées soient renseignées de façon appropriée pour que le moteur de recherche puisse les saisir.

Ces infrastructures ont également un autre enjeu : elles visent à améliorer la performance du travail scientifique en offrant des espaces de stockage et des outils de traitement des données. Le but consiste à offrir des moyens de façon à faciliter la coopération entre équipes éloignées géographiquement et de disciplines différentes. Ainsi la grille du TGE-ADONIS, en lien avec les CINES et l'IN2P3, permet le stockage, l'aide au calcul et la pérennité des résultats de recherches. En dehors d'ISIDORE, le TGE-ADONIS est essentiellement une infrastructure de soutien à l'activité de recherche. Une infrastructure se caractérise par sa capacité à relier des ressources différentes, de s'assurer de leur interopérabilité et de la pérennité du travail déposé.

A proprement parler, ISIDORE n'est pas un projet d'e-science, au sens où il peut être développé au Royaume Uni, [30]mais un service agrégé à ceux qui existent et qui permet de mettre en valeur, au travers d'un moteur de recherche dédié, les fonds des bibliothèques numériques.
Le projet même du TGE-ADONIS, relativement centralisé et lié au CNRS, constitue un cas relativement rare (voir les conclusions deK. Eccles & alii.[4]) : " What we can see even from this small sample of six e-Infrastructures is the diversity of these socio-technical systems, both on the social side where we have top-down discipline transcending federations (DRIVER-II, OSG, CineGrid) and national efforts (SND) as well as bottom-up efforts (SBG). [...]What is common to these efforts is that none has a hierarchical and centralized structure, and none is developing technology on behalf of a single group. Instead, they are creating longer term collaborative socio-technical structures. Whether these should be labelled infrastructures is questionable: none yet serve the whole research community and none have yet established a user-base that relies on this system".

La portée des infrastructures est encore du domaine des projections politiques, notamment au travers du programme « e-research 2020 ». Les travaux préliminaires à ce projet[31] font apparaître un réel intérêt des chercheurs pour le montage de ce genre d'outil de travail, considérant aussi les changements de pratique que cela comporte.

En France, une initiative de l'IMAG à Grenoble vise à créer une communauté de travail autour de l'e-science[32]. Nous reprendrons cette question plus loin, entre autre parce qu'elle implique un repositionnement de l'offre et du travail des bibliothécaires. Le TGE-ADONIS distingue clairement la mission d'aide à la recherche, via la grille, de l'accès à l'information

[29] http://www.rechercheisidore.fr/annuaire
[30] http://www.merc.ac.uk/
[31] http://www.eresearch2020.eu/eResearch2020%20Final%20Report.pdf
[32] http://e-science-mi2s.imag.fr/e-science/

électronique, via ISIDORE.

Les autres pays européens, et plus particulièrement la Grande-Bretagne, les pays nordiques et les Pays-Bas, ont lancé de programmes nationaux d'infrastructures, mais qui ne sont pas systématiquement associés à une très large segmentation disciplinaire, comme peut l'être le TGE-ADONIS (qui rappelons-le, couvre l'ensemble des SHS alors que le CINES constitue l'infrastructure pour les autres sciences).

De façon emblématique, les Pays-Bas, à travers l'agence Nederland e-science Center [33]lance un certain nombre de projets disciplinaires mais permettant systématiquement à une discipline, au regard de ses spécificités, de disposer d'une réponse e-science à ses questions de mise en commun. Parmi ceux-ci, le projet Generic e-Science [34]qui porte l'élaboration d'outils génériques pour l'e-science et notamment pour la gestion des données primaires, dont nous reparlerons.

On peut également citer les projets suédois [35]ou britanniques [36]et surtout l'organisme fédérateur JISC[37]) pour illustrer la prise en charge du développement de l'e-science au niveau national. En Grande Bretagne, le développement de l'e-science et des bibliothèques numérique s'intègrent pleinement à l'intérieur du JISC sous forme de treize programmes de recherches, quarante-deux projets et deux services.

Il n'existe pas de modèle unique d'infrastructure, mais différents modèles qui correspondent à des besoins différents (liés à la forme que peut prendre une coopération elle-même). Le rapport préparatoire au projet (eResearch2020_Final_Report, op. cit. p. 11/262) est à cet égard formel : « From this report, a number of patterns can be elicited, including understanding that e-Infrastructures should not be regarded as uniformly top-down efforts but also bottom-up efforts, both of which may emerge within but also across disciplines and fields of research. This heterogeneity, and a balance of leading-edge and more well established efforts, are highlighted at a number of points throughout this document as requiring a balanced approach in terms of support and planning. Further findings from the report include a selection of technical but mainly social bottlenecks to e-Infrastructures development, of which a current critical one is the sharing and re-use of data."
Il existe également des projets transnationaux comme le réseau Knowledge Exchange [38]qui permet d'échanger des solutions entre projets nationaux, et surtout impliquent les bibliothèques[39]. Ces projets bénéficient également du soutien du réseau EUROCris [40] qui est une association dont le but est de fédérer et échanger dans le cadre des projets d'e-science, de bibliothèques numériques et de l'apprentissage en ligne.

En définitive, si une politique volontariste veut être menée, doit se réaliser en prenant en compte la diversité des configurations scientifiques (notamment dans des cadres pluridisciplinaires), des demandes d'outils et de ressources. Dans ce contexte, le lien avec les bibliothèques numériques, et plus globalement l'offre informationnelle, ne peut être simple, à

[33] http://www.esciencecenter.com
[34] http://www.esciencecenter.com/projects/project-portfolio/generic/
[35] http://essenceofescience.se/infrastructures/
[36] http://www.merc.ac.uk/
[37] http://www.jisc.ac.uk/
[38] http://www.knowledge-exchange.info/
[3939] http://www.deff.dk/english/
[40] http://www.eurocris.org/Index.php?page=hometext&t=1

la différence de la proposition TGE-ADONIS/ISIDORE.
En réalité, les dimensions de recherche, de développement et de propositions de services se croisent au sein des différentes initiatives qui sont le fait à la fois de réseaux nationaux et européens. Cette complexité est liée en grande partie aux politiques de recherche, d'enseignement supérieur, de développement technologique et d'information scientifique et technique des différents pays européens. Elle relève aussi de cultures scientifiques et d'enseignement différents.

Pour conclure, notre point de départ, à savoir les bibliothèques numériques, croise très rapidement d'autres chantiers de la mise en ligne des ressources et de leur structuration : l'e-science, l'enseignement.

Cette convergence rapide s'explique en partie par l'interopérabilité des langages utilisés. La généralisation de l'utilisation des langages de représentation entraîne le développement d'outils divers pouvant aisément se connecter entre eux et produire de nouveaux services. La facilité de maniement et la souplesse de ces outils entraîne une productivité très forte et une réponse rapide à des problèmes divers. A contrario, cette facilité entraîne une faible structuration du domaine et par conséquent de nombreuses difficultés à découvrir et utiliser ces ressources.
L'objectif des projets européens et notamment de DARIAH, sera alors de fédérer ces expériences nationales, mais au vu du contexte que l'on vient d'évoquer, les infrastructures européennes auront essentiellement un rôle fédératif. Néanmoins, avant de décrire les projets européens, il convient de rappeler le dynamisme de certaines politiques nationales de développement des services à la recherche et à l'enseignement. Ces politiques de développement sont liées à des politiques de recherche proprement dites et des choix scientifiques. Ainsi, si l'on regarde les politiques nationales, on peut rapidement s'apercevoir d'une grande diversité de projets tant au niveau de leur structuration que des choix scientifiques. La seule conclusion est que ce développement prend la forme d'infrastructures permettant de fédérer des projets de recherche différents mais utilisant des outils semblables.

1.2.2. Développement de projets européens fédératifs.

Les projets européens s'inscrivent dans un cadre fédératif, à la fois pour des raisons politiques et pour d'autres, qui tiennent à la diversité des besoins, des projets et des finalités de recherche. Ils visent à mettre en relation les initiatives nationales et/ou disciplinaires, voire de consortiums pluridisciplinaires comme par exemple celui qui concerne les ressources linguistiques. L'ensemble des projets présentés ici ont émergé à partir du 7ème PCRC. La politique générale projetée à propos des infrastructures est redéfinie maintenant à partir d'une nouvelle "roadmap", GRDI 2020[41] . Nous y reviendrons, notamment à propos des stratégies relatives à la fédération des ressources des bibliothèques scientifiques.

Les bibliothèques numériques sont considérées comme étant suffisamment stabilisées pour pouvoir être mises en relation. Comme nous allons le voir tout de suite, le cadre initial des bibliothèques est largement intégré dans celui des ressources.
Nous présenterons d'abord ce cadre très général, puis les trois initiatives les plus proches de notre projet (CLARIN, DARIAH et EUROPEANA). Enfin, nous présenterons les résultats de l'initiative PEER, qui concerne en premier lieu l'édition scientifique et technique.

[41] http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf

Néanmoins, si l'on veut comprendre comment se structure cette offre, il nous faut distinguer trois problématiques différentes :

- La question de l'ouverture des données (qui constitue une condition politique au développement de l'offre numérique). Cette politique d'ouverture des données (Open Data Pilot) insérée dans le projet « Horizon 2020 » concerne en premier l'activité scientifique et les bibliothèques numériques. Cette politique est rendue possible, au niveau des réseaux techniques, par les grilles, qui permettent la circulation de quantités de plus en plus importantes de données.
- Les choix de formats. Les formats étant universels, élaborés, amendés et recommandés au sein de communautés scientifiques et industrielles, ils ne sont pas concernés directement par les politiques publiques.
- Les infrastructures assurent la mise en relation et donc en cohérence de nombreuses initiatives éparses, achevées ou en cours. Elles sont visibles grâce à des portails, et font le lien entre différentes plateformes. Leur capacité à structurer en font un outil important pour les politiques d'organisation de l'offre de services sur le web.

Nous nous intéresserons relativement peu ici à la première problématique, dans la mesure où cette question se pose de façon spécifique pour les bibliothèques, et que nous traiterons cette question plus loin. La question des formats est certes centrale dans notre travail, mais peu importante pour les politiques publiques. C'est donc la question des infrastructures qui sera centrale pour montrer la façon dont une entité politique affirme son rôle par le biais du déploiement d'outils spécifiques.

1.2.2.1. EUROPEANA : accès à des ressources muséales de plus en plus étendues.

Lorsque l'on consulte le programme de recherche de l'Union Européenne e-contentplus[42].
On constate, en ce qui concerne les contenus numériques, que la priorité est donnée à trois projets : EUROPEANA et la communauté de la créativité artistique, le libre accès (Open Data) à l'information scientifique, et enfin le e-learning. Cette tripartition reprend les la remarque précédente concernant la convergence entre les l'activité scientifique, les bibliothèques et l'enseignement.

La question des bibliothèques numériques est dès lors totalement intégrée dans celle de la circulation des données. Elle constitue un ingrédient d'une politique plus globale, qui concerne les données numériques et la culture, ou l'activité scientifique : « public data for re-use: geographical information, statistics, weather data, data from publicly funded research projects, and digitised books from libraries » (diapo14 de la présentation du projet[43]).
Dans ce cadre (ibid., dia. 20), l'accent est mis sur l'élaboration de nouveaux modèles permettant de caractériser l'ensemble des procédures de mise en valeur de cette information : « Explore new paradigms for accessing, querying, using and evaluating this information (e.g. article level assessment, text mining, research/research paper identifiers, etc.) Impact: Opening wider access to more scientific information, contributing to new ways to asses and re-use, evaluate scientific information ».

[42] http://ec.europa.eu/information_society/activities/econtentplus/currentcall/index_en.htm

[43] http://ec.europa.eu/information_society/activities/ict_psp/documents/call6_theme_2_presentation.pdf

Si l'on regarde de façon générale, en observant la participation de LIBER (Ligue européenne des bibliothèques de recherche) à l'intérieur des projets européens, on peut s'apercevoir que cette implication concerne huit projets. LIBER est impliquée d'abord dans la politique culturelle numérique via les projets EUROPEANA, dans les archives scientifiques pérennes (APARSEN[44] et ODE) et des projets liés à la structuration des infrastructures et plateformes de façon à s'assurer de leur sécurité, accessibilité et pérennité (voir le rapport du groupe AAA (authentication, authorization and accounting) :[45])

Après cette présentation générale du cadre institutionnel, nous nous intéressons plus précisément aux infrastructures développées depuis 2008 et qui organisent les données de certains domaines de recherche et dans le cas d'EUROPEANA, de culture. Nous présentons plus précisément EUROPEANA notamment parce qu'il s'agit d'un ensemble de projets impliquant les bibliothèques. Ces infrastructures s'organisent de façon relativement disciplinaire. Elles sont liées pour certaines à la circulation des données nécessaire à l'activité scientifique elle-même (physique nucléaire par exemple) ou à la publication de données statistiques (en économie, géographie et sociologie). Une part essentielle de ces infrastructures n'intègre pas directement la dimension documentaire. Notre choix d'infrastructure à présenter se limite à celles qui concernent le domaine des bibliothèques numériques et le champ dans lequel nous développons notre projet, à savoir les ressources linguistiques (CLARIN). Les langages documentaires ne font pas partie de CLARIN, mais les descriptions de documents, notamment les métadonnées, y sont intégrées.

EUROPEANA constitue un projet de plus faible ampleur par rapport aux initiatives CLARIN et DARIAH parce qu'il ne concerne que le patrimoine culturel. EUROPEANA constitue la suite du projet MINERVA, qui concerne plus fondamentalement la numérisation du patrimoine européen et l'adoption de normes communes. Il s'agit d'un projet bien plus ciblé que le précédent, ce qui explique aussi qui fait figure de pionnier pour les projets européens.
Comme ISIDORE, EUROPEANA constitue un moteur de recherche. L'objectif consiste à associer les différentes institutions muséales, d'archive et de bibliothèques disposant de ressources d'images numériques patrimoniales de façon à les rendre accessibles dans l'Europe entière. Le problème est au départ le fait chacune de ces institutions dispose de ses propres outils de structuration, de langue différente et quelquefois de culture professionnelle différente. La solution trouvée est avant tout sémantique, à savoir qu'elle vise à construire des relations en utilisant les outils fournis par SKOS. Ces relations constituent un « réseau sémantique » qui sera lié au moteur de chacune des institutions. On utilise alors les relations spécifiées entre deux concepts pour activer l'interrogation de deux organisations de connaissances. L'idée consiste ainsi à permettre une mise en valeur des différents outils d'organisation des connaissances en construisant des relations entre eux.
EUROPEANA a ainsi élaboré un modèle permettant de structurer et de représenter des données provenant d'institutions différentes. EDM (Europeana Data Model) qui constitue le modèle de départ d'EUROPEANA[46] est fondé sur les standards du W3C, ce qui le rend suffisamment générique pour pouvoir être adopté par les bibliothécaires dans leur projet européen (voir infra).
EUROPEANA a connu une première phase de réalisation, et un appel a été lancé par le CORDIS (ICT Policy Support Programme) concernant la version 2. Celle-ci se doit

[44] http://www.alliancepermanentaccess.org/
[45] https://confluence.terena.org/download/attachments/30474266/2012-AAA-Study-report-final.pdf?version=1&modificationDate=1355503760046&api=v2
[46] http://pro.europeana.eu/documents/900548/bb6b51df-ad11-4a78-8d8a-44cc41810f22

d'améliorer la description des documents, les usages et le passage à plus grande échelle[47].
Si l'on inclut également d'autres projets, plus sectoriels, liés à EUROPEANA (EUROPEANA libraries, fashion, 1914-1918, photography, newspaper, REGIA,[48] il apparaît que le modèle de départ se distribue dans des outils plus spécialisés, ce qui permet également de renforcer sa généricité.
EUROPEANA ne constitue pas seulement une infrastructure de recherche, mais un outil à destination du public. L'enjeu pour la construction de l'identité européenne est également important.

1.2.2.2. CLARIN : accès et mise en valeur des données linguistiques.

CLARIN[49] constitue une infrastructure européenne dont le but est de créer un réseau de conservation et d'échanges de données, centré sur les ressources linguistiques. CLARIN héberge des données et des services et assure la relation à d'autres bases de données, qui elles, hébergent des services. Un aperçu de cette organisation est décrit ici : http://www.clarin.eu/node/2971 . Elle vise à faire en sorte que les données soient intégrées et constituent une référence pour structurer un domaine tant du point de vue des ressources que des outils de description.
CLARIN distingue clairement dans son offre les services offerts (dépôt et consultation de données primaires) et les technologies mises à disposition (métadonnées, un certain nombre de services web –des outils de traitement de la langue ou d'annotation).
CLARIN est structuré autour de cinq grands chantiers : Infrastructure technique, ressources linguistiques, usages, aspects légaux et perspectives.
Pour donner une idée des volumes de données, on peut dire que 892 ressources ont été recensées et 236 outils.

Il s'agit d'un projet institutionnel mais qui reprend les idéaux des « humanités digitales ». Ainsi, le consortium assure la promotion des formats et des normes qui permettent l'interopérabilité et donc favorise la dissémination des ressources[50] . CLARIN structure un champ de recherches jusque-là éclaté : les utilisateurs de ressources linguistiques tendent ainsi à se constituer en communauté autour de mêmes données, mais également de mêmes vocabulaires, entre autre autour de l'outil ISO-cat.

CLARIN vise à ce que ces données soient interopérables (quel que soit leur format et les différences terminologiques entre les différentes langues de l'UE), qu'elles soient stables (disponibles sans interruption), persistantes, accessibles quelle que soit la méthode utilisée et que l'infrastructure puisse accueillir toutes sortes de nouvelles données. Elle est donc extensible. L'enjeu de CLARIN ait que l'on ait un accès unique à l'ensemble de ces ressources, qui par ailleurs sont disséminées.

Un enjeu de CLARIN consiste à faire en sorte que les ressources linguistiques ne soient pas seulement utiles aux linguistes, mais bien à l'ensemble des SHS et aux sciences de l'ingénieur utilisant les ressources linguistiques. En effet, un des enjeux de CLARIN, au-delà de la mise en commun des ressources linguistiques, ce sont les questions du multilinguisme en Europe[51]
.

[47] http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=270902
[48] http://ec.europa.eu/information_society/apps/projects/index.cfm?menu=secondary&prog_id=IPSP
[49] http://www.clarin.eu/
[50] http://www.flarenet.eu/sites/default/files/FLaReNet_Standards_Landscape.pdf
[51] http://workshops.elda.org/lrslm2010/

Nous serons amené à parler de certains projets inclus dans CLARIN, entre autre concernant les métadonnées IMDI. En effet, autour de CLARIN opère la mise en commun et la normalisation de la description des ressources linguistiques, qui sont des tâches impliquant les descriptions documentaires.
En ce sens, les questions de représentation de données, et par la suite celle de la description documentaire, constituent un enjeu crucial de CLARIN. Conçu au départ comme une mise en relation de ressources, CLARIN amène à la création d'une collection distribuée de ressources diverses impliquant un effort important de normalisation dans la gestion des métadonnées, via notamment le projet CMDI[52].
En définitive, une infrastructure crée une communauté scientifique, et par extension, un langage commun, qui sert d'abord à la description des objets que l'on rend publics. CLARIN correspond très précisément au rôle structurant des données primaires, des outils d'analyse et du langage de spécialité pour la formation d'un domaine de recherche. CLARIN s'appuie sur l'expérience des infrastructures crées depuis de longues années dans d'autres domaines (un panorama en France a été publié par le ministère de la Recherche et de l'Enseignement Supérieur[53]. L'intérêt spécifique de CLARIN réside dans le fait que l'autorité politique crée cette communauté.
Les ressources documentaires sont exclues de ces infrastructures, ce qui ne manque pas de poser le problème de la corrélation entre les données et les outils d'une part, et les résultats obtenus et publiés d'autre part. Notre projet de recherche contribue à combler ce fossé.

1.2.2.3. DARIAH : fédérer les ressources numériques en Sciences Humaines.

DARIAH est une infrastructure dont la phase exploratrice commence en 2008[54] (). Il constitue une fédération d'initiatives nationales relative aux infrastructures, à la recherche et à l'enseignement sur le web. DARIAH privilégie une approche orientée services et considère l'usager comme central. DARIAH soutient donc des projets d'accès aux données, favorise l'échange de connaissances, de méthodologies et de pratiques dans l'ensemble des champs des SHS. Plus techniquement, DARIAH recommande l'usage de standard et de bonnes pratiques.

Le rôle de DARIAH est beaucoup plus vaste que celui de CLARIN puisque le domaine concerné est celui de l'ensemble des SHS. DARIAH vise à mettre en relation des initiatives fondamentalement diverses de façon à les rendre accessibles. Il s'agit d'une infrastructure permettant l'échange et la valorisation de propositions élaborées dans des cadres nationaux. En France, c'est le TGE-ADONIS (en lien avec le CCSD, l'ABES et le CLEO) qui constitue le correspondant de DARIAH. DARIAH apparait comme une « méta-infrastructure » et permet de faire le lien entre les différentes infrastructures existantes dans les SHS.
DARIAH se propose de coordonner des actions nationales, et donc chaque état garde la maîtrise des actions envisagées. L'objectif est que ces innovations soient coordonnées et que l'expertise soit partagée. DARIAH s'appuie donc sur des projets plus limités dont il cherche assurer les transferts et le partage des compétences.
En ce sens, ce sont plutôt les projets soutenus par DARIAH qui peuvent apparaître les plus intéressants par rapport aux bibliothèques numériques. (Par exemple NeDiMAH (Network of Digital Methods in the Arts and Humanities[55]) est un projet qui vise à promouvoir l'échange de méthodes au sein et entre les différentes disciplines scientifiques en SHS. Le CESSDA

[52] http://www.clarin.eu/content/component-metadata
[53] http://cache.media.enseignementsup-recherche.gouv.fr/file/TGIR/29/6/infras_def3_243296.pdf
[54] http://www.dariah.eu/index.php?option=com_content&view=article&id=7&Itemid=119
[55] http://www.esf.org/index.php?id=8752

(Council of European Social Science Data Archives[56], en lien avec le réseau QUATELET en France constitue un autre assemblage, centré sur les Sciences Sociales, et plus particulièrement les données démographiques et socio-politiques. DARIAH reste centré sur les « Humanités ».
L'innovation technique que constituent les grilles et donc les circulations de quantités énormes de données, modifie considérablement l'accès aux documents historiques et littéraires. Cette opportunité permet une transformation complète de la façon dont les données, notamment historiques, sont disponibles. L'enjeu pour la discipline est aussi important que les corpus numériques pour la linguistique.
Les grilles (et les nuages) deviennent un enjeu scientifique majeur, qui construit sa propre communauté scientifique autour de la dénomination « big data ».
Surtout, ce qui est vraiment important, c'est le fait que les sciences humaines passent du statut de sciences à matériau pauvre (des archives papier disséminées) à des sciences à matériau riche (des archives numériques conséquentes et en accès libre, au-delà des frontières). On peut donc attendre de l'accroissement de l'amplitude des données une transformation des pratiques mais au-delà de la façon d'envisager un problème scientifique. Cette amplification quantitative des données disponibles (mais également des outils techniques pour les traiter) entraîne un changement important du statut des sciences humaines au sein des communautés scientifiques[57]. Il n'échappera pas que les enjeux relatifs à la construction de l'identité européenne sont également présents derrière la mise en commun des archives de l'ensemble des pays européens.
DARIAH a donc comme rôle d'accompagner cette mutation. Au vu de l'ampleur de l'objectif, la tâche actuellement la plus pertinente concerne l'affinement de la compréhension de cet environnement. En effet, DARIAH n'a vu son véritable démarrage qu'à partir du moment où les implications des grilles sur les SHS sont apparues fondamentales. Néanmoins, à plus long terme, l'objectif consiste à élaborer une plate-forme permettant de faciliter les échanges et de permettre à l'infrastructure de remplir son rôle de fédération des données, expériences et résultats. Cette mise en commun concerne également l'enseignement, avec de mêmes engagements dans la mise en place à la fois de ressources et de communautés. Les environnements virtuels de recherche sont encouragés dans leur mise en œuvre (mais non dans leur élaboration). Aux questions de stockage, d'interopérabilité, s'ajoutent celles concernant l'accès aux données et donc finalement, celles relatives à la description des documents. Enfin, ces questions engagent le programme dans des problèmes liés à la dissémination des objets scientifiques.
L'avancement des travaux fédérés au sein de DARIAH ne permet de donner une vision complète de l'offre, mais seulement d'une organisation qui se met en place. Néanmoins, comme CLARIN précédemment, elle montre la capacité des institutions européenne à construire des infrastructures fédératives, sachant que pour DARIAH, le projet inclut les bibliothèques et le monde de l'enseignement.

1.2.2.4. Plateformes complémentaires et à portée plus limitée.

Ces initiatives relativement globales ne doivent pas occulter des plateformes plus limitées dans leurs objectifs comme FLARENET pour les ressources linguistiques, ou META, une plate-forme dédiée au développement de l'information multilingue. Ces initiatives sont celles que justement une infrastructure comme DARIAH souhaite fédérer, selon un système qui peut

[56] http://www.cessda.org/
[57] (Voir par exemple : http://horizon2020projects.com/societal-challenges/the-role-of-humanities-and-social-sciences/ et http://horizons.mruni.eu/).

s'apparenter au rôle d'EUROPEANA par rapport aux bibliothèques des musées. Or, justement, le cas de META et de ses métadonnées spécifique montre que l'interopérabilité avec les infrastructures n'est pas facile : en effet, les plateformes ont pu développer des outils de description, des thésaurus ou adopter des formats qui ne peuvent être articulés aux outils des infrastructures qu'après un long travail de mise en relation concept par concept.
Si les initiatives restent localisées (régionales, nationales notamment), le niveau européen se caractérise par sa capacité à fédérer en facilitant les échanges. Ainsi, deux points émergent en dehors des questions techniques d'interopérabilité : d'une part la description des documents parce que c'est ce qui permet d'identifier les documents, d'autre part la recherche d'information, qui doit être envisagée comme un accès au plus grand nombre de ressources à partir d'un point unique.
Au niveau technique, la concurrence entre les plateformes (relativement légères) et les infrastructures a des conséquences importantes en matière d'outil et de langage de représentation. Entre des services relativement restreints et finalisés proposés par les plateformes, destinés à des usages limités et circonscrits, et des infrastructures regroupant une offre beaucoup plus étendue, utilisable par des publics différents, et qui est ouverte à des usages différents. Nous avons vu que ces différences d'objet recouvraient en grande partie des différences d'acteurs (institutions publiques de la recherche pour les infrastructures, acteurs plus proches du secteur privé pour les plateformes.

Nous nous sommes intéressés essentiellement aux projets en cours en SHS, et plus particulièrement les ressources linguistiques. Ces domaines sont en lien à notre projet de recherche. A titre d'exemple, d'autres infrastructures européennes sont développées pour d'autres disciplines : LIFEWATCH, qui est une infrastructure dévolue aux données et observatoires de la biodiversité, EURO-ARGO, pour l'océanographie. Ces projets sont développés dans l'esprit de la « feuille de route » d'ESFRI 2010[58] .

En définitive, les infrastructures manifestent une volonté forte de structuration de la recherche et de la mise en valeur du patrimoine culturel communautaire. Néanmoins, le fonctionnement est relativement complexe puisque les dimensions nationales sont les principales productrices des ressources. Par ailleurs, les infrastructures ne financent pas des recherches : elles utilisent des résultats de recherches (portant sur les métadonnées, ou sur les annotations) financées par de tout autres moyens et sans qu'elles interviennent nécessairement ; par ailleurs souvent ces outils ou ressources sont nettement antérieurs à l'initialisation de ces infrastructures (comme la TEI par exemple).
Par conséquent, si les infrastructures ont un rôle fédérateur, elles ne sont pas contraignantes. En effet, il est possible de développer des services et des plateformes scientifiques en dehors d'un lien aux infrastructures. Seule la popularité de l'usage aura un rôle contraignant pour les plateformes par rapport aux infrastructures.
Cette situation s'explique par le fonctionnement de la recherche sur les outils constituant des infrastructures. Les outils utilisés s'appuient sur les LINKED DATA, qui constituent une initiative totalement extérieure aux politiques publiques dirigeant les infrastructures. Cette indépendance de la recherche garantie la pluralité des propositions et donc l'adoption par les plateformes de solutions qui ne sont pas nécessairement celles des infrastructures. On détaillera plus loin cette question mais elle doit rester à l'esprit pour caractériser les infrastructures.

Si nous avons mentionné les initiatives émanant des institutions européennes, d'autres

[5858] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri-roadmap

initiatives supranationales telles que KNOWLEDGE EXCHANGE (que nous avons déjà mentionnée) qui regroupe les pays d'Europe du Nord[59]. Ces initiatives sont liées à des problèmes particuliers, notamment dans le cas présent, l'échange d'expertise entre des institutions chargées de développer les usages des nouvelles technologies numériques dans l'enseignement supérieur et la recherche. Ces projets, qui sont aussi des services réels, sont justement ce que les grands projets d'infrastructures (DARIAH, CLARIN) cherchent à fédérer.

Les projets que l'on vient de présenter sont fondamentalement disciplinaires ou pluridisciplinaires et sont fondés sur la transformation des pratiques scientifiques. D'autres projets sont en cours, eux sans spécificités disciplinaires et qui offrent un service relativement limité. Il peut s'agir d'environnements de recherche comme le TexGRID allemand[60], qui propose à la fois un outil d'archivage et une interface pour la menée d'un travail de recherche ("virtual research environment, will provide integrated access to both new and existing tools and services via a user friendly software. The TextGridLab will be improved according to the suggestions of the TextGrid user communities and supplemented by integrating new tools such as optical character recognition for Gothic print during the term of the project »), soit limités au dépôt de publications (comme OpenAIRE et OpenAIREplus[61] , thèses (DART-Europe)).

Concernant le stockage pérenne des données scientifiques, un autre projet est en cours : EUDAT[62]. Il s'agit de promouvoir l'accès open data des données de la science tout en s'assurant de la pérennité des données. CLARIN utilise EUDAT pour le stockage des données linguistiques. EUDAT comprend également le renseignement par les métadonnées.

On peut étendre l'horizon en mentionnant le projet BAMBOO[63] , qui constitue une initiative essentiellement américaine, proche, dans ses objectifs, de DARIAH. L'effort de structuration ne concerne donc pas seulement l'UE, mais bien l'ensemble des organismes chargés de structurer la recherche dans le monde. En effet, la question est mondiale et se concrétise aujourd'hui avec la création du RDA[64] . Il s'agit du partenariat, initié en 2012 entre l'Union Européenne, de l'Australie et des Etats-Unis pour réfléchir à des infrastructures mondiales de circulation de données. L'alliance RDA travaille avec le DC pour définir les caractérisations de document pertinentes pour ce type d'infrastructure.

On comprend mieux maintenant l'importance des formats d'échange et de représentation. Ils permettent à la multiplicité des projets de s'assurer d'une certaine visibilité.
Si l'on ne regarde que les projets nationaux, on aboutit à des entreprises relativement hétérogènes. Cette hétérogénéité est essentiellement marquée au niveau local et ne concerne que la nature des projets. Le rôle des projets européens consiste avant tout à structurer ces initiatives, sous forme d'un soutien, mais surtout en s'assurant de leur mise en relation. De façon pragmatique, l'aboutissement est bien la construction d'un réseau d'initiatives.
Cela dit, l'enjeu dépasse largement la proposition d'un accès facilité aux services proposés aux chercheurs. Comme la roadmap 2020 l'indique, cette structuration contribue à la mutation du travail de chercheur, considérée comme une nécessité stratégique (P. 18).

[59] http://www.knowledge-exchange.info/.
[60] http://www.textgrid.de/en/ueber-textgrid.html
[6161] http://www.openaire.eu/
[62] http://www.eudat.eu/
[63] http://www.projectbamboo.org
[64] http://rd-alliance.org/about.html

Ce changement de cadre implique très fortement les bibliothèques, jusque-là peu présentes dans des projets sectoriels visant notamment la structuration des données et la mise en place de services pour les chercheurs.

Les langages de représentation comme les méthodes de description, sont communs entre ces différentes initiatives. En effet, elles constituent le socle de l'ensemble des initiatives. La mise en relation entre ces différentes initiatives reste encore l'objet de nombreuses recherches (notamment au sein de CLARIN -voir LREC 2012 où différents jeux de métadonnées pour les ressources linguistiques sont présentés[65]-), ce qui veut dire que la mise en valeur de l'assemblage de ces différents outils (qu'est-ce que produit leur fédération) n'est pas encore pensée dans l'ensemble de ses implications. Enfin, un certain nombre de positions d'acteurs et d'implications ne sont pas encore définies. Si les bibliothèques semblent encore extérieures à ces infrastructures, leur rôle devient de plus en plus important, entre autre via LIBER, notamment parce que leur expertise en matière de numérisation, de description d'objets numérisés et de structuration de collections est essentielle, notamment pour DARIAH et EUROPEANA.
Les implications dont on vient de parler concernent essentiellement la recherche et, pour EUROPEANA, le monde de la culture. Les bibliothèques via LIBER sont associées à titre d'expert dans DARIAH et EUROPEANA. Dans DARIAH, l'implication pourrait être plus importante, notamment via les infrastructures nationales (comme par exemple le partenariat entre le TGE-ADONIS (qui pourrait s'appeler ADONIS/CORPUS) et les bibliothèques numériques, comme HAL par le biais du moteur de recherches ISIDORE).

En définitive, il semble bien que le projet de structuration de recherche, qui passe par les infrastructures, s'élargisse vers des types de données que traitent les bibliothèques. Alors, elles deviennent en tant qu'expert, et probablement bientôt en tant que gestionnaires de collections, de acteurs de plus en plus incontournables de la structuration de l'information scientifique numérique. Néanmoins, ces bibliothèques sont également confrontées à une autre évolution, celle de l'édition. Nous allons l'aborder essentiellement du point de la publication numérique, sachant que d'autres questions deviennent d'actualité comme les techniques qui permettent de tagger un texte de façon à ce qu'il soit relié automatiquement à ses sources, ses données primaires ou expérimentales. Cela dit, la question du document publié doit être pensée et résolue auparavant.

Les infrastructures apparaissent comme un outil permettant de fédérer des ressources existantes, de les rendre visibles, d'échanger des expériences et de construire une offre de service multiple à l'usager. C'est en ce sens que les infrastructures ont un rôle stratégique. Cette perspective de services fédérés requiert des ressources en nombre élevé et un objectif fédérateur, qui est d'ordre politique, important. C'est la raison pour laquelle la dimension européenne apparait la plus pertinente.
Néanmoins, comme on vient de le voir, ce n'est pas la seule possible. Des politiques nationales d'envergure incluant l'enseignement, comme en Grande Bretagne, ou fondées sur l'interdisciplinarité comme aux Pays Bas, peuvent néanmoins se développer.

1.2.2.5. Evolutions des problèmes de l'édition scientifique électronique.

Le panorama des implications des bibliothèques numériques dans le cadre des projets européens ne serait pas complet sans l'initiative PEER, qui constitue dans le programme e-

[65] http://www.lrec-conf.org/proceedings/lrec2012/index.html

content + de l'Union Européenne, le lieu dans lequel se réfléchit l'avenir de l'édition scientifique électronique. Il s'agit de trouver une solution à la difficulté posée aux éditeurs par la disponibilité des ressources numériques. PEER regroupe l'ensemble des partenaires et un travail très important est mené à propos des usages. En effet, le dépôt de documents dans des archives ouvertes et institutionnelles croit de façon exponentielle et constitue par ailleurs un critère pour l'évaluation de la production des chercheurs. Le projet PEER a duré de 2008 à 2012, et s'est conclu en mai 2012. Il permet de disposer d'un état des lieux à la fois récent et approfondi.

PEER distingue trois formes d'archivages : les prépublications (données primaires de la recherche, données brutes et manuscrits soumis à un éditeur), les manuscrits corrigés (manuscrits d'auteur, intégrant les apports des pairs et acceptés pour publication par l'éditeur) et les publications éditées (article publié sur le site de l'éditeur : intégrant le process éditorial, la mise en page, ouvert à la citation, au référencement, aux liens avec d'autres articles ou matériaux de recherche). La dernière voie est celle que préfèrent les chercheurs.

Le but de PEER a été l'établissement d'un observatoire et pour cela a requis un dépôt, à la fois des auteurs et des éditeurs. Il a permis la constitution d'un corpus satisfaisant pour mener une étude de comportements et d'usages.
Il est apparu essentiel de comprendre comment les utilisateurs (qui sont en même temps des auteurs) appréhendent l'archivage des publications et leur accès. Ces analyses se sont fondées sur la publication de manuscrits corrigés, à savoir la "voie verte", qui est celle qui mobilise le plus les différents partenaires (activité de dépôt volontaire), et qui par ailleurs ne remet pas totalement en question le modèle économique des éditeurs (puisque ces publications ne contiennent ni la pagination ni la mise en forme finale des articles).

Ces analyses ont été menées par deux équipes indépendantes, et avec deux problématiques distinctes :
- les comportements des chercheurs et la façon dont le dépôt transforme le fonctionnement scientifique. "the role of stage-two manuscript repositories in the scholarly communication system by exploring perceptions, motivations and behaviours of authors and users. »(Rapport final [20120618_PEER_Final_public_report_D9-13], p. 9).
- Les usages concernent la façon dont les usagers sélectionnent les de plateformes pour la recherche et le téléchargement des publications (sites d'éditeurs, bibliothèques numériques). Considérant la multiplicité des lieux de dépôt des publications, l'étude vise à comprendre ceux qui sont privilégiés par les chercheurs : « usage trends at publisher and repository platforms on the basis of article level usage data, whether stage-two deposits increase access and use and which effects large-scale deposits may have on journals ». (Rapport final[66] , p. 11).

En résumé, les comportements montrent une communauté scientifique à la fois favorable aux archives ouvertes mais ne souhaitant pas que les règles du fonctionnement scientifique (et notamment l'importance des revues) soient modifiées.
Les usages visent à mesurer, sur une période de deux ans, les visites et téléchargements sur les sites éditeurs et sur PEER en se posant notamment la question de l'impact d'une prépublication sur le site PEER sur le téléchargement de l'article sur le site de l'éditeur[67] : p. 5 « The purpose of the experiment was to see whether this allocation made any difference to use at the publisher web site? »

[66] http://www.peerproject.eu/fileadmin/media/reports/20120618_PEER_Final_public_report_D9-13.pdf
[67] http://www.peerproject.eu/fileadmin/media/reports/20120618_D5_3_PEER_Usage_Study_RCT.pdf

Les conclusions de l'étude montrent globalement que les dépôts en bibliothèques numériques et sur les sites des éditeurs sont complémentaires : « that are visible in PEER are associated with higher average downloads than those that are hidden, at least in the case of the life and physical sciences where the probability that these results may have arisen by chance is very low, less than one in a hundred. For these subjects, and in the case of larger publishers, we reject the `no effect' publisher hypothesis. Visibility does make a difference and in a positive direction for publishers.
In the case of medicine and the social sciences and humanities, visibility in a PEER repository was also associated with more publisher downloads, although here (and in the case of smaller publishers) this could be explained by chance rather than a strong clear signal. » (ibid. 18)

Cette complémentarité ne doit pas être trop affirmée : les auteurs sont nettement plus prudents et circonspects que ne semblent l'indiquer les conclusions précédentes, d'autant plus que des paramètres comme les moteurs de recherche, les métadonnées utilisées et reconnues par ces derniers, etc. peuvent fausser les résultats de l'enquête.

Ces biais sont reconnus dans l'exposé de l'équipe CIBER lors de la rencontre-bilan de Bruxelles, en mai 2012 : « Overall, PEER is associated with a significant, if relatively modest, increase in publisher downloads, in the confidence range 7.5% to 15.5%.
The likely mechanism is that PEER offers high quality metadata, allows a wider range of search engine robots to index its content than the typical publisher, and thus helps to raise the digital visibility of scholarly content ». [68](p.24)
En même temps, l'équipe CIBER soulève un certain nombre de questions relatives au rôle des descriptions bibliographiques et des choix de méthode pour une prise en compte par les moteurs de recherche. Un certain nombre d'enjeux peuvent se cristalliser sur des questions de description bibliographique.

On remarquera enfin qu'une attention très particulière a été portée aux métadonnées et à l'instauration de pratiques standardisées de renseignement. Plus globalement, les rédacteurs du rapport final insistent sur les convergences techniques qui concernent, outre les métadonnées, les formats et les interfaces d'accès.

Les enquêtes du projet PEER sont limitées à la question des publications (ce qui est l'enjeu du programme). Elles ne permettent pas de comprendre comment les différents services du web scientifique sont utilisés et par qui, et pourquoi. D'ailleurs, la question de l'e-science n'est pas évoquée dans les rapports de PEER. Or, il s'agit d'une transformation fondamentale de la pratique de la recherche, et qui concerne entre autre les pratiques informationnelles.
On pourra considérer que les infrastructures constituent une étape dans la transformation des pratiques de recherche. Elles structurent l'offre de ressources, données primaires, outils et vocabulaires. Cette offre sera ensuite tout naturellement intégrée à l'intérieur des procédures d'e-science.

En conclusion, l'initiative PEER dans son ensemble a permis de résoudre un certain nombre de problèmes relatifs à l'intégration des publications scientifiques dans le cadre de l'ouverture des données. Ainsi, la barrière que constituaient les éditeurs scientifiques à l'accès public aux résultats des recherches est levée. Par conséquent, les perspectives de mise en ligne de l'ensemble des ressources nécessaires à l'activité scientifique se font plus précises.

[68] http://ciber-research.eu/download/20120631-PEER_CIBER_Brussels.pdf

Plus globalement, les projets d'infrastructures de services à la recherche montrent un dynamisme important et impliquent des acteurs très différents et des configurations originales, sans structuration unique. On peut considérer qu'il s'agit d'une véritable opportunité pour l'émergence de services pertinents et innovants. Le fait est que les bibliothèques sont aujourd'hui intégrées à l'intérieur d'un mouvement qui dépasse largement leur seul domaine d'activité, et qui englobe l'ensemble des données numériques intéressant la recherche, l'enseignement et la culture.
Un tel développement est également rendu possible par l'usage des standards du W3C, notamment RDF et OWL et donc toutes les possibilités d'interopérabilités qui leur sont liées.

1.3. La question de l'e-science.

La question de l'e-science constitue l'un des enjeux majeurs pour l'ensemble des acteurs impliqués dans le cadre de l'information scientifique et technique parce qu'il s'agit d'abord d'une transformation des méthodes de travail des usagers, les scientifiques. En même temps, il s'agit d'une politique publique visant à réduire les coûts de la recherche et à opérer des économies d'échelle en favorisant le réemploi des données de travail. Il s'agit enfin d'efficience par un accès facilité à l'ensemble de la production scientifique pertinente.
Cette perspective est particulièrement pertinente pour les acteurs des bibliothèques numériques parce qu'elle renouvelle l'approche que l'on peut avoir des fonds (notamment l'usage des données primaires), de leur mise en valeur (par le biais de leur description), et de leur usage. L'enjeu de l'e-science explique à la fois le foisonnement, la diversité des acteurs et des politiques, et les objectifs des infrastructures précédemment présentées.

On peut reprendre la citation de SIR JOHN TAYLOR, director office of science and technology, uk, [reprise en exergue de CAROLE GOBLE dans iswc 2005 8th november 2005[69]], pour résumer les enjeux d' l'e-science : "e-science is about global collaboration in key areas of science and the next generation of [computing] infrastructure that will enable it."

L'hypothèse première consiste à considérer que la validation des hypothèses gagne en rapidité et en crédit en utilisant des analyses informatiques et des outils de traitement automatique. Cela permet d'accentuer l'importance des tests fondés sur des collections (ou populations) et des simulations, en plus des expérimentations traditionnelles. Cette première ambition est aujourd'hui relayée par la mise à disposition en ligne de l'ensemble des autres ingrédients de l'activité de recherche.
Cette multiplication des données d'études et des outils transforme considérablement l'activité scientifique, qu'il s'agisse de l'utilisation et de la réutilisation des données, de la de validation des hypothèses ou encore de la vérification des résultats. Cette question fondamentale sera traitée après avoir présenté les Linked Data. En effet, le rôle des Linked Data dans l'e-science est fondamental puisqu'elles introduisent de nouvelles méthodes de travail. Néanmoins, nous en restons pour le moment aux seules questions de réseau et d'infrastructure. En effet, toutes ces données occupent une place considérable (en termes de mémoire) et mobilisent pour leur activation des flux de données très importants. Par conséquent, au-delà de la disponibilité des données, se pose la question des infrastructures de transfert d'information. Pour répondre à cela, l'union européenne a mis en place le programme CHIST-ERA qui vise à construire les grilles (ou « grid »). CHIST-ERA est un programme de recherche européen fondé sur les enjeux à long terme des sciences et technologies de l'Information et de la Communication. L'élaboration des contenus de cette infrastructure est

[69]« e-Science and the Semantic Web » opening keynote at 4th Intl Semantic Web Conference 2005, Galway Ireland, 6-10 November 2005 (19Mb), http://www.semanticgrid.org/presentations/ISWC2005keynote-final.ppt

Code de champ modifié

déjà en cours : [70]
P. 4 « the challenge is to produce new computational concepts, models, tools and methodologies to automatically and reliably extract new knowledge from large amounts of heterogeneous, unstructured data. Typical data include multilingual and multimedia data such as found on the web (text, speech, image, video, …) and data generated by human organisations in the course of scientific, industrial or service activities (medical data, 3d object representations, advanced manufacturing data, …). The data are processed to produce higher level new knowledge, typically a semantic description of the content of the data, or elaborate models, scripts or experiential knowledge, which might in turn be used to process other data. Such production of new knowledge involves complex processing systems whose reliability cannot be determined analytically but can be estimated through confrontation against representative data sets ».

Cela dit, le projet n'est pas seulement technique. Il est même avant tout social. Il s'agit de renforcer les collaborations entre équipes, mais également de fédérer des travaux autour de données et d'outils communs. D'ailleurs, le programme CHIST-ERA 2012 place au centre les questions d'interface et d'usage[71] .
Une telle perspective permet de travailler sur une quantité de données et d'outils nettement plus importants. Surtout, elle permet une meilleure vérification des résultats, la reproduction des expérimentations et l'éventuelle détection d'erreurs ou de biais. Ainsi, elle permet de mieux évaluer la qualité des publications. Les SHS utilisent néanmoins de plus en plus ce type d'outils, notamment la linguistique (linguistique computationnelle, linguistique de corpus, traitement de corpus oraux, analyseurs). CLARIN peut ainsi apparaître comme la première pierre pour la construction d'une structure d'e-science pour les utilisateurs de ressources linguistiques.

L'e-science est avant tout un projet, à savoir qu'il s'agit de transformer à la fois des outils, des pratiques et des méthodes de travail. Le projet a été formulé entre autres par [TRisse]:
« For such an infrastructure, there are various areas with a potential for improved innovation process support like:
•a more effective re-use of scientific results, information and various other kinds of "innovation" resources
•facilitating the collaboration in dynamically created multidisciplinary teams
•an accelerated and more effective fostering of technology transfer
•intelligent support for routine tasks enabling the researcher to focus on the creative parts of innovation
•a flexible management of the innovation process, which takes into account the dynamics and the creative character of this process ».

Ces possibilités sont offertes également par l'interopérabilité des descriptions de ressources hétérogènes, corpus et outils d'analyse : les perspectives d'une intégration de ressources diverses dans le cadre de pratiques d'e-science, comme celles projetées par D-SPIN (german infrastructure for language resources and tools) ou TEXT-GRID (virtual research environment for the art and humanities), constituent des avancées importantes en matière d'outils pour la gestion de l'e-science. D-SPIN est intégré dans l'offre de CLARIN [72]. TEXT-

[70] HTTP://WWW.CHISTERA.EU/SITES/CHISTERA.EU/FILES/CHIST-ERA%20CALL%202011%20-%20CALL%20ANNOUNCEMENT.PDF
[71] http://www.chistera.eu/call-2012-topics-and-keywords
[72] http://www.clarin.eu/node/1472

Code de champ modifié
Code de champ modifié
Code de champ modifié

GRID y est répertorié comme une collection de ressources primaires, autrement dit de corpus.

Les possibilités techniques de traitement offertes par les grilles (grid) ne concernent pas seulement les données (ou « big data ») : les espaces et interfaces de travail sont également repensées dans cette perspective (voir le programme CHIST-ERA). Ainsi, des espaces de travail scientifiques commencent à voir le jour, comme par exemple l'environnement de recherche open-source eSciDoc[73]. L'ensemble de ces projets est énuméré dans le « virtual research environment collaborative landscape study » de janvier 2010 (auteurs : A. Carusi et T. Reimer)[74].

ESCIDOC présente l'avantage d'être une plate-forme pluridisciplinaire et modulaire. D-spin[75] est l'infrastructure d'e-science dévolue à la linguistique et incluse dans CLARIN. Ces plateformes (comme également en France le TGE-ADONIS entre autre au travers d'ISIDORE[76]) ont vocation à fournir des services, notamment des répertoires. La particularité de l'assemblage de services de sSciDoc et D-SPIN réside dans sa spécialisation relative aux ressources linguistiques.

Le développement de l'e-science est plus ou moins accentué en fonction des pays, sachant qu'en Europe, l'ensemble des outils tend à être fédéré dans le cadre des grandes infrastructures. En effet, si l'on observe les projets en cours proposés en Hollande[77], on s'aperçoit qu'il s'agit bien de projets disciplinaires. Les méthodes, sources et outils d'analyse sont spécifiques à certaines disciplines. Par contre, les outils, les expériences et les modèles qui servent à les penser puis à les réaliser sont eux faciles à échanger et réutiliser.
Les enjeux de l'e-science sont différents entre les domaines scientifiques, du fait de procédures de validation et de vérification différentes. On peut citer la reproductibilité des expériences dans le domaine biologique, la vérification des sources chez les historiens, les données et leur traitement en économie, les corpus et leur annotation en linguistique, etc.

Nous ne développerons pas l'ensemble des dimensions de ces projets. Nous en donnerons deux aperçus importants dans notre domaine de connaissances, l'Human Language Project et la question des bibliothèques d'ontologies. Nous poursuivrons par l'implication des bibliothécaires et des documentalistes. En effet, l'e-science transforme considérablement la quantité, la nature et la structuration de l'information dont les chercheurs auront besoin. Ainsi, le rôle des bibliothécaires est considérablement transformé, notamment dans son interaction aux chercheurs. Enfin, nous aborderons l'activité des associations de bibliothécaires dans le cadre du développement de l'e-science (et donc des infrastructures).

1.3.1. Mise en ligne et traitement de ressources primaires. Quelques exemples de conséquences sur la recherche.

Nous proposons un très rapide aperçu de certaines des conséquences engendrées par le traitement en ligne des données numériques. En effet, les transformations auxquelles on assiste ne concernent pas seulement les outils de la recherche mais bien les données sur lesquelles elle se fonde et la façon dont elles peuvent être construites.

[73] HTTPS://WWW.ESCIDOC.ORG/).
[74] WWW.JISC.AC.UK/MEDIA/DOCUMENTS/PUBLICATIONS/VRELANDSCAPEREPORT.PDF
[75] HTTP://WEBLICHT.SFS.UNI-TUEBINGEN.DE/INDEX.SHTML
[7676] WWW.RECHERCHEISIDORE.FR
[77] http://www.esciencecenter.com/projects/project-portfolio/

Code de champ modifié
Code de champ modifié
Code de champ modifié

Une question fondamentale se pose à l'organisation du traitement de grandes quantités de données. Cette question se pose, par exemple, pour l'annotation de corpus. Néanmoins, le problème est relativement différent s'il s'agit d'un travail qui ne peut être mené qu'à petite échelle (notamment les annotations manuelles ou semi-automatiques) ou d'un ensemble de données qui ne peut être appréhendé que parce qu'il est organisé en prenant en compte la quantité des données.
Un exemple de ce dernier cas est celui de l'Human Language Project[5]. Son objectif est une collection de données intégrant l'ensemble des langues du monde. Ainsi, sont concernées des données primaires et non des publications avec un objectif précis : permettre le traitement automatique de données quantitativement importantes et diversifiées. Parallèlement à cet objectif de traitement, un autre est celui de la préservation des données primaires et par extension, de la connaissance de certaines langues.
L'originalité d'un tel projet réside également dans sa structuration. En effet, les langues ne sont pas traitées par famille, mais par séquences. Ces séquences sont faites de structures morphologiques dites d'apprentissage non supervisé. (...)
L'intérêt de ce projet, qui s'inspire du génome, réside dans le fait qu'il va traiter cet ensemble de données que sont les réalisations linguistiques (orales, écrites) en n'utilisant pas les taxonomies et classifications des langues.

Dans ce cas, la quantité des données disponibles rend possible un traitement complètement différent. Par ailleurs, ce traitement n'est pas totalement anodin: il montre qu'un certain nombre de classifications, ici les classifications de langues, peuvent ne plus être les seules opératoires pour caractériser des liens entre des langues. Cette autre approche, qui n'invalide pas les classifications traditionnelles mais montre qu'un autre point de vue est possible sur les relations entre les langues, peut modifier considérablement les approches des phénomènes linguistiques.

Néanmoins, l'exemple le plus important d'organisation d'outils numériques est constitué par les bibliothèques d'ontologies. Une bibliothèque d'ontologies peut être définie de la façon suivante : « An *Ontology library system* is a library system that offers various functions for managing, adapting and standardizing groups of ontologies. It should fulfill the needs for re-use of ontologies. In this sense, an ontology library system should be easily accessible and offer efficient support for re-using existing relevant ontologies and standardizing them based on upper-level ontologies and ontology representation languages. For this reason, an ontology library system will, at the very least, feature a functional infrastructure to store and maintain ontologies, an uncomplicated adapting environment for editing, searching and reasoning ontologies, and strong standardization support by providing upper-level ontologies and standard ontology representation languages."
[6], p. 63).

Les auteurs s'intéressent au système de bibliothèque, et non aux contenus qui sont déposés. Ils explorent huit bibliothèques, mais ignorent ONTOBEE[7] qui est l'un des plus aboutis de ces projets, et surtout celui qui héberge les ontologies biomédicales et la Basic Formal Ontology.

Ding & Fensel identifient un certain nombre de critères destinés à évaluer les bibliothèques d'ontologies. Trois grands critères sont retenus : (1) au niveau du « management », sont considérés l'accessibilité, l'identification de l'ontologie et de ses versions. (2) Au niveau de l'adaptation à l'utilisateur, sont considérés les critères pour la recherche d'une ontologie, l'édition de modifications, propositions et actualisations d'une ontologie, et enfin les

conséquences, à savoir la caractérisation de résultats de l'ontologie, et notamment les possibilités d'une évaluation. Enfin, (3) au niveau dit de la standardisation, sont retenus les critères de la généralité du langage utilisé et du haut niveau d'abstraction.

Cette étude déjà ancienne a à la fois comme intérêt d'avoir défini le champ et d'avoir proposé des critères d'évaluation toujours pertinent. Dans leur étude beaucoup plus récente (et qui reprend le cadre de Ding et Fensel), d'Acquin et Noy[8] montrent que l'adoption des standards du W3C a résolu de nombreuses questions relatives à l'interopérabilité, mais que les bibliothèques se sont considérablement diversifiées (en même temps que l'offre a considérablement augmenté). Ils intègrent la notion de collection dans la définition, de façon à distinguer la bibliothèque d'ontologie de l'indexation des ontologies par des moteurs de recherche. La notion de collection sert aussi à rendre compte de la structuration des objets de façon à répondre à un besoin de l'usager.
Cette précision à propos des critères permet de lier les bibliothèques de ressources aux autres bibliothèques. Simplement, le besoin de l'usager peut être défini différemment entre des bibliothèques de documents et celles de ressources. Néanmoins, la différence fondamentale réside dans la capacité d'intervention des utilisateurs sur les ressources, qu'il s'agisse de propositions ou de de mapping d'ontologies. En effet, l'interopérabilité des ressources permet à n'importe quel acteur, producteur d'ontologie, de mettre en relation différentes ontologies existantes.

Les bibliothèques de ressources peuvent être considérées au même titre que les autres bibliothèques numériques. La différence notoire est le fait qu'elles intègrent la dimension d'actualisation de façon systématique. De même, elles construisent une dimension communautaire par l'adjonction de listes de discussion intégrées dans des wikis associés aux sites.

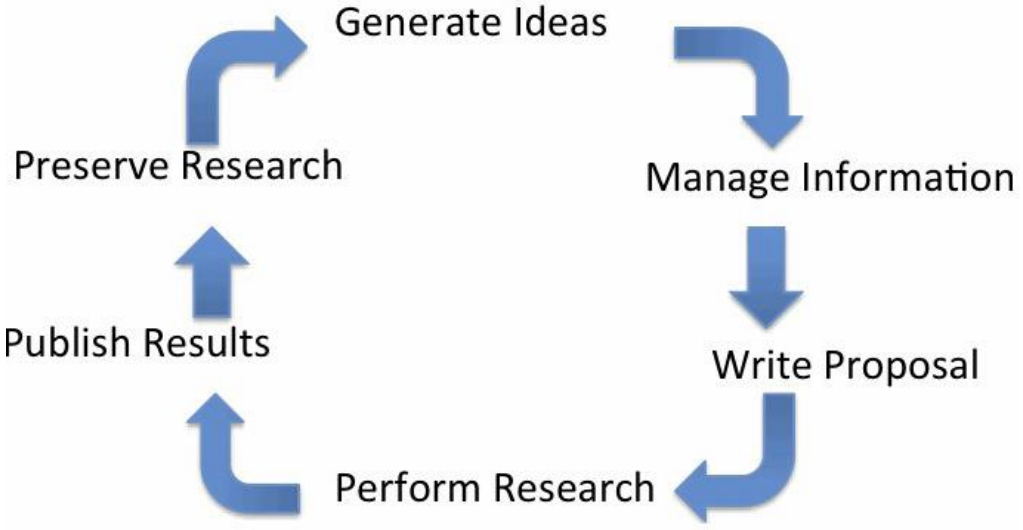
En définitive, ces deux exemples montrent les difficultés qu'il y a à structurer des données mais en même temps toutes les possibilités offertes à de nouvelles formes d'organisation. Si d'un côté on a la possibilité de traiter des données de façon totalement renouvelée, d'un autre se pose la question de la structuration de ces données et de leurs différents traitements. Ces questions, qui émergent aujourd'hui, font apparaître un champ de recherches en matière d'organisation des connaissances particulièrement riche, et qui ne se limite pas aux métadonnées : les classifications, thésaurus et lexiques structurés seront fortement mobilisés.

1.3.2. Questions pour les bibliothécaires scientifiques.

La perspective de l'e-science met en lumière et en perspective le développement des bibliothèques numériques dans le cadre d'une évolution globale des pratiques scientifiques. On a pu considérer l'importance des bibliothèques au travers de leur expertise au sein des projets d'infrastructure, et à l'intérieur de la politique de l'UE. Il s'agit là de positions d'organismes comme LIBER chargés de faire du lobbying. Or, les transformations dont nous parlons sont d'abord observables sur le terrain. Elles mettent également en lumière un rôle nouveau dévolu aux bibliothécaires, qui n'ont guère participé aux projets d'infrastructures. Bart Ragon (directeur associé à la bibliothèque de médecine Claude Moore)[78]
part du constat d'une faible implication des bibliothécaires dans les projets d'e-science, vus du point de vue des infrastructures. Il n'en va pas de même si l'on considère la gestion des

[78] http://library.med.utah.edu/or/pmayden/2012_Feb_22_Mayden_Escience_PPT_Ragon.pdf

projets de recherche et leur mise en œuvre au travers des infrastructures. Le point de vue de B. RAGON est certes fondé sur une expérience professionnelle dans le domaine médical.
Néanmoins, l'investissement des bibliothécaires dans le cadre de l'e-science est un élément essentiel du développement de la profession. Un portail[79], maintenu par une association professionnelle, est dédié à ces pratiques[80]. L'ensemble produit une revue scientifique : Journal of e-science librarianship[81]. L'approche développée dans cette publication est nettement plus pragmatique et fondée sur des projets locaux d'association entre des bibliothèques scientifiques et des centres de recherche développant des procédures d'e-science. Comme le notent ROMANO & alii, l'apport des bibliothèques est d'abord celui d'une expertise : "It is gaining momentum as a venue for librarians to collaborate with researchers and scientists like never before. The data curation necessary for eScience activities will provide librarians a new platform for demonstrating their expertise in data retrieval, collection, and storage"[9]. Plus précisément, cette intégration concerne la gestion des ressources sur l'ensemble du cycle de vie d'une recherche. Celui-ci est schématisé ainsi par HAMASU & alii[10] :



Cette stratégie d'intégration à l'intérieur de projets locaux ou régionaux constitue une approche différente de celle qui se met en place en Europe. Elle ne doit surtout pas faire oublier le travail de l'Association européenne des bibliothèques scientifiques[82], dont une partie est centrée sur l'intégration des bibliothèques scientifiques dans les projets d'e-science, eu égard à la distinction opérée entre infrastructure de recherche et bibliothèque. Les projets de cette association sont en grande partie fédératifs, au sens où ils visent à s'intégrer à l'intérieur de programmes impliquant d'autres dimensions que les seules bibliothèques. En premier lieu, CHAMBERS & SCHALLIER[11] propose un agrégateur de données intégré dans EUROPEANA.

Cette question de l'e-science n'est pas simplement pour les bibliothèques une question d'organisation des connaissances ou de spécialisation des services ; c'est avant tout une transformation du travail quotidien, du lien aux usagers et de la participation à leurs projets.

[79] http://esciencelibrary.umassmed.edu/about
[80] http://esciencecommunity.umassmed.edu
[81] http://escholarship.umassmed.edu/jeslib/
[82] http://www.libereurope.eu

1.3.3. Projets menés par les associations de bibliothèques et bibliothécaires.

Les questions que nous venons d'évoquer sont associées au niveau européen à l'engagement des associations professionnelles de bibliothécaires, notamment de celle qui est la plus concernée par les projets dont nous venons de parler, LIBER. On ne traduit pas exactement la position de l'ensemble des bibliothécaires, mais ceux qui sont les plus proches et les plus impliqués dans l'évolution de l'information scientifique et technique en Europe.

L'Association européenne des bibliothèques scientifiques constitue un groupement professionnel qui s'inscrit à l'intérieur du processus que l'on a caractérisé jusqu'à présent. Nous avons aperçu le point de vue américain. Maintenant, nous présentons la façon dont les bibliothèques s'intègrent dans le cadre des transformations de la recherche au niveau européen. La spécificité de l'engagement des bibliothécaires s'explique par le fait que derrière le développement de l'e-science, se poursuit le projet d'intégration européenne, et que celui-ci se traduit aussi dans le cadre global et spécifique des bibliothèques. Le développement d'EUROPEANA constitue une prémisse à la construction de cet espace commun de bibliothèques scientifiques. (Il est à noter que la BNF participe également à EUROPEANA). Cette stratégie est différente de celle qui consiste à intégrer les bibliothèques dans les programmes de recherche des universités et de formuler des solutions locales. Nous reviendrons sur certains engagements de ce réseau par rapport aux LINKED DATA ou sur les métadonnées.
La position de l'association par rapport à les questions de l'open access a été rendu publique le 24 juillet 2012 [83] : elle consiste à approuver pleinement la "voie verte", donc l'open access et la libre circulation de l'information à l'intérieur de l'union. La « voie verte » consiste à rendre accessibles gratuitement les publications scientifiques acceptées par un journal et corrigées, mais sans la mise en forme éditoriale qui constitue le propre de l'activité de l'éditeur. La version mise en forme reste elle, payante et n'est accessible qu'auprès de l'éditeur. Les recommandations de la commission ont été élaborées à la suite des conclusions de PEER notamment[84] .
La conséquence de cette position est la participation active aux projets d'infrastructures : le déclencheur politique est bien évidemment la politique d'ouverture des données au niveau européen, par l'émergence de solutions vis-à-vis des éditeurs et par l'élargissement de la vocation des infrastructures, au-delà des ressources des communautés scientifiques. On pourra encore mentionner le rôle d'Openaire, qui constitue l'infrastructure pour le dépôt des publications et des données mis en place au niveau européen. Openaire est partenaire de l'activité de LIBER.

LIBER participe à la seconde phase d'EUROPEANA ("Libraries" et "newspapers"). Cette participation est la première de l'association à propos des contenus.

Un article de LIBER[12] fait le point sur les rapports entre e-sciences et bibliothèques : « Research libraries' engagement with RIs has been low. While this could be understandable in 2005 when the first priorities for RI investments were defined, it now represents a big gap

[83] http://www.libereurope.eu/news/liber-response-to-european-commission-communication-and-recommendation
[84] http://www.peerproject.eu/

in the European strategy. Key initiatives such as the ESFRI Research Infrastructures involve no participation by research libraries, except for DARIAH." (p. 314)
A ce constat, quelques lignes plus loin, répond par ce qui constitue le nouveau projet de l'association : "Research libraries need to become visible actors in strategic discussions on RIs and should actively explore their engagement in research data infrastructures. Open Access, open science (data), research data infrastructures and management are the catalysts to get research libraries back into the awareness of researchers beyond the humanities and social sciences" (p. 314)

La question de l'intégration des bibliothèques numériques à l'intérieur des protocoles et infrastructures d'e-science constitue un point essentiel de la ROADMAP 2020[85] (p.8 : « Digital Data Libraries: Increasingly, the volume of data produced by high-throughput instruments and simulations is so large, and the application programs so complex, that it is cheaper to move the user's programs to the data than the data to the programs -that is, to keep most everything in a central library, while only the scientist's questions and answers move back and forth across the network. These central service stations are called Digital Data Libraries, or Science Data Centres. Each of these libraries holds one or more massive datasets, manages the programs that provide access to the data, and provides staff who understand, add to and improve the data. DDLs can come in many flavours. **There are Research DDLs** that hold the direct output of one or more focused research projects - typically, data without much processing. An example might be a DDL for environmental data on the Waddenzee in the Netherlands or Lake Como in Italy. **Discipline or Community DDLs** serve a larger group of scientists or engineers for instance, providing a pan-European library for data on all major bodies of water. **Reference DDL**s are broader still, serving large segments of science or education - scientists, students, educators and others from a variety of disciplines, institutions and places; an example might be a comprehensive European environmental data centre. Finally, **Specialised Service DDLs** go beyond data to provide services specific software tools to map all that aquatic and environmental data from across Europe, or look for correlations among the databases.
Digital Data Archives: As the name suggests, archives keep data and methods for the longer term, rather than active use. They can be for future reference or regulatory compliance; and they are more important in some fields than in others – for instance, a good chronological record of weather readings is vital for climatology. Indeed, the whole concept of digital data archiving is a modern development in the history of science, made possible by advances in information technology.
Digital Research Libraries: These store the results of research, printed or not, in electronic format and organise them for long-term use -like a modern university library. Their mission is to acquire information, organise it, make it available and preserve it. And they require a long-termcommitment -in financing, institutional backing and organisation. »

Les bibliothèques se retrouvent ainsi devoir à la fois participer aux dimensions de stockage et de référencement d'un nombre de plus en plus important de données, de s'assurer de leur pérennité et par ailleurs de devenir le point de contact des chercheurs pour le management de leurs données au niveau institutionnel pour toutes les disciplines académiques. De ce fait, les bibliothèques scientifiques assurent le lien entre les infrastructures et les utilisateurs de cette information.
Si l'on observe l'offre des infrastructures, on s'aperçoit soit qu'elle concerne des outils (lexiques, ontologies, etc.) mais qu'elle se préoccupe peu de les structurer (notamment dans le

[85] http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf),

cas de CLARIN), soit qu'elle repose sur des structurations de connaissances préexistantes comme dans le cas d'EUROPEANA. Dès lors que l'on envisage une perspective d'e-science, qui mettrait en relation l'ensemble des ressources pour la recherche, une structuration classificatrice et descriptive, dynamique dans le sens d'une mise en relation de données hétérogènes contribuant à la menée d'un travail de recherche, se posent à la fois des questions de structuration et de description documentaire qui sont le domaine de compétence des bibliothèques. Cette opportunité se double d'une autre, qui concerne le service que peut rendre la bibliothèque à ses utilisateurs. Comme on l'a vu, ce service ne concerne pas seulement l'accès à l'information, mais la mise en forme, la valorisation et l'archivage de l'information. En ce sens, le rôle de médiation assuré par les bibliothèques pourrait être essentiel dans un avenir proche. A titre indicatif, il constitue un aspect fondamental de notre projet et de notre collaboration avec les équipes d'informaticiens.

En ce sens, comme dans les expériences précédentes, l'e-science constitue un enjeu que s'approprient ces bibliothèques[86].
Ainsi, on voit apparaître dans le cadre de LIBER Quaterly des articles relatifs à des actions de recherche collaboratives associant des bibliothécaires[13].

En définitive, la convergence actuelle que l'on observe amoindrit la séparation que l'on avait observée au début de ce travail entre bibliothèques, bibliothèques numériques et perspectives d'e-science. Le niveau européen, apparaît ainsi comme l'espace d'une dynamique associant les différents acteurs de l'information scientifique.

1.4. Politiques publiques et initiatives non gouvernementales.

Nous avons largement traité des politiques publiques liées aux bibliothèques, aux chercheurs et à la structuration du web de données. Nous avons donc occulté une catégorie d'acteurs essentiels : les intérêts privés et leur réponse face aux stratégies des acteurs publics, envisageant de structurer un web gratuit et interopérable.

Les intérêts privés ne sont pas systématiquement contraires aux intérêts publics. Nous l'avons vu dans le cadre du projet PEER, à propos des éditeurs. Les questions d'équilibre de la recherche ont joué un rôle essentiel dans la coopération qui s'est engagée à propos de l'édition en ligne. La "voie verte" constitue le résultat d'un consensus entre les intérêts de chacun et permet la préservation du système éditorial sans nuire au programme d'open data.
Si cet accord est d'abord une question économique et organisationnelle (au sens de "l'écologie " de la recherche), elle aura quelques impacts sur la façon dont les documents sont décrits et sur le rôle de chacune des versions dans le cadre d'une recherche d'information.

Une complémentarité apparaît également au niveau de la recherche d'information : les applications spécialisées de GOOGLE (SCHOLAR GOOGLE ou ACTUALITES notamment) requièrent l'usage de métadonnées descriptives des contenus. La description des contenus par des métadonnées constitue une façon extrêmement rapide et précise de répondre à une question sans avoir à entrer dans les contenus du document.
En ce sens, il y a complémentarité entre les intérêts des acteurs élaborant des moteurs (et les services qui les accompagnent) et ceux des bibliothèques numériques.
Cette complémentarité ne doit néanmoins pas faire oublier les divergences fondamentales en termes de politiques, entre les deux types d'acteurs, et qui concernent les collections de

[86] http://www.slideshare.net/libereurope/escienceroledesbibliothequesderecherche

documents numérisés. Néanmoins, c'est bien GOOGLE qui, s'inspirant de DBPedia, met en place un "graphe de connaissances" [87]. Les métadonnées de SCHEMA.ORG [88] reprenant les IPTC ne sont pas le seul exemple de l'intérêt du moteur de recherches pour les propositions du web de données.
Au travers du wiki du W3C sur les schémas[89] , les responsables de la recherche de GOOGLE (R.V.Guha et D. Brickley) proposent des extensions des métadonnées de SCHEMA.ORG (qui elles-mêmes sont issues des IPTC) vers le Dublin Core, EUROPEANA, les métadonnées de la DPLA (Digital Public Library of America) ou encore LOD-LAM.

Ces convergences cachent aussi des enjeux très différents entre ces acteurs : ce qui importe pour les acteurs publics en Europe, notamment gouvernementaux, c'est de mettre le web au service de l'activité économique (et notamment de la recherche). Pour cela, un certain nombre d'outils de structuration sont développés, ces outils étant évidemment publics et pouvant être utilisés par n'importe quelle organisation. Il va de soi que les ignorer ne peut être que préjudiciable pour une entreprise privée qui par ailleurs, en se les appropriant, peut diversifier son offre de produits internet et rester concurrentielle sur le marché des moteurs de recherche généralistes (ASK, BING, YAHOO notamment sont des concurrents sérieux). L'ensemble des schémas en cours d'élaboration ou publiés sont répertoriés[90].

Pour les acteurs privés, la réalité la plus tangible du web de données et de l'open access est celle des LINKED DATA. Il ne s'agit pas de créer une infrastructure au sens où les politiques publiques peuvent le faire, mais simplement un réseau d'outils interopérables et complémentaires. Les LINKED DATA constituent un produit du fonctionnement collégial du W3C. Il s'agit de propositions d'écriture (quatre au total) qui font qu'un outil puisse être reconnu comme faisant partie des LINKED DATA[91] . Les LINKED DATA constituent le versant accessible à tous des projets développés dans les infrastructures : les outils sont disponibles et gratuits, mais ils sont issus de politiques publiques, notamment de développement d'infrastructures.
Si aujourd'hui le graphe représentant l'ensemble des outils connectés est difficile à observer tant il est complexe du fait du très grand nombre de ses objets, par contre, si l'on regarde simplement au niveau des ressources linguistiques[92] , on s'aperçoit que des outils ayant des origines très différentes (comme par exemple les métadonnées IMDI et les représentations lexicales de Wordnet) peuvent être mises en relation parce qu'elles utilisent des langages de représentation interopérables. En ce sens, les LINKED DATA ne sont pas simplement un répertoire d'outils. Elles peuvent être considérées comme une boîte à outil. Les capacités d'innovation associées aux LINKED DATA sont accentuées par le partage des données et des règles.

Les LINKED DATA constituent les fondements pour l'open access ; bien que les LINKED DATA constituent des règles d'écriture et l'open access une attitude par rapport à des données, le lien entre les deux est important : des données ne peuvent être effectivement accessibles que si les outils qui les structurent sont eux aussi transparents. Par conséquent, même si les acteurs sont différents, on assiste à la convergence de deux mouvements à l'intérieur du web.

[87] http://newsbreaks.infotoday.com/NewsBreaks/Google-Unveils-Knowledge-Graph-82816.asp
[88] http://schema.org/docs/schemas.html
[89] http://www.w3.org/wiki/WebSchemas
[90] http://www.w3.org/wiki/WebSchemas/SchemaDotOrgProposals
[91] http://www.w3.org/DesignIssues/LinkedData.html
[92] http://nlp2rdf.lod2.eu/OWLG/llod/llod.svg

Comme nous l'avons déjà présenté, les politiques publiques européennes se traduisent dans la mise en place d'infrastructures qui mobilisent, au travers d'enjeux scientifiques et techniques comme la normalisation, l'interopérabilité et la fédération des ressources, des acteurs publics et privés. Le cas des ressources linguistiques est à ce titre exemplaire : dans le réseau réuni autour des workshops LREC, on retrouve les agences privées comme les ELRA/ELDA, les consortiums liant des acteurs publics et privés comme METASHARE et des organismes de recherche comme le MaxPlanck Institute de Niemigen.
Le travail en commun, le partage des données et la production de services transforment considérablement les règles du rapport entre entreprise privée et organisme public ; par exemple, l'édition de corpus payants constitue une position de moins en moins tenable (ELRA a ainsi dû rendre libre d'accès le 23 mai 2012 un nombre important de ses ressources[93]).
L'entreprise a intérêt à participer à ce mouvement collaboratif quitte à partager ses produits. Rester dans une position de rétention empêche d'accéder à l'innovation et au développement.

La Commission Européenne, au travers de la directive HORIZON 2020[94], développe une politique volontariste relativement à l'accès libre : « To make progress in science, we need to be open and share. (…) But sharing data, and having the forum to openly use and build on what is shared, are essential to science. They fuel the progress and practice of scientific discovery. That's why scientists have long sought out new tools and new ways to share their knowledge". Kroes, EC Vice Présidente de l'UE responsable de l'agenda numérique lors du workshop de Rome « l'Infrastructures for Open Science", 10-12 April 2012[95].

La réaffirmation de cette volonté est justifiée par la perspective d'un développement centré sur la recherche et l'innovation, et dont les moteurs essentiels sont le partage des données et le travail collaboratif.

Recommandations (de type "linked data") et open access ont en commun entre autre une propriété très importante. Il s'agit de la réutilisation. Chaque schéma créé est destiné à être réutilisable, et ainsi contribue à la structuration du web. Cette réutilisation constitue un trait fondamental à la fois de la structuration du web mais aussi de l'activité professionnelle, et plus particulièrement scientifique, qui l'accompagne.
L'interopérabilité et la disponibilité évitent la création de produits ou services aux fonctionnalités identiques. Il suffit qu'une équipe crée le produit ou le service et que d'autres utilisateurs puissent l'améliorer et l'adapter à des fonctionnalités plus particulières. Le gain pour la gestion de la recherche n'en est que plus impressionnant.

Si ces principes sont ceux qui président au logiciel libre, ce qui est différent, c'est la façon dont cette offre est structurée : les LINKED DATA regroupent des outils et des services répondant à des critères très précis, internes au web, définis par le W3C. Cette contrainte n'empêche pas une structuration autre, permettant l'émergence des outils et des services : il s'agit de l'OPEN KNOWLEDGE[96] , une communauté dont l'objectif consiste justement à promouvoir les projets en open access au travers d'un réseau de compétences permettant de faire évoluer et les produits[97]. L'OPEN KNOWLEDGE est une fondation et fonctionne par chapitres et centres d'intérêt. Ces derniers concernent à la fois la science, la culture et le

[93] http://www.elra.info/Free-LRs.html
[94] (voir
http://ec.europa.eu/research/infrastructures/pdf/icri_conclusions%20_final.pdf#view=fit&pagemode=none
[95] http://europa.eu/rapid/press-release_SPEECH-12-258_en.htm?locale=en
[96] http://okfn.org/about/
[97] http://wiki.okfn.org/Handbook/Governance

gouvernement. Dans ces centres d'intérêt, on trouve des groupes de travail, comme par exemple celui sur les ressources linguistiques[98] ou les métadonnées bibliographiques[99]. Nous reviendrons plus loin sur l'activité de ces différents groupes. Considérons simplement le fait que le modèle qui fonde l'open access s'étend depuis les pouvoirs publics, les organisations structurant le web, jusqu'aux communautés scientifiques. Ainsi, on dispose d'une synergie d'acteurs complémentaires qui permet de convaincre de l'intérêt de la démarche et donc l'adhésion des acteurs « privés ».

Les projets, largement publics, que nous venons de présenter, constituent une certaine forme d'appropriation des recommandations de formats et de l'organisation du web. Comme nous l'avons indiqué, notamment au travers des IPTC, d'autres formes, professionnelles et privées sont possibles. Elles sont bien entendues, ouvertes et compatibles avec les premières.
En définitive, si le W3C constitue le laboratoire produisant les langages du web, les LINKED DATA sont la boîte à outils dans laquelle se servent les différents acteurs qui se doivent de collaborer (de façon plus ou moins irénique) de façon à construire le web de données et à profiter des positions et des services des uns et des autres.

1.5. Les langages de structuration du web.

Nous mettons en perspective maintenant les recommandations du consortium W3C. Elles constituent les langages d'expression des outils que l'on a pu présenter et donc le fondement de l'interopérabilité. C'est ce soubassement qui permet aujourd'hui l'élaboration des infrastructures ouvertes, non seulement techniquement mais également stratégiquement : on ne peut atteindre les effets escomptés (résultats, produits, services) qu'à partir du moment où les services, les outils, les documents proposés sont à accès libre.

Les nouvelles propositions de standards intègrent de plus en plus des propriétés relationnelles : les recommandations évoluent depuis la structuration et la publication des données vers la mise en relation de celles-ci. Notons enfin que ces langages effectuent une trajectoire depuis la syntaxe vers la sémantique et les connaissances. Enfin, cette mise en relation, prometteuse de nouveaux produits et services, repose sur l'open access.
Les recommandations qui suivent constituent les conditions des projets que nous venons de présenter. Nous les présentons après parce que ce qui nous intéresse d'abord est la façon dont elles sont appréhendées et adoptées par les sociétés, notamment au travers des bibliothèques numériques et l'insertion de ces dernières dans un projet encore plus vaste, qui est celui de l'intégration du travail scientifique sur le web.

Cette évolution se comprend aisément : les premières recommandations visaient essentiellement à structurer les pages web de façon à les rendre lisibles par les moteurs de recherche et par tout utilisateur. Elles sont destinées à être utilisables par des personnes qui ne sont pas directement des experts de l'informatique (HTML, SGML). Leur adoption a permis l'explosion du nombre de documents présents sur le web et leur référencement.

Par ailleurs, le W3C a cherché à intégrer au web d'autres outils numériques et d'abord les Bases de Données. Un certain nombre de protocoles comme MySQL ont été élaboré. Cette intégration a permis d'élargir le champ d'intervention du web. L'objectif est que l'ensemble des objets numériques puissent être accessibles sur le web. Cette prise en compte a permis

[98] http://linguistics.okfn.org/
[99] http://openbiblio.net/

l'émergence et l'explosion du web commercial et du web de services.
XML est un langage de représentation dévolu à la description des données sur le web. XML est d'abord une syntaxe permettant l'expression de la prédication.

1.5.1. RDF et SKOS : des langages de représentation pour structurer des relations.

RDF[100] ne constitue pas un langage de représentation mais un cadre de travail permettant de relier des données. Il s'agit donc d'un modèle de données qui ordonne de façon logique des ensembles de connaissances. En effet, on ne peut toujours pas entrer dans les contenus des documents. Par contre, on peut donner des informations sur un document en le reliant à un autre, ou à une entité qui donne du sens à ce document.

RDF est maintenant structuré à l'aide d'une syntaxe particulière, TURTLE[101]. Aujourd'hui, TURTLE passe du statut de soumission à celui de dernière version provisoire avant validation définitive comme standard[102] . TURTLE évite la complexité des formulations XML et ainsi de retrouver une fluidité proche du langage naturel.

Les relations que propose RDF sont améliorées par RDFS qui propose quelques indications de relations : RDF caractérise des classes, des sous-classes et des propriétés. Ces schémas permettant de systématiser les relations entre les données qui sont proposées par les prédications RDF.
Les schémas se définissent par la typification des descriptions RDF. Ainsi, on dispose grâce au schéma d'une structure type réutilisable sur l'ensemble des instances acceptées par ce type.
OWL vise à intégrer dans le web les ontologies, à savoir les outils de structuration des connaissances. Par rapport aux schémas RDFS, OWL précise les relations de classes. Une telle perspective est également celle de SKOS, qui constitue une structure à la fois d'édition et de mise en relation de systèmes d'organisation de connaissances.
Les schémas sont reconnus à l'aide de requêtes SPARQL, ce qui permet de faire gérer directement la recherche d'information par les seuls outils web de données.

Les schémas RDFS connaissent un succès grandissant, surtout ceux qui sont répertoriés et reconnus par les moteurs de recherche. Des bibliothèques comme SCHEMA.ORG[103] proposent des schémas pour des situations fréquentes du web. Leur intérêt réside dans leur reconnaissance par les moteurs de recherche du web (Bing, Google, Yahoo! Yandex notamment). Ainsi, il est possible de faire reconnaître par les moteurs de recherche « traditionnels » des structures sémantiques. Utilisable depuis septembre 2011, cet outil permet de faciliter la reconnaissance autant de types d'entités que de scènes (événements divers). La présentation RDF concerne des frames, qui peuvent être plus ou moins spécialisés. Ainsi, on peut caractériser des événements et même les spécialiser (sport, vente, etc.).
Cette stratégie d'intégration des outils du web de données dans l'actualité est également celle de rNews[104], qui constitue le groupement de recherche des IPTC. RNews est un outil fondé sur des schémas RDFS qui permet d'insérer des métadonnées dans des pages HTML de façon à ce que leur contenu soit immédiatement reconnu et donc utilisé pour des recherches

[100] (pour une introduction : http://www.rdfabout.com/intro/#Comparing%20RDF%20with%20XML)
[101] http://www.w3.org/TeamSubmission/turtle/http://www.w3.org/TeamSubmission/turtle/.
[102] http://www.w3.org/TR/turtle/
[103] http://schema.org/
[104] http://dev.iptc.org/rNews

d'information immédiate. RNews constitue un outil relativement original puisqu'il réutilise des attributs de métadonnées, mais qu'il construit également des classes RDF et enfin qu'il utilise des vocabulaires contrôlés pour spécifier les valeurs considérées.
On peut ainsi noter que l'évolution des outils du web ne suit pas une trajectoire linéaire qui irait de HTML vers les connaissances. Au contraire, notamment pour les domaines de l'information d'actualité, le web de données s'insère à l'intérieur des pages HTML. De cette façon, les schémas RDFS peuvent être utilisés comme des tags, et rejoignent les problématiques de l'annotation de documents.
On assiste dès lors à une multiplication des schémas. Comme un schéma répond à des besoins liés à une activité précise, il n'est pas possible de limiter cette prolifération. Par contre, les lexiques, limités aux possibilités de la langue naturelle, constituent les bases fédératives de l'ensemble du web de données[14].

SKOS permet de mettre en relations différents points de vue sur un domaine et ainsi de permettre une couverture plus importante. En effet, il s'agit par SKOS d'établir des relations entre des structures de connaissances établies (comme par exemple des thésaurus). Le premier objectif, outre l'édition d'organisations de connaissances, est leur mise en relation, notamment dans le cas expressément prévu, des relations entre structures de langues différentes. Le second consiste à donner un cadre à l'alignement d'outils structurés.
Par ailleurs, la liste des institutions ayant adopté une représentation par SKOS est impressionnante[105]. Elle montre surtout que des référentiels ayant des usages extrêmement différents puissent être organisés selon un même format. Ainsi un jeu de métadonnées comme les IPTC et la classification DEWEY n'ont pas grand-chose en commun, si ce n'est le fait qu'elles se doivent de caractériser leur structure, vocabulaires contrôlés et champs, comme des connaissances. Des outils différents disposent ainsi d'un même format de représentation ; à la fois il permet une éventuelle utilisation commune sans que par ailleurs la pertinence locale de chacune ne soit altérée. Le langage d'expression est indépendant de ce qu'exprime la structuration comme des usages.

Le succès très rapide de SKOS peut être expliqué également par son double niveau d'organisation : d'une part les concepts, d'autre part les lexiques. Cette propriété permet une utilisation de SKOS pour lier des structures de différents niveaux d'abstraction : lexiques et thésaurus, taxonomies et ontologies.
Par ailleurs, la complexité des thésaurus et des relations hiérarchiques entraîne la création d'extensions à SKOS, comme par exemple les extensions proposées par FinnONTO[106]

Il nous faut revenir sur un impératif qui est celui de l'élaboration d'outils de recherche d'information qui exploiterait les descriptions fournies par les expressions RDF, SKOS et OWL. SPARQL offre cette possibilité-là. L'usage de SPARQL est guidé par le principe de la recherche d'information fondée sur le contenu sémantique[107].

1.5.2. Remarques à propos des recommandations du W3C et de la signification.

Les langages de représentation du web prennent de plus en plus en compte la signification, entendue au sens de la prédication. Il ne s'agit pas d'une caractérisation linguistique mais

[105] voir http://www.w3.org/2001/sw/wiki/SKOS/Datasets
[106][106] http://schema.onki.fi/skosext/.
[107] http://www.w3.org/DesignIssues/Metadata.html, http://www.w3.org/DesignIssues/Semantic.html

logique, au sens des logiques prédicatives. Ce principe est rappelé en introduction de la recommandation RDF et spécifié dans l'ensemble des documents explicatifs[108] : "RDF is based on the idea that the things being described have properties which have values, and that resources can be described by making statements, similar to those above, that specify those properties and values. RDF uses a particular terminology for talking about the various parts of statements. Specifically, the part that identifies the thing the statement is about (the Web page in this example) is called the subject. The part that identifies the property or characteristic of the subject that the statement specifies (creator, creation-date, or language in these examples) is called the predicate, and the part that identifies the value of that property is called the object ».

Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié

.

En ce sens, ces formats peuvent servir à représenter l'ensemble des raisonnements fondés sur des argumentations (rapports prédicat-argument)[109] : « The semantics given here restricts itself to a formal notion of meaning which could be characterized as the part that is common to all other accounts of meaning, and can be captured in mechanical inference rules". Leur intérêt réside dans la pluralité des niveaux d'expressions : concepts, classes, unités lexicales. En ce sens, ces recommandations permettent de gérer différents niveaux d'abstraction et donc de modalités de recherche.

Code de champ modifié
Code de champ modifié

Ce sont également des recommandations, ce qui signifie qu'elles peuvent être utilisées pour finaliser des résultats de recherches et les rendre utilisables. Par exemple, un raisonnement peut être caractérisé dans un langage logique précis. Il n'est pas immédiatement utilisable en dehors de son cadre logique natif tant qu'il n'a pas été exprimé dans les schémas RDFS qui permettent de rendre utilisable ce qu'il propose. Leur portée est extrêmement importante parce qu'ils permettent, par la publication des schémas, de rendre immédiatement disponible et utilisable le résultat d'une recherche. En ce sens, il s'agit d'un changement essentiel dans les modalités de diffusion d'un produit scientifique (et sans parler des publications évidemment).

Nous n'avons pas insisté sur la portée de l'un des aspects de cette mise en relation : il s'agit des opérations d'alignement, de matching et de fusion d'ontologies ou de thésaurus. (L'alignement consiste à relier deux structures par des liens entre les concepts, le matching à associer des relations structurantes entre ces deux structures, et la fusion à ne construire plus qu'une seule structure à partir des deux). En effet, toute écriture d'une ontologie ou d'un thésaurus suivant les recommandations du W3C permet de réaliser certaines opérations des unes par rapport aux autres. Ainsi, il est possible dès lors de mettre en relation différents points de vue sur un domaine. Il est également possible, grâce au projet DATALIFT[110] , de convertir toute donnée brute dans le format RDF, en utilisant une ontologie appropriée. Le processus consiste ainsi aussi à mettre en relation systématiquement ces données décrites.

Enfin, un aspect fondamental du travail du W3C est la façon de travailler, notamment par le biais des incubateurs. Le fondement du travail est collaboratif, ouvert et repose sur des finalités clairement exprimées. A titre d'exemple, on peut citer le Library Linked Data Incubator Group[111] qui a permis de faire travailler pendant un an des experts autour de la mise en relation des outils des bibliothécaires et des LINKED DATA. Les conclusions de ce groupe seront importantes pour la partie suivante de notre travail.

[108] Voir notamment http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.
[109] http://www.w3.org/TR/2004/REC-rdf-mt-20040210/
[110] http://datalift.org/wiki/index.php/Main_Page, http://datalift.org/fr/node/11
[111] http://www.w3.org/2005/Incubator/lld/

1.6. Outils documentaires. Montée en puissance des métadonnées.

Nous traitons maintenant des outils, notamment documentaires, qui constituent le lien intermédiaire entre les formats de représentation et les moteurs de recherche. Ces outils sont pour la plupart très largement antérieurs au web. Ils voient leur pertinence s'accroître par le web, tout simplement parce que leur rôle dans la recherche d'information devient fondamental. Nous traiterons ici essentiellement des métadonnées dans la mesure où cette expansion des notices bibliographiques répond à l'ensemble des besoins de médiation. Nous présentons ici les évolutions les plus générales des métadonnées, sachant que nous reviendrons sur les propriétés de certains jeux.

Considérées au départ comme un outil de recherche d'information concurrent et d'un usage moins facile que le référencement des sites par les moteurs de recherche, les métadonnées sont néanmoins beaucoup plus précises pour l'identification des documents. Par conséquent, les moteurs intègrent les métadonnées de façon de plus en plus accentuée. (Voir notamment la carte GOOGLE).

Issues des notices bibliographiques, les métadonnées étaient relativement éloignées des questions linguistiques. Le choix qui avait été fait d'utiliser des langages contrôlés, notamment dans le cadre du Dublin Core mais aussi des IPTC et du MPEG 7, restreint les possibilités d'expression. Par exemple, dans le cadre des métadonnées des ressources linguistiques, le souci de précision tout autant que du fait de la diversité des usages ont abouti à l'utilisation de terminologies de référence ; ainsi, ISO-cat sert de fondement lexical aux métadonnées IMDI[15]). Par ailleurs, l'émergence de schémas tels que ceux qui sont présentés dans SCHEMA.ORG va également dans le sens d'une substitution des langages contrôlés par des catégories linguistiques et des terminologies plus précises et proches du langage naturel.

Ce lien traduit une évolution du rôle des métadonnées. Parce que les initiateurs des bibliothèques numériques tout autant que les personnes chargées de renseigner ces publications ne sont pas systématiquement des bibliothécaires mais des experts du domaine, les considérations lexicales et terminologiques sont devenues fondamentales. Parallèlement, elles se doivent d'être plus proches des requêtes formulées par des utilisateurs.

Les LINKED DATA, et l'ensemble des outils linguistiques proposés, permettent de réduire la distance entre les vocabulaires de l'utilisateur et celui reconnu par les machines.

1.6.1. Fondements des métadonnées.

On distingue traditionnellement les métadonnées informatiques, associées à tout document et comprises dans le format des données, des métadonnées professionnelles, qui relèvent de jeux élaborés par une profession et qui sont renseignés dans un but de description documentaire.

A l'origine, les métadonnées sont fondamentalement associées à l'identification des documents électroniques. Leur véritable expansion sera leur assimilation à un outil de description de documents par l'intégration d'une sémantique et d'éléments de description de contenu. Le Dublin Core[112] est le premier jeu à articuler une logique de la description formelle du document à celle de ses contenus. Les métadonnées informatiques étaient considérées comme minimales et sans portée descriptive. Le Dublin Core va articuler l'identification des documents à leur description. Néanmoins, la limite à cette description est le fait que les contenus sont uniquement un discours porté par un ou des expert(s) sur ces contenus.

[112] http://dublincore.org/

Les métadonnées sont des outils de description des documents interopérables, complémentaires et d'un usage facile. Leur difficulté réside dans le fait qu'elles ne sont pas automatisées et donc qu'elles impliquent toujours l'intervention humaine. Une partie de notre projet vise à répondre à ce type de problème, en proposant quelques éléments pour une automatisation du renseignement de ces métadonnées.

Les métadonnées partent du constat qu'il est impossible d'entrer dans le contenu des documents et donc que la façon la plus pertinente de les décrire est de s'intéresser à leur contexte. La description la plus large de ce contexte permet au mieux de cerner l'identité de ce document.

Les métadonnées, comme la plupart des outils du web, ont une histoire antérieure à lui. Ainsi, les représentations RDF ou SKOS sont des versions web d'outils élaborés antérieurement. La représentation RDF est suffisamment souple pour exprimer des fonctionnalités et des systèmes de pensée très différents. L'évolution des langages de représentation permet aux outils de perdurer voire même de trouver une plus large audience. Elle permet également à de nouveaux outils d'apparaitre avec une grande facilité.
Cette standardisation des représentations permet également un décloisonnement des usages. L'identification des schémas sur le nuage des LINKED DATA permet à n'importe qui de les utiliser partiellement ou totalement, ou encore à les intégrer dans d'autres jeux. Par exemple, un modèle comme FRBR conçu par l'IFLA pour caractériser un document en lien à l'ensemble de ses versions, des ressources liées à son auteur et son contexte, peut être utilisé pour représenter des objets numériques distribués, comme par exemple dans l'ontologie FABIO[113] .

1.6.2. Des formats MARC vers une multiplicité de l'offre.

Sans vouloir faire un retour au format MARC, qui constitue depuis 1963 un outil pour la description des documents, au sens d'une version électronique des notices bibliographiques, on pourra noter un mouvement des notices vers l'utilisation des formats du web (mise en place d'une DTD[114]) et le développement de dimensions conceptuelles associées aux notices bibliographiques. Ce dernier développement a donné lieu au modèle FRBR de notices bibliographiques[115] . Ce modèle, élaboré par l'IFLA, est en cours de redéploiement, pour deux raisons, selon la BNF :

1. d'une part l'élaboration d'un format international unique en lien avec RDA ("Ressources : description et accès").
2. d'autre part l'interopérabilité avec les formats de description des objets muséaux.

RDA est pensé dès l'origine pour la caractérisation des documents électroniques « *RDA: Resource Description and Access* [1] is in development as a new standard for resource description and access designed for the digital world. (…) A specific focus of RDA is the description of elements of the content and carrier of a resource that will help users to identify and select the resource to meet their needs with respect to the form of content, subject, volatility, etc., on the one hand, and the physical characteristics of the carrier, the formatting and encoding of the information, etc., on the other »[116].

[113] http://www.essepuntato.it/lode/http://purl.org/spar/fabio
[114] http://www.loc.gov/standards/marcxml/
[115] http://www.bnf.fr/fr/professionnels/modelisation_ontologies/a.modele_FRBR.html
[116] http://www.dlib.org/dlib/january07/dunsire/01dunsire.html

RDA constitue une norme succédant à l'Anglo-American Cataloguing Rules. C'est d'abord une initiative régionale qui vise maintenant à devenir interopérable avec les propositions de l'IFLA.

L'enjeu dans cette reformulation des principes du catalogage est l'évolution de la recherche d'information d'une part, celle des catalogues d'autre part. Une telle évolution était déjà sensible dans le MARC 21, qui introduit des éléments nouveaux de description des documents électroniques.

Depuis fort longtemps, les bibliothèques travaillent en réseau et les descriptions des documents sont depuis longtemps écrites sous forme électronique. Les formats METS développent le MARC sous forme XML[117]. Les formats METS ont comme intérêt de pouvoir intégrer des éléments très divers dans leur schéma, notamment des éléments du Dublin Core. Ils constituent des éléments essentiels à l'articulation des descriptions bibliographiques et des métadonnées.
Le format MARC est en cours d'évolution et se traduit par l'élaboration d'un modèle traduisant plus spécifiquement les réalités du web : il s'agit de BIBFRAME, projet piloté par la Bibliothèque du Congrès, et qui existe depuis novembre 2012 comme recommandation[118]
Son objectif est formulé ainsi (p.3) : "The new model is more than a mere replacement for the library community's current model/format, MARC. It is the foundation for the future of bibliographic description that happens on, in, and as part of the web and the networked world we live in. It is designed to integrate with and engage in the wider information community while also serving the very specific needs of its maintenance community - libraries and similar memory organizations. It »
BIBFRAME apparait ainsi très proche du modèle FRBR de l'IFLA. C'est un modèle relationnel représenté en utilisant RDF.

MODS (Metadata Object Description Schema) constitue un schéma de description bibliographique qui étend les possibilités de MARC 21. Il s'agit d'un schéma de description bibliographique plus riche que METS.
MADS/RDF (The Metadata Authority Description Schema) constitue un schéma pour la caractérisation des agents, événements, lieux : MADS est fondé sur des listes d'autorités et constitue un système d'organisation des connaissances.
MADS et MODS sont complémentaires au sens où MADS fournit à MODS des vocabulaires contrôlés adaptés.

Un des arguments de MADS/MODS par rapport à SKOS est la spécificité des besoins des bibliothèques (« Unlike SKOS, however, which is very broad in its application, MADS/RDF is designed specifically to support authority data as used by and needed in the LIS community and its technology systems »[119]) (. Il s'agit d'un système d'organisation des connaissances qui repose sur les listes d'autorité des bibliothèques. Ils rendent opérable sur le web les organisations de connaissances pré-coordonnées (comme les classifications) alors que SKOS est prévu pour les systèmes post-coordonnés comme les thésaurus.

Concernant plus particulièrement le document électronique, de nouvelles métadonnées sont en cours d'élaboration à la Bibliothèque du Congrès, et qui constituent des expansions de

[117] http://www.loc.gov/standards/mets/.
[118] http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf.
[119] http://www.loc.gov/standards/mads/rdf/

MODS[120].

Enfin toujours dans le cadre des formats de la bibliothèque du Congrès, SRU est un protocole de recherche d'information associé à CQL, qui constitue un langage d'interrogation contextuel. SRU/CQL est en cours de standardisation dans OASIS, un service de recherche web (ou SWS)[121].

En définitive, les initiatives de la bibliothèque du Congrès ont comme objectif de déplacer sur le web et d'adapter les systèmes classificatoires et de description documentaire mis en place dans le cadre des bibliothèques.
Elles permettent de structurer l'ensemble de la chaîne depuis la description des documents jusqu'à la recherche d'information.

Dans le cadre des infrastructures de recherche, le jeu de métadonnées VIVO[122] permet d'enregistrer un ensemble de renseignements importants relatif à une publication, un projet, et surtout, pour les auteurs, éditeurs et lieux de publication. Il s'agit ainsi de métadonnées relatives à la description des communautés scientifiques : « VIVO supports direct data entry. Research data records can enter in directly into VIVO, and then linked via relational statements to existing people, grant, and organizational statements. Using semantic structures, records about people, grants, org units etc., can also be harmonized at a later date if required."

Néanmoins, les métadonnées de type VIVO ne sont pas associées à une recherche d'information de type documentaire ; elles sont limitées à une lecture par les moteurs de recherche classiques du web.

On ne peut s'empêcher de parler du Dublin Core, qui constitue le premier jeu de métadonnées pour le document électronique. Son succès est lié à sa simplicité et à l'absence de recours à des listes et classifications structurées. En d'autres termes, son indépendance par rapport au langage des bibliothèques. Les travaux actuels, tels que l'on peut les saisir au travers des conférences annuelles (La Haye 2011, Kuching 2012, Lisbonne 2013) traitent à la fois de l'interopérabilité et du lien aux LINKED DATA[123]. .
Les autres conférences accentuent cette dimension, qui de plus en plus associe le DC à d'autres outils, de façon à proposer une description relationnelle des documents sans pour autant remettre en question le modèle central. Les solutions proposées par le DC pour construire une description relationnelle comportent un lien aux modèles lexicaux et terminologiques de représentation des domaines qu'aux ontologies, considérées comme trop restrictives.

Ainsi, l'apparence de la multiplicité des initiatives cache en fait une forte différenciation des objectifs. Par ailleurs, nous n'avons pas traité des spécificités des métadonnées des archives (EAD) et des musées.
Si les différences entre les jeux de métadonnées semblent limitées, elles n'en restent pas moins liées aux objets décrits et aux langages contrôlés utilisés.

Comme on peut s'en apercevoir à la lecture de cette évolution, les consortiums responsables

[120][120] http://www.loc.gov/standards/mdc/index.html
[121] http://www.loc.gov/standards/sru/oasis/
[122] http://blogs.unimelb.edu.au/vivoands/about/
[123][123] http://dcevents.dublincore.org/index.php/IntConf/dc-2011/schedConf/presentations

des jeux ont systématiquement adopté les formats proposés par le W3C. Néanmoins, la portée de cette traduction n'a pas été totalement intégrée dans sa dimension sémantique, au moins dans un premier temps. Ce n'est qu'aujourd'hui, avec les projets LINKED DATA du DC ou l'apparition sur marché de la Bibliographic Ontology[124], que s'est faite la prise en compte de la dimension relationnelle du RDF.

Néanmoins, ces possibilités de relations ne concernent pas seulement les descriptions de documents mais les relations que ces dernières peuvent établir avec les langages documentaires. En ce sens, les effets de la publication de SKOS sont encore loin d'être exploités.

SKOS est un outil web d'édition et de mapping de données structurées et d'abord de thésaurus. L'objectif d'une interopérabilité entre différents thésaurus, contenu dans SKOS, transforme considérablement l'accès à l'information et surtout la mise en commun des ressources documentaires.

Les métadonnées constituaient au départ des adaptations des notices bibliographiques à la disponibilité des documents numériques. Avec l'intégration d'autres outils documentaires, notamment ceux de structuration des connaissances comme les thésaurus, à l'intérieur du web, on assiste non seulement à un « passage » au web, mais véritablement à la fertilisation des fonctionnalités des outils documentaires et des propriétés du web. D'un côté, le web accède aux connaissances. D'un autre côté, les outils documentaires peuvent être fédérés et fonctionner de façon complémentaire, notamment pour la recherche d'information. Par exemple, les capacités d'alignement de thésaurus permettent de mettre en réseau différents outils. De cette façon, en lien à l'articulation institutionnelle des institutions des bibliothèques et du W3C, on assiste à la mutation des outils proprement documentaires par leur intégration des propriétés du web[16].

Cette intégration ne modifie pas les fonctionnalités, mais permet déjà au web d'entrer de plein pied dans la question des connaissances. Par ailleurs, les outils de structuration de connaissances sont de fait mis en relation avec ceux qui assurent la représentation des connaissances. Une telle connexion a déjà été réalisée par Unified Medical Language System[125]. Elle assure la complémentarité entre des réseaux sémantiques et des thésaurus dans le domaine médical.

Ainsi, on voit poindre la possibilité d'une interopérabilité généralisée, qui permettrait d'enrichir considérablement les relations déjà établies, dans le cadre du DC notamment, entre certains éléments et les langages contrôlés les plus utilisés : DDC (classification Dewey), IMT (ensemble des types de médias répertoriés par la Internet Assigned Numbers Authority), LCC (Library of Congress Classification), LCSH (Library of Congress Subject Headings), MeSH (Medical Subject Headings), NLM (National Library of Medicine Classification), TGN (Getty Thesaurus of Geographic Names) et enfin l'UDC (Universal Decimal Classification).
Simplement, ces ressources sont considérées uniquement au titre de vocabulaire pour les schémas d'encodage. Autrement dit, il s'agit de restreindre et de normaliser les choix de valeurs associés certains éléments.
Ces choix sont effectivement restreints par rapport à l'ensemble des possibilités offertes par les langages de représentation.

[124] http://bibliontology.com/
[125] http://www.nlm.nih.gov/research/umls/

En définitive, les métadonnées évoluent très vite vers une dimension relationnelle de plus en plus marquée, liée à la fois à la multiplication des versions d'un document et à la publication sur le web d'objets divers en lien à ces documents. On remarquera que cette mise en relation insérée dans la description des documents ne comporte pas d'enrichissement particulier de la description des contenus. Elle se marque essentiellement comme un accroissement de la qualité de description du contexte.

1.6.3. Métadonnées professionnelles ou liées au support.

L'adoption généralisée des recommandations de représentation et la diversification des documents publiés sur le web entraînent la multiplication des jeux et leur spécialisation. On étend maintenant le champ pour passer des bibliothèques généralistes vers les centres spécialisés et leur utilisation de métadonnées spécifiques. Cette diversification est également accentuée par l'évolution des formats de représentation des documents audiovisuels et multimédias, et enfin l'élaboration de jeux empruntant à plusieurs formats.

Certaines de ces métadonnées sont relativement anciennes et sont liées à la circulation mondiale des informations de presse comme les IPTC[126] . Les IPTC sont plus complexes que le Dublin Core dans la mesure où la circulation internationale des nouvelles et de leurs images requièrent de façon immédiate certaines informations sur leur statut, leurs conditions de vente, etc.
Les IPTC sont constituées de champs et de vocabulaires contrôlés visant à décrire les informations d'actualités et leurs images. Comme nous l'avons vu, elles fournissent la base de schémas caractérisant les textes d'actualité qui permettent d'accélérer la recherche d'actualités.

Néanmoins, à partir du moment où on n'est plus dans le cadre très structuré des bibliothèques et de leurs associations mais dans celui, concurrentiel, des médias (presse écrite, télévision, actualités web), les métadonnées constituent aussi des enjeux économiques. En témoigne la mise en œuvre, par la BBC, d'un standard pour la description des événements sportifs concurrent des IPTC[127]
Si ces différents acteurs travaillent de concert dans le cadre du W3C[128] , il n'en reste pas moins que chacun se différencie par une caractérisation différente de l'événement sportif : soit il s'agit d'un résultat, comme dans le cas des propositions de la chaîne ESPN et GOOGLE[129] , soit une structure de relations conceptuelles, comme dans le cas de l'ontologie de la BBC[130] , élaborée notamment pour les JO de Londres en 2013. Cette dernière se fonde sur la notion d'événement inséré à l'intérieur de protocoles et de manifestations réguliers.

Néanmoins, cette multiplication des jeux de métadonnées spécialisés est également observable à l'intérieur des domaines scientifiques, en lien en grande partie aux données primaires utilisées. On doit mentionner les DDI[131]: « The **Data Documentation Initiative (DDI)** is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. »

[126] http://www.iptc.org/site/Home/
[127] (pour un résumé de la situation voir http://www.iptc.org/site/Home/Newsletter/Sports_standards_bodies_compete_for_converts).
[128] http://www.w3.org/wiki/WebSchemas/Sports
[129] http://insidesearch.blogspot.fr/2011/08/microdata-sports-stats-happy-fans.html
[130] http://www.bbc.co.uk/ontologies/sport/2011-02-17.shtml
[131] http://www.ddialliance.org/what

Un jeu encore différent est proposé par Statistical Data and Metadata eXchange [132] : il concerne le domaine des données statistiques (ou plus précisément, tout ce qui concerne la description des données statistiques à l'intérieur de documents de domaines différents[133]).
Ces jeux sont spécifiques à une certaine organisation des données scientifiques. Le problème est totalement différent des métadonnées spécifiques au temps ou aux noms de lieux[134]. Ce qui est important, c'est que ces métadonnées caractérisent un domaine scientifique, incluant les données primaires et les besoins des utilisateurs.
Ces jeux sont à lier aux infrastructures que nous avons présentées plus haut. En effet, par domaine de spécialité se construisent des infrastructures qui ont besoin de métadonnées propres de façon à identifier les documents spécifiques qu'elles publient et mettent en relation. Il ne s'agit pas seulement ici de caractériser un vocabulaire de spécialité mais une prédication sur ce qui informatif à propos d'un document pour une communauté précise.

Le problème est légèrement différent dans des cas de formats spécifiques aux documents audiovisuels et multimédias, notamment le MPEG7. En effet, le problème de la description des contenus et de l'organisation des objets multimédias est envisagé dans l'élaboration du format lui-même. Les métadonnées sont jusqu'à présent le seul moyen que l'on ait pour accéder à ces documents, notamment dans le cadre d'une recherche d'information. Les moteurs de recherche classiques utilisent alors nécessairement ces métadonnées.
Nous n'entrerons pas de façon très détaillée dans la multiplicité des enjeux des métadonnées liées à l'image, animée ou fixe. Rappelons que les métadonnées constituent jusqu'à présent, les seuls outils permettant de décrire le contenu des images. Dans le cas de l'image animée, ces métadonnées peuvent être associées à des annotations reposant sur des segmentations du flux des images. Remarquons également que les images animées ou fixes constituent des objets numériques extrêmement nombreux et qui ont des usages très divers. La dimension sémantique de la description en est encore accentuée.
Les métadonnées associées aux images peuvent reposer sur une ontologie, comme le propose l'INA[17]. Une telle solution est également celle qui a été retenue par la BBC.

En définitive, on s'aperçoit que les métadonnées suivent un triple développement :

1. l'adaptation des jeux traditionnels aux perspectives du web de données (hétérogénéité des ressources, dynamique relationnelle notamment)
2. Intégration de plus en plus importante des connaissances dans les jeux généralistes et couramment utilisés comme le Dublin Core (au travers des profils et des relations aux organisations des connaissances), en concurrence (ou en complémentarité) avec l'apparition de descriptions bibliographiques fondées sur des ontologies
3. Développement d'ensembles de métadonnées spécifiques à un domaine d'activité et aux types de documents qui y sont utilisés.

Ainsi, le développement des métadonnées s'inscrit pleinement dans le cadre du web de données. Cette évolution n'entraîne pas la disparition de schémas, mais au contraire une reformulation et une actualisation de l'existant. Cette posture permet aux bibliothèques de s'insérer pleinement dans le cadre du web de données.

Cela dit, tous ces jeux de métadonnées ne sont pas de même grain, à savoir qu'ils décrivent les documents à différents niveaux de précision. Cette caractéristique rend leur compatibilité assez difficile à envisager.

[132] http://sdmx.org/
[133] http://sdmx.org/?page_id=81
[134] http://www.geonames.org/ par exemple

1.7. Différents points de vue sur le document et différentes formes de référencement du document numérique.

Nous présentons enfin le contexte élargi de la description des documents, mais également des ressources permettant de travailler pour ces descriptions et enfin la question de la recherche d'information. Nous terminerons cette présentation par quelques précisions à propos des ontologies et des structures lexicales. Ces deux derniers points seront développés dans les parties ultérieures.

1.7.1. Métadonnées et autres outils de description des documents.

En dehors des objets numériques qu'ils décrivent, les différents jeux sont distincts relativement au point de vue qu'ils adoptent à propos du document. Avant de nous préoccuper des transformations majeures en ce qui concerne les méthodes outils d'organisation du web, nous aimerions présenter la diversité des initiatives concernant la description des documents. Nous devons également ajouter que cette diversification est associée à des modes nouveaux de référencement qui viennent encore accentuer les liens entre les documents.

Ainsi, la DOI (Digital Object Identifier) présente des métadonnées apparemment proches du DC. Un DOI constitue seulement un identifiant pérenne (à la différence par exemple d'une adresse URL) pour un document numérique mais également un ensemble de métadonnées. En ce sens, il s'agit d'un outil particulièrement intéressant pour un éditeur, d'autant plus que le schéma DOI est un format universel (« ISO 26324:2012 specifies the syntax, description and resolution functional components of the digital object identifier system »)[135]. Le jeu de métadonnées associées à la description de l'objet numérique [136]a cela de particulier qu'il ne concerne pas les contenus du document mais seulement ses propriétés formelles. La DOI constitue un seul outil d'identification de l'objet et ne propose aucune description ni classification.

La TEI[137] présente des caractéristiques inverses : si au départ la TEI est conçue comme un outil d'annotation d'archives numériques, notamment de textes anciens, elle comprend un jeu de métadonnées qui lui permet de se présenter comme un outil complet à la fois d'édition électronique et de description de documents. Elle permet des recherches fines à l'intérieur d'ensembles documentaires importants, notamment si les annotations sont supervisées par des lexiques et des terminologies partagés. Malgré la complexité de son appareil d'annotation, liée entre autre à ses domaines d'application littéraires et historiques, le TEI constitue un outil d'actualité pour la description sémantique des documents électroniques.

Dans le cadre du web de données, les métadonnées dépassent ainsi largement le rôle qui leur a été conféré de simple description documentaire. Leur structuration à la fois classificatoire et relationnelle les associe à des structures fondées sur des ontologies. Cette équivalence s'illustre particulièrement pour SEM, structure de représentation d'événement[138]. Ces équivalences permettent aux descripteurs de s'inscrire dans des systèmes relations permettant de représenter événements, des phénomènes et toutes sortes d'autres faits que la simple description de traits de publications. Ces équivalences sont porteuses d'opportunités encore

[135] http://www.datacite.org/node/64
[136] http://www.doi.org/doi_handbook/4_Data_Model.html#4.1
[137] http://www.tei-c.org/index.xml
[138] http://semanticweb.cs.vu.nl/2009/04/eventExtended/

largement inexplorées, et dans lesquelles nous situons le projet que nous développons tout au long de ce travail. (Néanmoins, ces équivalences ne doivent pas tromper : entre la définition de l'événement dans le DC[139] () et dans le cadre de SEM ou de tout autre vocabulaire d'ontologie, il existe de nombreuses différences, et d'abord de domaine d'application.

Les métadonnées s'inscrivent ainsi dans un espace entre les identifications de documents et les annotations. Elles jouent donc un rôle irremplaçable dans le cadre des liens aux moteurs de recherche, mais également dans l'objectif d'une structuration des bibliothèques numériques. Elles peuvent être associées à des classifications et à des outils relationnels.

1.7.2. Métadonnées de travail et de partage : les LINKED DATA.

Nous avons déjà présenté les LINKED DATA dans le cadre général des politiques de structuration du web. Nous analysons maintenant un peu plus précisément leur rôle dans l'élaboration d'un ensemble de services en lignes. Nous développerons plus précisément un aspect des LINKED DATA qui concerne plus particulièrement les bibliothèques, à savoir le management des ressources linguistiques. En effet, comme nous l'avons vu, les représentations terminologiques et lexicologiques (ontologies terminologiques et lexiques électroniques en premier lieu) sont cruciales pour l'intégration des ressources des bibliothèques à l'intérieur de l'offre du web et il s'agit justement d'un domaine dans lequel les LINKED DATA jouent un rôle essentiel.
Les LINKED DATA, dont nous avons vu précédemment les fondements, constituent maintenant un cadre de travail essentiel pour assurer la visibilité sur le web des contenus des bibliothèques. L'objectif des LINKED DATA est néanmoins beaucoup plus large : il s'agit du répertoire structuré de l'ensemble des ressources et outils permettant de structurer le web de données.
Cet investissement est mis en perspective par la question de la recherche d'information, et par conséquent ses évolutions futures.

Les LINKED DATA permettent de construire des outils et des ressources interopérables et ouvertes, de telle sorte que l'on puisse aisément passer de l'un à l'autre et profiter d'une exploitation commune. On peut considérer de différentes façons les LINKED DATA: on peut les considérer comme une nouvelle façon d'envisager les traitements des données, de nouveaux outils et donc de nouvelles méthodes de travail. En effet, à la différence de la conception de bases de données dans laquelle prime la cohérence interne de l'outil, le critère contraignant fondamental de réussite est, en adoptant certains formats, de rendre les données représentées explicites à l'ensemble des outils partageant ces formats.
A cette dimension technique, on peut ajouter celle des contenus. En effet, l'interopérabilité ne constitue une réalité qu'à partir du moment où les contenus sont sémantiquement explicités entre les différents outils. La difficulté consiste alors à spécifier cette dimension symbolique : qu'est-ce que représentent exactement en terme de raisonnement, de signification, ces outils mis à disposition et qui s'écrivent par de mêmes règles ?
Le cadre de travail RDF avait bien spécifié la nature prédicative et informationnelle de RDF. Néanmoins, RDF reste un outil neutre et peut représenter des raisonnements de nature très différente. Le problème de l'interopérabilité n'est pas aujourd'hui syntaxique mais sémantique : quel sens attribuer aux relations établies entre structures ?

Pour notre propos ici, les LINKED DATA constituent avant tout une méthode de travail; en

[139] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=dcmitype#Event

effet, elles permettent de lier des outils différents en utilisant toutes les possibilités de l'interopérabilité. De cette façon, il est possible d'élaborer des outils hybrides, fonctionnant à l'aide d'autres. Ce principe économique (qui permet d'éviter de construire des prototypes quasiment identiques) constitue également un moyen pour élaborer une communauté de services.

La montée en puissance des Linked Data dans le cadre des bibliothèques numériques est liée en partie à la définition d'intérêts communs, résumés dans l'incubateur du W3C « Library Linked Data Incubator Group »[140]: d'une part la plus grande visibilité sur le web, d'autre part, en interne, la réorganisation des bases de données. Comme cela est mis en évidence dans l'ouvrage de Tom Heath and Christian Bizer (2011)[18] *Linked Data: Evolving the Web into a Global Data Space*, les LINKED DATA permettent, grâce au principe des wrappers (des programmes permettant d'activer d'autres programmes), mais surtout de la structure même de RDF, de mieux mettre en évidence les contenus des bases de données par rapport aux outils de type XML : « Just as traditional Web browsers allow users to navigate between HTML pages by following hypertext links, Linked Data browsers allow users to navigate between data sources by following RDF links. For example, a user may view DBpedia's RDF description of the city of Bristol (UK), follow a *hometown* link to the description of the band Portishead (which originated in the city), and from there onward into RDF data from Freebase describing songs and albums by that band. The result is that a user may begin navigation in one data source and progressively traverse the Web by following RDF rather than HTML link ». (Chap. 6.1.1.).
L'un des enjeux les plus profonds des LINKED DATA tient dans la modification opérée dans le rôle et la structuration des bases de données. Cette transformation, due à l'utilisation systématique de RDF, permet de relier des contenus distribués qui ne pouvaient être directement reliés. A propos des bases de données, il faut ajouter que les outils disponibles, et en premier lieu le DC, ne sont pas associés à des bases de données particulières : ils sont indépendant des plateformes de développement des bibliothèques numériques. Enfin, les LINKED DATA constituent uniquement des répertoires : les schémas sont stockés dans d'autres structures comme PURL.ORG.

L'article de Jerry Parson sur les métadonnées bibliographiques [141] peut être considéré comme fondateur d'une nouvelle perception du rôle des LINKED DATA dans le cadre des bibliothèques[142] . Le réseau de l'OPEN KNOWLEDGE relatif aux bibliothèques relate l'augmentation constante du nombre des bibliothèques qui vont choisir les formats ouverts et qui vont donc rendre publiques leurs données.
Les LINKED DATA permettent d'augmenter les performances de la navigation. En rendant la totalité des ressources disponible pour un sujet, via les URI, on accroît considérablement l'offre pour un utilisateur, notamment en permettant un lien entre des données de bibliothèques et d'autres, comme des dictionnaires, des encyclopédies. Ce point est essentiel pour l'intégration des produits de bibliothèque dans l'offre globale du web.

Nous avons déjà évoqué les transformations que les principes du Linked Data entraînent dans la structuration des bases de données. Nous pouvons préciser en prenant par exemple de DRUPAL[143] (qui constitue une plate-forme permettant de construire des sites incluant des bases de données conformément aux principes du LINKED DATA. A propos de DRUPAL, le

[140] http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/
[141] http://www.diglib.org/archives/3167/
[142] http://semanticweb.com/the-linked-data-landscape_b30943
[143] http://drupalfr.org/apropos),

DC a mis en place une session spéciale de sa conférence de 2012,[144]
, concernant la façon dont on peut lier les possibilités de la plate-forme et les propriétés des LINKED DATA. En effet, DRUPAL est utilisée par de nombreux producteurs de métadonnées et de structures de connaissances. Concernant les métadonnées en général, on s'aperçoit qu'elles s'intègrent peu à peu les processus et les règles des LINKED DATA, notamment VIVO et les AGROVOC[145].

L'interopérabilité n'est pas seulement une question technique ; la méthodologie choisie a une importance fondamentale dans la nature des résultats obtenus[146]. En effet, les LINKED DATA permettent de proposer des méthodologies d'interopérabilité et d'échange fondées sur des liens et non sur le principe du plus petit commun dénominateur, ce qui permet une nettement meilleure sauvegarde de la signification.

Comme on l'a évoqué, les principes des LINKED DATA permettent une prise en compte plus fine de la structuration des entités linguistiques, de leurs relations et de leurs dépendances. Par conséquent, au vu de ces potentialités, les outils élaborés dans le cadre des LINKED DATA peuvent devenir un fondement de la construction des infrastructures.
Ainsi, par exemple, dans le cadre de CLARIN, la connexion entre les métadonnées de l'IMDI et ISO-cat a permis l'élaboration de CMDI, un outil permettant de renseigner des documents (contenant des données primaires de la linguistique) en opérant des choix nettement plus larges que par le biais des seules possibilités offertes par l'IMDI. Cette connexion permet ainsi aux métadonnées d'être enrichies par les lexiques du domaine. Plus généralement, l'élaboration des lexiques communs, de terminologies et de métadonnées. Ainsi, l'un des objectifs de CLARIN et de son association au LINKED DATA est la construction de ressources linguistiques communes à l'ensemble des participants aux différents processus de construction du web de données : e-science et open data notamment.

Néanmoins, l'univers des Linked Data est loin d'être simple : la multiplication des outils et des ressources entraîne une difficulté certaine à les retrouver. Le nuage[147] () représentant l'ensemble des ressources liées montre une complexité telle qu'il est difficile de s'y repérer. Un certain nombre de travaux, comme ceux de Nikolov et D'Aquin [19]fondent les bases d'une recherche d'information à l'intérieur des Linked Data de façon à retrouver la ressource la plus appropriée. Ils proposent une recherche par index raffinée puis filtrée par une ontologie.
L'évolution des Linked Data se dirige vers une structuration par domaine de plus en plus appuyée[148]. On n'entend pas ici des domaines académiques, mais des domaines d'application (Media, Geographie, Publications (publications d'objets numériques), Gouvernement, inter-domaines et sciences de la vie)). Cela dit, les ressources linguistiques ont donné lieu à une représentation spécifique, ce qui introduit une structuration disciplinaire des outils[149]. Là encore, les métadonnées jouent un rôle fondamental puisque ce sont elles qui ont seules la possibilité de décrire les ressources.

Les Linked Data constituent le principal outil de structuration des contenus pour l'e-science. En effet, l'interopérabilité généralisée tout autant que la diversité des contenus (ontologies, terminologies, services) que l'on peut éditer dans le cadre des Linked Data en font un

[144] http://dcevents.dublincore.org/index.php/IntConf/index/pages/view/specialSessions-2012
[145] http://aims.fao.org/standards/agrovoc/about
[146] http://dcevents.dublincore.org/IntConf/dc-2011/paper/view/70/40
[147] http://lod-cloud.net/
[148] http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.html
[149] http://nlp2rdf.lod2.eu/OWLG/llod/llod.svg et http://linguistics.okfn.org/resources/llod/

ensemble de ressources particulièrement riche. La structuration même des Linked Data pose un certain nombre de problèmes.

On explique cette primauté des Linked Data pour l'e-science d'une double façon : par l'évolution de la pratique scientifique, et par l'évolution des supports de la recherche.
L'évolution de la pratique scientifique se marque par la réutilisation des données et une collaboration de plus en plus forte entre partenaires. Elle requiert un effort particulier à la fois de description des données et de leur mise en relation, de façon à ce que les ressources soient découvertes, évaluées et utilisées à bon escient. Des questions de connaissance, de vocabulaire et de relations spécifiées entre les données doivent trouver une réponse.
L'évolution des supports de la recherche se rapporte à la numérisation et surtout au fait que cette dernière est native des documents. Cette technologie permet également de décrite très précisément comment les observations, expérimentations et recueils de données sont menés.
De cette façon, le commentaire entre à l'intérieur du document lui-même. Si un tel traitement est possible à propos des données, on peut aisément faire le lien avec les publications. Une telle perspective change totalement l'un des principes des bibliothèques numériques, à savoir la stricte séparation entre données et métadonnées.

1.7.3. Métadonnées et recherche d'information ; évolution des problématiques.

Le principe des moteurs fonctionnant exclusivement à partir de métadonnées réside dans la notion de moissonnage. C'est le cas par exemple d'ISIDORE. Néanmoins, ce type d'outil est toujours restreint à un ensemble identifié de bibliothèques numériques. ISIDORE fonctionne grâce à un moissonnage de métadonnées. Celles-ci permettent ensuite à un moteur de répondre aux requêtes formulées par un utilisateur en employant uniquement des outils sémantiques. Un principe similaire a été retenu pour EUROPEANA. L'enjeu consiste à mettre en œuvre un outil de recherche d'information qui a terme puisse être concurrentiel des moteurs généralistes. Les outils comme SPARQL, élaborés dans le cadre du W3C, permettent de formuler les requêtes interrogeant spécifiquement les métadonnées structurées en utilisant les formats RDF.

Néanmoins, ces exemples restent limités et ne permettent pas de rendre compte de la complexité de l'évolution de la recherche d'information. Comme on l'a vu, les moteurs de recherche généralistes comme GOOGLE améliorent constamment leur algorithme de façon à prendre en compte les données structurées, ce qui permet d'envisager des systèmes mixtes, reposant à la fois sur des probabilités et des structures sémantiques. Surtout, comme on l'a vu, l'enjeu consiste à intégrer les dimensions de la navigation dans la présentation des résultats de la recherche d'information.

En proposant d'étendre la navigation, les LINKED DATA transforment également considérablement la portée et les enjeux de la recherche d'information. Dans le rapport de l'incubateur W3C, l'un des enjeux des LINKED DATA se situe également du côté de la recherche d'information : "Linked Data is not about creating a different Web, but rather about enhancing the Web through the addition of structured data. This structured data, expressed using technologies such as RDF in Attributes (RDFa) and microdata, plays a role in the crawling and relevancy algorithms of search engines and social networks, and will provide a way for libraries to enhance their visibility through search engine optimization (SEO). »

Au-delà, l'effort porte sur le lien entre métadonnées et lexiques. Les travaux menés à propos des ressources linguistiques convergent vers une standardisation des descriptions des ressources linguistiques dans le monde.
Cette élaboration rend de plus en plus pertinente une recherche d'information fondée sur des critères sémantiques. Cela dit, l'état des choses n'est pas aussi simple: les approches statistiques, probabilistes notamment, ne sauraient être contradictoires de celles qui utilisent des schémas. Au contraire, elles peuvent être complémentaires. Et c'est justement le pari des grands moteurs de recherche notamment GOOGLE. Les équipes du moteur se fondent sur les Snippets (c'est-à-dire des portions de code source, notamment des parties de schémas RDFS de métadonnées, ou des parties de textes formellement caractérisées, comme par exemple utilisant des annotations TEI) pour argumenter la montée des dimensions sémantiques dans l'algorithme de recherche[20].

On assiste, du fait des LINKED DATA, à un élargissement du rôle de la navigation et de son articulation à la recherche l'information.
La prise en compte des métadonnées, et plus globalement, des données structurées, constitue un second aspect de l'évolution de la recherche.
Enfin, on doit mentionner la prise en compte de plus en plus marquée du contexte dans le cadre de l'extraction et de la recherche d'information. Nous reviendrons plus tard sur les questions d'extraction d'information et de text mining ; néanmoins, les travaux dans ces domaines montrent également la pertinence des approches intégrant le contexte d'unités lexicales pertinentes.

Cela dit, la pertinence de recherches d'information fondées sur les seuls principes sémantiques n'est pas encore complètement acquise[21], de même les méthodes purement statistiques, qui se doivent d'être couplées à d'autres pour fournir des résultats. SPARQL est un outil d'interrogation, et ne peut prétendre à d'autres fonctionnalités que celle-ci. Par conséquent, l'intégration d'ontologies dans les outils de recherche d'information dynamise les recherches à fondement booléen, traditionnelles dans le champ de la recherche d'information. Elles sont généralement corrélées à des modèles formels de représentation tels que l'Analyse Formelle des Concepts. Néanmoins, il n'existe pas d'application à grande échelle et en autonomie de ces principes[22].

Par conséquent, les évolutions dans le cadre de la recherche d'information ne montrent pas de solutions utilisant une méthodologie unique, mais plutôt des associations d'outils interdépendants. Par ailleurs, du fait de l'annotation et des métadonnées, les données sur lesquelles il est possible de travailler se sont largement diversifiées. Les différentes approches de la recherche d'information s'inscrivent ainsi dans des contextes nouveaux et diversifiés (concernant les données, leur structuration et leur représentation) et eux-mêmes fortement évolutifs.

Cette remarque n'empêche pas l'existence de moteurs de recherche utilisant les outils du web de données, notamment WATSON[150] , SIG.MA[151] , VISINAV[152] (). Quel que soit leur intérêt, ces moteurs restent relativement confidentiels dans leurs usages, même si, comme SIG.MA, ils utilisent le web comme une base de données, autrement dit, comme les moteurs de recherche standard.
Le fonctionnement de ces outils est quelque peu différent. WATSON[23] fonctionne sur la base

[150] http://watson.kmi.open.ac.uk/Overview.html
[151] http://sig.ma/
[152] http://semwebtec.wordpress.com/, http://semwebtec.wordpress.com/tag/visinav/ ou http://visinav.deri.org/

d'un moissonnage de ressources sémantiques sur le web, à savoir d'ontologies et de documents décrits en utilisant les formats RDF et OWL. Ces ressources collectées sont alors analysées de façon à produire des métadonnées qui elles-mêmes constitueront les mots-clés des recherches. A ce moment-là l'utilisateur pose les questions qui l'intéressent. Une telle structuration demande un travail important d'analyse des différentes ontologies. Ce point est également soulevé par les concepteurs de SIG.MA : « The task at hand is however particularly complex. Assuming that an entity is indeed sufficiently described by available Semantic Web data, these descriptions can often be very heterogeneous and exhibit problems such as different describing ontologies, missing links between descriptions, little or no reuse of identifiers for the same entity, data errors, poor RDF publishing practices, and more. »[24]).
SIG.MA moissonne à la fois les documents représentés en RDF et utilisant les micro-formats.
SIG.MA présente les données trouvées de façon à ce qu'apparaisse une agrégation de liens structurés. A l'utilisateur de suivre les liens correspondant le plus à sa recherche. A la différence des autres moteurs de recherche du web de données, SIG.MA se présente donc à la fois comme un outil de recherche et de navigation. VISINAV possède les mêmes propriétés.
En définitive, les moteurs de recherche fondés sur des ontologies et plus globalement des structures conceptuelles amènent une transformation importante de l'idée que l'on peut de faire de la représentation des résultats d'une recherche d'information sur le web. En effet, plutôt que de simples sites, on peut représenter leur description dans une certaine ontologie, et donc caractériser leurs contenus. Ce système fonctionne fondamentalement sur l'agrégation et par conséquent permet de présenter comme résultats des ensembles de ressources liées et décrites. L'exploration par navigation peut alors commencer.

Dominée par les grands moteurs de recherche du web d'une part (GOOGLE et YAHOO) et les applications sur des ensembles de données restreintes d'autre part, la recherche d'information intègre progressivement les évolutions générales, sachant que celles-ci, au moins jusqu'à présent, ne remettent pas en cause les fondements théoriques et méthodologiques de la recherche dans le domaine.
Par contre, les outils développés utilisant les formats RDF et les micro-données constituent un changement potentiel assez radical dans la façon dont on peut représenter les résultats d'une recherche d'information et les exploiter. Cette innovation, qui constitue le fondement de SIG.MA n'est pas prise en compte dans ISIDORE, qui reste un moteur encore largement dépendant des modes traditionnels de représentation des résultats de recherche d'information.
Or les micro-données, qui sont des outils extrêmement légers puisqu'il s'agit de schémas de tags que l'on insère à l'intérieur de pages HTML constituent un élément important de l'évolution de la recherche d'information sur le web, mais surtout pour la caractérisation des contenus des documents et plus particulièrement la façon dont on peut lier ces caractérisations de contenu avec les descriptions.
Cette question de la recherche d'information reste complexe dans la mesure où elle est dépendante des pratiques des usagers. Or celles-ci sont encore pour une large part dominées par les moteurs de recherche généralistes, lesquels s'approprient progressivement les outils du web de données. Par conséquent, la mise en œuvre de modèles alternatifs n'est pas simple.
Par ailleurs, les bibliothèques numériques justifient leur existence par leur utilisation, et donc leur visibilité (et d'abord celle de leur offre documentaire) sur les moteurs de recherche. Cet impératif oblige à des stratégies de description de l'information compatibles avec celles utilisées par les moteurs.

1.7.4. Des outils de représentation des connaissances vers les ontologies.

Nous traiterons ici des outils rangés sous l'appellation « Knowledge Organization System » ou KOS. Il s'agit de thésaurus et d'ontologies, c'est-à-dire de structures conceptuelles qui organisent un domaine. On peut y intégrer également les vocabulaires contrôlés et les taxonomies. Le trait fondamental de ce domaine est bien évidemment la montée en puissance des ontologies, qui constituent des outils généralisées de structuration des connaissances, distincts et complémentaires des thésaurus et autres modèles conceptuels. Nous ne traiterons pas de l'ensemble des questions soulevées par les ontologies : nous nous limiterons à indiquer certaines de leurs applications dans le cadre des bibliothèques numériques (et plus généralement de plateformes de ressources), et qui ont donc un intérêt particulier pour notre propos : nous choisissons d'illustrer cette montée en puissance par des exemples d'utilisation contrastés.

Au départ, le mouvement s'est amorcé par une reprise des concepts de l'IA vers des outils relationnels : la montée en puissance des ontologies repose sur les structurations relationnelles des concepts, ce qui permet d'élaborer des représentations relativement simples et utilisables, entre autre dans le domaine de l'information. L'idée fondamentale consiste à intégrer des raisonnements dans les relations établies à l'aide des descriptions fournies par XML et RDF. Le langage OWL remplit cette fonction. Le lien entre les raisonnements de l'IA, notamment ceux qui sont caractérisés par les graphes conceptuels, et le RDF, a été largement démontré par DIENG[25]. En élaborant la recommandation OWL, le W3C rend aisée l'intégration des connaissances à l'intérieur du web, et donc l'utilisation d'ontologies dans le cadre de la structuration des services du web.

A partir du moment où ces différents outils peuvent être mis en relation dans le cadre d'outils de recherches fédératifs, ce sont des problèmes théoriques qui se posent : caractérisation des domaines, niveau de représentation des données, vocabulaires. Les questions d'ontologies ne sont pas isolées de celles du langage et des autres outils de structuration des connaissances. Ainsi, les questions qui se posent sont liées à la référence, à l'interopérabilité et au raisonnement.

Comme on a pu le voir à propos d'EUROPEANA, l'utilisation de SKOS permet de mettre en relation différentes structures de connaissances de façon à les utiliser simultanément. En ce sens, les questions de signification et de traduction deviennent fondamentales dans la mise en relation des structures de connaissances. Cette mise en relation permet de construire un moteur de recherche interrogeant systématiquement les différentes bases de données muséales. L'ontologie sert à unifier l'interrogation de bases utilisant des thésaurus différents linguistiquement et culturellement.

Dans le cadre des bibliothèques numériques, le choix de structurer les collections par une ontologie tend à se répandre. Cela a d'abord été le cas de DBLP[153] : l'ontologie sert d'abord d'outil de structuration des notices répertoriées dans la base et par ailleurs d'outil pour l'interrogation.

Pour METASHARE, les ontologies permettent ainsi une meilleure navigation à l'intérieur d'une collection. METASHARE n'est pas une bibliothèque numérique mais un ensemble de ressources destinées à la communauté des chercheurs préoccupée par les questions du

[153] http://swat.cse.lehigh.edu/resources/onto/dblp.owl

traitement de la langue. METASHARE fonctionne à partir d'un moissonnage de métadonnées, et une ontologie qui permet ensuite de les structurer[26]. L'ontologie sert ici d'abord à structurer l'ensemble des métadonnées à propos de ressources pertinentes[27]

Les ontologies identifiées ici sont descriptives de collections (de ressources, de notices ou encore de thésaurus) ; d'autres applications entraînent l'élaboration d'ontologies descriptives de structures d'information, comme par exemple celle de la BBC. On s'approche alors des schémas de SCHEMA.ORG.

Une autre caractéristique fondamentale des ontologies est lié à leur expression dans de mêmes formats. Il est donc relativement aisé de relier des ontologies différentes de façon à mettre en relation les objets qu'elles décrivent chacune. Cette possibilité offerte est l'alignement d'ontologies[28]. Une autre est la fusion d'ontologie et une autre encore le mapping. Il s'agit d'une cartographie qui permet de lier plusieurs structures différentes. On utilise alors les treillis (FCA, FRA[29]). Par ailleurs, RDFS permet de disposer de propriétés et de classes, mais également de sous-classes et de sous-propriétés qui constituent autant d'outils permettant de mettre en relation des structures.

L'idée fondamentale est que sur un portail, on puisse trouver un ensemble complet de services complémentaires, de sorte que l'utilisateur puisse facilement, dans le domaine qui le concerne, réaliser l'opération qu'il souhaite. Une ontologie permet de façon relativement aisée de le faire.
Cette idée se réalise en grande partie dans le fait que les structurations produites en utilisant OWL, donc des ontologies, sont beaucoup plus riches et proches du langage naturel que celles élaborées avec des bases de données relationnelles : « RDF, its query language SPARQL and its logical extension OWL provide features and expressivity that go beyond relational databases and simple graph-based representation strategies. This expressivity poses a performance challenge to query answering by RDF triples stores, inferencing by OWL reasoners and of course the combination thereof.”[30]

Néanmoins, les ontologies constituent un niveau abstrait de structuration des données. Il ne peut être véritablement opérationnel que lié à un niveau beaucoup plus proche du langage naturel. Si les ontologies restent des outils privilégiés pour représenter des domaines et les relations que l’on peut y trouver, leur généralité ne permet pas de proposer des outils descriptifs totalement adéquats. Sachant qu’il existe une longue tradition de lexiques et dictionnaires numériques, un champ important de recherche s’est constitué autour de la mise en relation des ontologies et des lexiques. Ces travaux ont une portée plus vaste que les seules relations entre attributs et éléments de métadonnées d’une part, et les valeurs associées à des langages contrôlés d’autre part.

1.7.4. Outils lexicaux et terminologiques.

Nous n'avons guère développé ici les questions liées à l'annotation des documents, à l'extraction d'information, et plus généralement au traitement automatique des textes. Nous aimerions donc reprendre et approfondir ce que l'on a pu en dire à propos des LINKED DATA. De plus, nous étions restés sur les questions de métadonnées, sans poursuivre à propos des autres usages des lexiques, notamment dans le cadre de l'annotation.
En effet, avec plus ou moins de complexité, les langages de représentation du web sont caractérisés par une structure binaire, prédicative, spécifiées par l'articulation d'instances et de types. Le niveau typique est marqué par des unités et des structures conceptuelles et celui des

instances par des unités linguistiques.

Sachant que le point de départ du web de données est que l'on ne peut entrer dans la signification des documents mais seulement la caractérisation des contenus, on ne peut se passer d'une attention particulière au niveau des représentations lexicales. En effet, il s'agit bien ici des termes qui décrivent les documents individuels.
Egalement, dans le cadre de la recherche d'information, mais aussi de l'évolution de l'édition des documents, on pourra faire apparaître qu'un effort très important est entrepris afin de caractériser et de normaliser les descriptions lexicales.

Ces travaux, conceptualisés notamment par Laurent Romary (op. cit), engagent un mouvement d'équilibrage entre les dimensions les plus abstraites du web et celles qui apparaissent au niveau le plus concret, celui des lexiques et des terminologies. On peut distinguer les lexiques ne contenant que des listes de mots des lexiques terminologiques. Dans le cadre des problématiques d'interopérabilité sémantique, ce sont ces derniers qui sont utilisés, notamment parce qu'ils permettent de lier des structures lexicales entre elles, et de relier ces structures lexicales aux concepts formant les ontologies. La définition du sens des unités lexicales permet effectivement la normalisation.

Les projets d'infrastructures s'appuient sur des initiatives visant à normaliser les vocabulaires de description, notamment dans le cadre de CLARIN: ISO-cat et LEMON[154] de LexInfo. Comme nous l'avons déjà largement présenté, ISO-cat constitue un outil permettant de structurer un lexique caractérisant les ressources linguistiques.
A l'appui de ce que nous avons pu dire des infrastructures, ISO-cat[155] constitue une fédération de ressources existantes au niveau européen. L'objectif étant d'assembler des initiatives antérieures, les objectifs propres de ces projets ont été sauvegardés, les différences de niveau (terminologique, ontologique, lexical) sont parfaitement assumées.
Ce qui est commun entre ces initiatives (ou leur adaptation) est le fait qu'elles utilisent les mêmes formats de représentation (ISO/IEC 11179[156]) et qu'elles proposent des définitions des termes qu'elles intègrent. La structuration du travail se fait par groupes relevant à la fois d'un découpage disciplinaire (morphosyntaxe, discours) et relatif à des niveaux de description des ressources (métadonnées, terminologie, management multilingue de l'information). En définitive, ISO-cat est une collection de données qui permet des usages multiples, de la description de document à l'annotation.
ISO-cat est donc un outil permettant d'unifier une offre diverse et éparpillée en même temps qu'elle oblige à un certain format. ISO-cat montre que l'unification, la normalisation de la syntaxe n'altèrent pas la diversité des projets.
Les structures lexicales sont soumises à des formats eux aussi normalisés.
Lexical Markup Framework (ISO 24613)[157] constitue un standard de représentation lexicale utilisant XML. Il peut être lié à LexInfo, un outil qui représente l'information lexicale relative à une ontologie ; il utilise des catégories lexicales, et possède une dimension non prescriptive qui permet des usages souples, comme par exemple dans le projet MONNET[158]. Le modèle LEMON proposé dans ce projet, constitue un outil particulièrement complet. L'ensemble de ces outils sont fondés sur la complémentarité du niveau lexical et du niveau conceptuel de représentation des informations et des connaissances.

[154] http://www.lemon-model.net/
[155] http://www.isocat.org/
[156] http://metadata-standards.org/11179/
[157] http://www.lexicalmarkupframework.org/
[158][158] http://www.monnet-project.eu/Monnet/Monnet/English?init=true

Les outils lexicaux servent enfin à la standardisation des infrastructures de métadonnées, notamment dans le cadre de CLARIN (581 dans LREC 2012). Il s'ensuit la possibilité de lier l'ensemble de ces structures dans le cadre des LINKED DATA (544 dans LREC2012).

De façon plus classique, la représentation lexicale est utilisée dans le cadre de l'extraction d'information et de l'annotation. Dans ce cadre des outils comme FRAMENET ont depuis un certain temps montré leur efficacité et leur adaptation à des langages différents[31]. Si ces méthodes à base lexicale sont fondées sur des modèles linguistiques, elles ne sont pas néanmoins utilisées en isolation. En effet, une extraction lexicale à grande échelle peut articuler des outils linguistiques et métriques, et donc améliorer la qualité des tâches impliquant l'exploration du document[32]. (493 dans LREC 2012).

De façon générale, l'élaboration de dictionnaires électroniques transforme considérablement la méthodologie de leur élaboration. Face aux approches fondées sur des communautés d'experts, on peut mentionner les approches fondées sur les wiki, notamment dans le cadre de la Wikipédia, avec le lexique Wiktionary[159] . Il s'agit d'un outil reposant sur la construction de lexiques et de définitions fonctionnant de façon collaborative. Néanmoins, l'utilisation de ce dictionnaire est difficile dans un contexte de tâches automatisées[33]. Dans cette communication, il est proposé de traiter le Wiktionary de façon à ce qu'il s'intègre dans le cadre des LINKED DATA via un modèle approprié.

De tels travaux sont pour nous de la première importance parce que le web de données travaille sur deux niveaux d'abstraction différents et que par ailleurs les structures d'information se caractérisent autant dans leur dimension linguistique que par rapport à des schémas.

1.7.5. Enjeux sémantiques liés aux métadonnées.

La généralisation d'une écriture RDF permet l'élaboration de jeux de métadonnées fondés sur la mise en relation de documents et l'association de différents jeux. De plus, la multiplication des points de vue et des descriptions associées au document pose le problème d'une trop grande hétérogénéité des descriptions et des formats utilisés. C'est pour cela que des tentatives d'unification sont apparues, recourant à des schémas RDF, OWL et SKOS. Ces schémas uniques assemblent des éléments des différents jeux et certains éléments propres de façon à offrir aux indexeurs un outil unique dans lequel il peut choisir. Ainsi, la Bibliographic Ontology[160] associe à ses propres descripteurs (qui concernent essentiellement les aspects contextuels et formels des documents), les schémas des autres métadonnées, notamment le DC, de façon à fournir un ensemble de descripteurs plus précis que ceux existant (par exemple en détaillant les renseignements sur les auteurs des publications scientifiques : nom de laboratoire, email des auteurs, ou en intégrant les numéros de page d'un articule, etc.)[161].).
L'enjeu de la Bibliographic Ontology dépasse largement la simple description documentaire pour l'articuler aux éléments de citation : « The Bibliographic Ontology describe bibliographic things on the semantic Web in RDF. This ontology can be used as a citation ontology, as a document classification ontology, or simply as a way to describe any kind of document in RDF. It has been inspired by many existing document description metadata formats, and can be used as a common ground for converting other bibliographic data sources ».

[159] http://www.wiktionary.org/
[160] http://bibliontology.com/
[161][161] Voir : http://bibotools.googlecode.com/svn/bibo-ontology/trunk/doc/index.html

Même si aujourd'hui les possibilités offertes par la Bibliographic Ontology dépassent largement celles des infrastructures en cours de réalisation, le lien entre citation, classification et description commence à se mettre en place. Les évolutions que l'on peut en attendre sont fondamentales concernant le rapprochement entre les outils de recherche d'information fondés sur les citations, ceux qui reposent sur des classifications et ceux qui sont fondés sur les métadonnées.

Si les métadonnées s'inscrivent totalement dans les mouvements que nous décrivons, elles rencontrent aussi une difficulté à laquelle aucune réponse n'a été jusqu'à présent portée : la difficulté du travail humain de renseignement. Si les éditeurs sont en général convaincus de l'intérêt majeur de cette tâche, J. Dinet[34] a montré la difficulté posée par les usagers pour l'alimentation de ces métadonnées.
Cette difficulté peut être contournée tout simplement par la proposition de métadonnées sans description des contenus. C'est le cas de la DOI notamment. C'est en ce sens que ce jeu rencontre l'intérêt des éditeurs. Par contre, pour tout ce qui concerne la recherche d'information, les métadonnées telles que le DC ou encore la TEI ont un avantage certain dans le cadre de recherches d'information. Comme on a pu le noter à propos des rapports PEER, les métadonnées apparaissent comme un enjeu important dans l'articulation entre les éditions électroniques opérées par les bibliothèques électroniques et celles opérées par les éditeurs. En effet, si les premières sont les plus faciles à identifier par les moteurs de recherche, les exemplaires éditeurs contiennent l'ensemble des informations pertinentes pour la citation dans le cadre d'une nouvelle publication.

Une seconde évolution est liée à l'utilisation de RDF. En effet, les relations <attribut-valeur> de XML sont complexifiées par les triples RDF et la possibilité d'inscrire des prédicats ou des valeurs issus de vocabulaires divers. RDF relance la question de l'intégration des lexiques, terminologies et des structures d'organisation des connaissances dans les descriptions de documents et les relations entre documents.
Ce sont donc des questions relatives au bas niveau d'abstraction qui se posent maintenant. En effet, les structures lexicales et terminologiques peuvent donner lieu tout autant que les structures d'organisation des connaissances comme les ontologies et les thésaurus à des représentations utilisant les standards du web. Par ailleurs, dans de nombreux travaux relatifs aux ontologies (notamment SMITH[35]), il est affirmé la nécessité de coupler une ontologie à une représentation terminologique ou taxonomique de façon à ce que le lien opéré entre l'ontologie et ce qu'elle représente soit validé par une structuration linguistiques des entités lexicales.
Un autre enjeu des métadonnées, lié au précédent, est aujourd'hui leur association avec les langages documentaires. Jusqu'à une date récente, les deux univers ont vécu un développement séparé : « The field of Library and Information Science and the Semantic Web community have been turning circles around one another ever since structured search for information on the Web became a major field of research in the second half of the 1990s. Both Library and Information Science and the Semantic Web community have been evolving significantly ever since. From the end of the nineties, the two communities invested considerably in the standard making process of metadata schemas and ontologies. However, the practice of gathering domain and technology experts to debate over a period of years to develop and fine-tune metadata schemas and ontologies has in both worlds lost a lot of its institutional support. The eContentplus funding program of the European Commission, for example, explicitly did not fund the development of metadata schemas and the creation of metadata itself (van Hooland et al., 2010). The early-to-mid 2000s economic downturn in the US and Europe forced both fields to adopt a more pragmatic stance and to deliver short-term

results towards grant providers. It is precisely in this context that the concept of Linked and Open Data (LOD) has gained momentum. » pp.1-2][36].

Enfin, cette présentation ne saurait être complète que si l'on intègre la question de l'intégration dans le web des bases de données traditionnelles. Dans ce cadre, les groupes de travail W3C Library Linked Data Incubator Group et International Linked Open Data in Libraries, Archives, and Museums Summit (LOD-LAM) ont permis un certain nombre de rapprochements que l'on va pouvoir observer. Les auteurs partent du constat suivant : « Library data today resides in databases which, while they may have Web-facing search interfaces, are not deeply integrated with other data sources on the Web. There is a considerable amount of bibliographic data and other kinds of resources on the Web that share data points such as dates, geographic information, persons, and organizations. In a future Linked Data environment, all these dots could be connected. ». (3.1.1.)[162]. Le problème posé ici est bien l'intégration des grands outils des bibliothèques aux autres ressources du web de façon à apparaître plus clairement dans les résultats des recherches d'information. Cette première remarque est suivie d'autres à propos des outils eux-mêmes, conçus pour les bibliothécaires et non pour d'autres communautés, la spécificité de la terminologie et enfin la dépendance des bibliothèques par rapport à des prestataires externes.

Cette dernière sous-partie permet d'amorcer ce qui constitue un enjeu fondamental de notre travail, à savoir l'appréhension des questions de sémantique à l'intérieur du web de données, en considérant plus particulièrement la description bibliographique (ou documentaire). En ce sens, ces propos seront développés tout au long de notre travail.

CONCLUSION.

Nous avons voulu montrer dans ce rapide état de l'art de l'évolution des bibliothèques numériques et des outils qu'elles utilisent quelques traits essentiels :

1. Le passage de questions liées aux documents et aux collections vers des problématiques de mise en relation entre les documents et les collections.
2. Le passage de questions de données vers des caractérisations de connaissances et de relations entre ces connaissances.
3. Le passage d'un web de dépôt à un web de travail.

Nous retiendrons également le fait les outils du web de données s'intègrent facilement à l'intérieur des productions communes du web, et notamment les pages HTML et les outils de recherche traditionnels. L'exemple de schema.org est à ce titre évocateur d'une intégration des nouveaux outils dans le web déjà établi.
L'ensemble de ces questions est encore en cours de reformulation, en lien à la construction de réseaux d'échanges de données. Nous avons parlé d'infrastructures et d'outils spécifiques de description associés à ces infrastructures. Nous avons vu les enjeux européens liés à ces infrastructures.

D'un point de vue plus stratégique, nous pourrons conclure à la montée en puissance des principes des Linked Data, parallèle à l'effort en direction de l'open data. Ainsi, on peut

[162] http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

considérer la participation d'IBM[163] au groupe de travail pour la plate-forme des Linked Data[164] () comme un changement stratégique essentiel; d'après A. Le Horz, responsable de la recherche chez IBM, c'est l'échec de produits tels qu'IBM Rationale, dû à la concurrence d'outils gratuits similaires sur le web, qu'IBM a décidé de s'investir dans des projets de LINKED DATA. Deux autres raisons sont avancées :

comme IBM Rationale est un outil de gestion de cycle de vie, l'apport de solutions et de révisions est beaucoup plus facile en utilisant des outils ouverts que des logiciels propriétaires. Dans le même ordre d'idée, d'autres outils comme IBM Watson recourent à des données issues des Linked Data.

l'intégration des industriels dans le processus des Linked Data. (Cette intégration est aussi une façon de chercher à peser sur certains processus, notamment ceux qui sont liés à l'évolution des formats : adoption de Turtle comme syntaxe requise pour RDF,

Cette présentation a permis de mettre en évidence le rôle fondamental de RDF comme outil de structuration du web de données. En intégrant dans le web des objets qui existent en dehors du web (qui sont des objets du monde comme les lieux, les choses ou les personnes), RDF participe d'une construction sémiotique complète. Cette propriété permet aux langages de représentation de se rapprocher des langages naturels et de leur capacité d'expression et de raisonnement.

Enfin, de façon plus contextuelle, cette partie a permis de montrer l'importance des grands projets d'infrastructure et par-delà les pouvoirs publics dans la montée en puissance du web de données. Les raisons en sont à la fois politiques et économiques. Les bibliothèques numériques constituent un pilier essentiel de ces politiques, qui concernent certes les infrastructures, mais qui ont un impact très important sur les problématiques de recherche. Ces dernières, nous les avons juste effleurées dans cette partie. Elles seront nettement développées dans les parties suivantes. En effet, cette partie, dont l'objectif consistait d'abord à décrire un cadre de travail, a permis de soulever en ensemble de questions dont les réponses ne seront fournies, au moins partiellement, qu'au cours de ce travail. Evidemment, certains points seront précisés.

[163] http://www.w3.org/QA/2012/05/interview_ibm_on_a_linked_data.html
[164] http://www.w3.org/2012/ldp/wiki/Main_Page

PREAMBULE AU PARTIES 2, 3,4 et 5.

Nous allons maintenant proposer un certain nombre de théories, d'approches, qui semblent à priori éloignées des bibliothèques numériques. Or ce sont ces recherches qui seront ultérieurement utilisées pour répondre à certaines questions posées par les bibliothèques numériques. Il convient donc de les exposer, de les expérimenter, avant de voir quel pourrait leur intérêt dans le cadre des bibliothèques numériques.

Comme ces propositions scientifiques ne sont pas systématiquement connues, nous passerons un certain temps à les présenter et à les expérimenter indépendamment de notre contexte de travail. Nous reviendrons dans la dernière partie sur ces travaux et nous verrons la façon dont ils contribuent à la formulation d'un projet concernant les bibliothèques numériques.

Cette présentation concerne trois domaines liés :

- les théories de l'information et les flux,
- la sémantique et les structures d'information.
- L'anthropologie cognitive située et distribuée.

Nous commencerons néanmoins par la clarification de notre usage de certains concepts, de façon à éviter des confusions d'interprétation.

Rappelons ici deux points qui ont été évoqués dans la partie précédente et qui ont leur importance dans l'évolution des bibliothèques numériques, mais aussi et surtout dans notre recherche :

- La question sémantique. Comme nous l'avons vu, RDF est neutre par rapport aux présupposés cognitifs et sémantiques des différents outils qui l'adoptent. Cette faible contrainte permet de produire des outils très différents quant à leur présupposé en matière de raisonnement et de signification.
- La structuration d'un travail de recherche. Une première rupture dans le travail de recherche est la facilité d'élaboration et de publication d'un outil : élaborer une ontologie et publier un schéma constituent ne prennent ni un temps ni ne requièrent des moyens très importants. Par conséquent, les projets sont rapidement menés à terme et disponibles pour des applications. Cette publication très rapide ne permet pas toujours d'expliciter les dimensions sémantiques des outils ni les caractéristiques d'usage. On peut toujours arguer que ces sont les utilisateurs de ces outils (les plateformes et les infrastructures) qui vont spécifier les usages et la sémantique. Néanmoins, on ne peut les ignorer, surtout s'il s'agit de produire un nouvel outil.

Ces deux questions peuvent sembler éloignées. Elles sont corrélées parce qu'elles se traduisent pour nous dans la construction d'un projet et d'une méthode. En effet, si par exemple les ontologies disposent de fondements théoriques stables au travers des propositions de Barry Smith et Nicola Guarino, le champ est nettement moins structuré pour caractériser la perspective relationnelle dans laquelle on positionne notre projet. Nous estimons donc fixer un cadre théorique pertinent pour caractériser le raisonnement entre structures de données hétérogènes. Les parties qui suivent s'intéresserons à cela.

PARTIE 2. Méthodologie pour la conception à partir d'analyses

Nous proposons maintenant un certain nombre de concepts pour formaliser une méthodologie susceptible d'articuler les dimensions d'analyse et de méthodologie d'élaboration. Partant du constat que l'évolution essentielle actuelle du web est la mise en relation des documents (pour la recherche d'information, le travail collaboratif, la navigation), nous recherchons à concevoir un outil qui permettrait de spécifier les liens entre les documents de façon à ce que les apports informationnels associés à chaque relation soient spécifiés.
Néanmoins, on ne peut pas facilement concevoir un tel projet sans avoir pu observer des phénomènes similaires dans le monde. Pour nous, cette mise en relation ne doit pas seulement apparaitre pertinente dans un cadre logique ou technique mais doit pouvoir être attestés dans le monde, dans des activités quotidiennes hors du web. Cette condition est celle de la condition d'une validité humaine du raisonnement proposé. La méthodologie que l'on adopte sera proche de celle qui est souvent proposée en robotique, où l'observation de phénomènes naturels (les colonies de fourmis, les insectes, etc.) servent à concevoir des robots. Toute cette partie visera à caractériser le lien entre des phénomènes observables et des modèles.
Nous nous appuierons sur le fait que d'élaborer des modèles puis des systèmes sur la seule base de leurs fonctionnalités et des possibilités des langages de représentation constitue un risque important vis-à-vis de l'appropriation par l'usager. Par contre, les outils qui sont fondés sur des raisonnements déjà largement observés ont beaucoup plus de chance d'avoir une pertinence pour les utilisateurs parce qu'ils reproduisent une habileté cognitive.

Les questions de sciences humaines et sociales deviennent relativement complexes au sein de ce travail. Elles concernent les usages des dispositifs, mais également le fait qu'un outil est l'externalisation d'une compétence cognitive. Elles concernent enfin les questions de langage.
Globalement, la question des usages concerne le modèle et caractérise ce à quoi il sert et comment il se positionne dans l'offre documentaire. La dimension cognitive (et plus précisément de cognition sociale) caractérise l'adéquation de l'outil par rapport à une compétence cognitive, ce qui permet de caractériser plus précisément le raisonnement que l'on représente. Les questions linguistiques lient l'ensemble du dispositif puisqu'il est question tout au long du travail, de langage.
Dans le cadre de notre projet, les problèmes dont il est question maintenant sont de l'ordre de l'élaboration d'un outil (ou plutôt le transfert d'une compétence depuis le monde des activités finalisées vers celui du web). Il ne s'agit pas de l'étude de l'usage d'un outil déjà élaboré.

On caractérisera d'abord la façon dont on entend l'usage et les méthodes que l'on peut utiliser pour son analyse. Une approche dite sémantique, incluant certaines définitions du langage et de ses propriétés, sera présentée ensuite. Enfin, nous établirons les limites de l'analyse, notamment par rapport aux sciences cognitives et aux hauts niveaux d'abstraction qui y sont liés.

Notre hypothèse est que les modèles relationnels actuels qui sont utilisés dans le cadre du web de données ne représentent des raisonnements de que façon très limitée. Cette hypothèse est partagée par ceux qui en logique notamment, visent à asseoir la conception d'outils du web de données sur des bases formelles comme les théories de types et les logiques de description. Nous cherchons donc à montrer que la solution que l'on propose est beaucoup plus élaborée et puissante si elle est fondée sur des assises théoriques élaborées. Néanmoins, pour cela, on doit pouvoir montrer les qualités de ce modèle. Et avant tout, construire un appareillage

scientifique qui soit à même de le montrer. C'est à cette première étape que nous nous employons maintenant.

Lier analyse et conception : postulats fondamentaux.
Nous postulons que l'on ne peut véritablement caractériser un phénomène que l'on veut représenter sous forme de modèle que si l'on dispose d'une définition explicite de ses limites et de ses usages. Pourquoi ce modèle-là serait-il plus pertinent qu'un autre, à quel usage il correspond dans le monde ?

Afin de traiter un problème d'information, on envisagera d'abord sa traduction dans des observables, à savoir des phénomènes du monde qui représentent le problème dans l'ensemble de ses dimensions. Cette hypothèse nous sert à fonder la modélisation des flux (de relations inférentielles entre deux structures de données (ou d'objets) hétérogènes. Cette modélisation doit apparaître comme une représentation pertinente de phénomènes culturels. Un objet scientifique comme les flux implique de caractériser dans le monde un objet matériel ayant des propriétés similaires à l'objet abstrait défini dans notre problématique. Cette démarche est liée au fait que la question des flux a d'abord été posée dans un cadre d'ingénierie, puis reprise dans des cadres mathématiques et philosophiques relativement abstraits. Par ailleurs, une critique généralement formulée (notamment par J. Collier[37]) à la théorie centrale pour aborder les flux de Barwise et Seligman, réside dans la trop grande imprécision de ses objets d'application. Une caractérisation de ce modèle dans le monde permettra au mieux de cerner sa portée et de préciser les phénomènes qu'il représente.

On présuppose que les questions d'information constituent un problème culturel universel et que l'analyse de l'existant (de phénomènes observables réguliers) permet la spécification du problème et la transmission d'un certain héritage. En d'autres termes, à partir du moment où l'on disposera d'un modèle de l'information suffisamment abstrait pour s'appliquer à toute situation d'information, alors on pourra caractériser l'information dans le cadre de l'élaboration de services. Ainsi, on utilise un acquis culturel. (Les outils et les procédures techniques sont des entités culturelles et les connaissances qu'elles contiennent peuvent être transmises[38]). Cette généralité et l'ampleur des objets impliqués rendent possible le lien entre des approches apparemment très éloignées du web de données comme l'anthropologie cognitive.

Nous nous appuierons sur notre expérience, à savoir le fait que l'observation de certaines régularités dans le cadre du fonctionnement d'un certain service de pharmacie hospitalière a permis de faire émerger un modèle des flux et par extension une représentation de la structure d'information. Nous ne considérons pas la pharmacie comme un cadre d'application du modèle, mais comme un espace dans lequel les phénomènes que l'on cherche à expliciter peuvent être observés sans ambigüité[165]. Cette modélisation, qui constitue une représentation des régularités de circulation d'information, n'est d'aucune utilité immédiate pour la pharmacie hospitalière (au moins pour ce qui est de la conception d'outils d'information dans la période actuelle).
Cette étude empirique permet de caractériser précisément le modèle que l'on propose et de justifier sa pertinence non seulement pour caractériser les relations entre deux ensembles

[165] Nous y reviendrons plus amplement, mais il est important de l'avancer déjà : la pharmacie hospitalière a dans le cadre de ce travail, un rôle proche de celui que joue un terrain dans le cadre des études anthropologiques. Nous reviendrons également sur le lien qu'il peut y avoir entre d'une part le lieu d'identification et de description des phénomènes servant à construire la modélisation, et d'autre part l'espace dans lequel s'applique la modélisation, et dans lequel elle joue un rôle pour la conception d'un outil.

structurés et hétérogènes de données, mais également pour montrer ce que chaque ensemble de données transmet comme information à l'autre. Cette étude empirique permet de mettre en évidence la pertinence d'un modèle fondé sur les flux pour l'enrichissement d'une structure par une autre, à laquelle elle est liée. Comme le modèle sera conçu à un niveau d'abstraction suffisant pour être indépendant d'un contexte particulier, il pourra alors s'intégrer dans tout domaine d'activité où un flux d'information peut être observé, modélisé ou encore inséré dans un outil. Ce modèle peut servir à toute situation présentant les mêmes caractéristiques fonctionnelles. Nombres de situations du web de données possèdent ces propriétés.

Nous pourrons donner l'impression ici de nous éloigner des questions de bibliothèques numériques pour envisager les questions de circulation d'information de façon plus abstraite. Ce mouvement est pour nous nécessaire dans la mesure où l'on souhaite construire un cadre théorique pour appréhender un phénomène que l'on ne peut limiter à une seule question de bibliothèques.
Cette abstraction des phénomènes prend deux chemins dont on s'assurera de la convergence : les Sciences humaines et Sociales d'une part, les représentations formelles à fondement logique d'autre part. L'ensemble est lié à la question de la prédication, dont on a vu qu'elle était partagée entre ces différentes perspectives scientifiques, et surtout, qu'elle constituait un point fondamental de la structuration du web de données au travers des capacités d'expression de RDF.

2.1. Différentes définitions de l'usage.

On reprendra ici une définition de l'usage, considéré comme la façon dont des entités sont mises en œuvre dans des contextes fonctionnels. Une telle définition est adaptée à des contextes qui ne sont pas seulement linguistiques, mais de toute activité requérant le langage pour atteindre son but. Cette définition de l'usage (et celles que nous allons présenter), se distinguent de la définition qui a cours en Sciences de l'Information, mais que nous allons utiliser dans la partie 6 de notre travail.

Afin d'éviter tout confusion, il est au préalable nécessaire de distinguer deux acceptions du terme « usage » dans le domaine des Sciences Humaines et Sociales.
Dans la tradition fonctionnaliste linguistique, l'usage caractérise la façon dont un outil, qui est la langue, est activé de façon à remplir une mission, qui est de transmettre des connaissances, de diffuser des informations, de fabriquer une représentation du monde (relativement à une certaine intention).
Dans une perspective qui est celle de l'usage des outils, il s'agit de caractériser comment des fonctionnalités sont utilisées, appropriées et interprétées de façon à susciter de nouvelles pratiques, menées d'activité (ou cours d'action[39]) et résolutions de problèmes. A ce moment-là, l'usage sera identifié autour du rapport entre un individu ou un groupe social et un dispositif.
Ces deux définitions sont utiles dès lors que notre matériau est symbolique et qu'il s'inscrit dans le cadre d'une activité.

Une communauté se définit alors par l'ensemble des personnes partageant le même usage d'une langue. Cette communauté se spécifie par des pratiques particulières, comme dans le cadre de la pharmacie hospitalière l'usage des noms propres pour ancrer en extension des collections de données. Ainsi, tout nom d'individu sert à assembler des données et construire un point de vue particulier sur ces données : un nom de bactérie ou de molécule, ou encore un nom de patient permettent d'assembler des données physiologiques, et plus précisément de

construire chacun un point de vue différencié sur de mêmes données. Un résultat d'analyse de prélèvement sanguin peut être interprété de façon complémentaire si l'on prend le point de vue du patient, de la molécule qui circule dans ce sang ou encore de la bactérie qui infecte le patient et que combat la molécule. Ainsi, une fonctionnalité linguistique comme la référence directe peut donner lieu à des usages spécifiques inscrits dans des activités d'information.

Le centre de cette distinction entre usage linguistique et usage communautaire est l'appropriation totale du langage par les individus quels que soient les apprentissages communautaires, et une disposition d'un groupe à élaborer des réponses concrètes à des problèmes et ainsi constituer un réseau de partage et d'apprentissage. Par exemple, l'usage du nom propre dans la pharmacie hospitalière est relatif à la fois à l'individualisation du patient (comme fondement épistémologique de la pratique scientifique du pharmacien) et au classement d'un ensemble de données (résultats d'analyse, prescriptions, correspondances échangées entre praticiens sur ce cas) répertoriées sur des feuilles de papiers indexées sous ce nom propre. Cette indexation papier est dupliquée à l'intérieur d'une base données où sont représentées et stockées toutes les données patient et molécule. Enfin, ces données assemblées sous un code sont classées à partir de la base de données dans une population statistique. Ainsi, si l'usage du nom propre consiste à associer un constituant symbolique à une valeur de vérité et à accréditer à cette relation l'application de propriétés et l'ancrage d'événements (par exemple), son rôle dans la constitution de dossiers constitue un usage fonctionnel classificatoire lié à une communauté de pratiques professionnelles. Alors le nom propre ne sert pas à référer mais à classer des données. L'usage s'ajoute à la fonctionnalité linguistique du nom propre et donc permet des pratiques sociales communautaires. Entre les « dossiers patients » et l'inscription de l'individu dans une population, l'usage est différent puisque dans le dernier cas, l'identification de l'individu est purement formelle et sans conséquence pragmatique pour ce dernier, ce qui bien évidemment n'est pas le cas du dossier patient. (Rappelons que l'on définit par population un ensemble défini et fini d'éléments observés qui sont tous de même nature ; ici en l'occurrence, il s'agit de données physiologiques et comportementales de patients, regroupés dans une base de donnée, et sur lesquels il s'agit de calculer statistiquement des probabilités de comportement d'élimination).

La distinction avec d'autres approches des usages en ergonomie, (surtout en ergonomie cognitive) s'opère sur un autre versant de la problématique : on ne s'interroge pas sur les modalités cognitives de cet usage ni sur les dimensions individuelles ou collectives de cette appropriation. En d'autres termes, on ne s'intéresse pas directement au comportement de l'usager final par rapport au dispositif et à la présentation de l'information. Par contre, les outils théoriques utilisés par cette discipline peuvent nous intéresser, notamment les théories de la cognition.

Caractérisation de l'usage : dimensions techniques et linguistique, du quotidien au web de données.

La caractérisation de l'usage que l'on vient de présenter concilie les propriétés du langage avec les fonctionnalités et rôles des outils d'information dans des activités professionnelles. Nous présentons maintenant de quelle façon ces propriétés relatives au langage et à l'activité peuvent s'inscrire dans un cadre de conception hétérogène. Nous donnons ici un cadre général, que nous détaillerons en 2.3.

Nous avons choisi de fonder notre caractérisation de l'usage à partir de la dimension linguistique parce que les régularités observées peuvent être abstraites dans des règles utiles

pour l'élaboration de dispositifs et d'outils. Cette capacité à abstraire et à formuler des règles prédictives est une longue tradition en linguistique et a permis la réalisation de toutes sortes d'outils. Néanmoins ces outils (comme les analyseurs syntaxiques et sémantiques) ne nous intéressent pas directement. Nous prenons comme point de départ des outils élaborés relativement à des usages communautaires, comme le sont les outils documentaires. Les communautés permettent alors de spécifier dans leur contexte professionnel particulier l'usage des termes, en lien à des pratiques et des techniques.
Dans le cadre des services web, les dimensions technologiques et celles de l'usage social sont également prégnantes. Dans la mesure où les usages communautaires des termes sont établis dans le cadre d'outils déjà existant (thésaurus, bases de données, ontologies) Un projet à propos du web de données peut très bien reprendre ces acceptions. Si les outils comme les jeux de métadonnées et les descriptions documentaires correspondent à des usages de la langue dans le cadre de processus professionnels, alors il apparait pertinent d'utiliser cette définition de l'usage. On peut alors y adjoindre le transfert des observations à propos de situations exemplaires observables aujourd'hui dans le monde.

Ainsi, la proposition dans laquelle on s'inscrit consiste à lier ces possibilités techniques offertes par les services (et l'accroissement de leur puissance liée aux grilles) à des outils plus anciens de description et d'organisation de l'information et des connaissances. (Les grilles constituent des infrastructures techniques et logicielles permettant le transfert coordonné de ressources, la résolution dynamique de problèmes dans des organisations virtuelles multi-institutionnelles. Les grilles coordonnent des ressources sans système de contrôle centralisé. Elles utilisent des interfaces standardisées, ouvertes et des protocoles généraux, de façon à délivrer un service à haute valeur ajoutée. L'infrastructure DARIAH, notamment, fonctionnera sur ces principes. Le programme européen CORDIS porte de nombreux projets relatifs aux grilles et à leur mise en œuvre). Ces grilles permettent un transfert massif de compétences et de connaissances depuis le monde des activités réelles vers le web. Avant même de parler des phénomènes que l'on transfert, on pose cette condition technique. Celle-ci est largement exploitée par les projets actuels d'e-science, dont nous reparlerons en partie 6.

Spécificités de la problématique de l'usage relativement au web.
Si l'on suit G. Antoniou et F. van Harmelen[40], les problématiques de l'usage dans le web de données posent des questions de l'usage de façon très différente de l'informatique et de l'IA traditionnelles notamment par rapport à la recherche d'information ou à l'interface homme-machine. En effet, le HTML, qui est le langage fondateur du web, est adapté à l'humain, à savoir que son usage est nettement plus aisé, intuitif et de sens commun que les langages traditionnels de l'informatique. De plus, les standards proposés évoluent de plus en plus vers la prise en charge possible du sens. Or, il faut bien que des outils sensibles aux contenus existent afin de faire émerger un web intelligent, capable de prendre en charge les connaissances et donc d'amener sur le web le plus grand nombre d'activités (ou tout simplement de parties d'activités). C'est dans cet enjeu-là que l'on inscrit notre projet.

En d'autres termes, les machines du web sémantique, XML/RDF, OWL (développées à la suite de HTML) visent à donner au web la capacité à gérer des informations et des connaissances. La problématique est en quelque sorte inversée par rapport à l'IA classique, puisque ce sont des langages adaptés à l'humain, comme le HTML, mais inadaptés aux raisonnements des machines, auxquels on ajoute des niveaux de structuration et de description supplémentaires, de façon à effectuer les raisonnements en question, qui commencent à être proposés et pour lesquels il est nécessaire de trouver des modèles. (Ainsi, on retrouvera les modèles de la cognition de Kintsch et Van Dijk pour l'annotation modulaire de documents[41], [

Wulf, J., Jorm, D., Casperson, M., & Newson, L. (2012, April). *Automated Assembly of Custom Narratives from Modular Content using Semantic Representations of Real-world Domains and Audiences*., pp. 66-78] ou encore le modèle sémiotique de Pierce chez A. Gangemi.

Ainsi, le web se structure par couches, depuis les données vers les connaissances, par niveaux d'abstraction successifs. A ces différents niveaux correspondent des langages spécifiques : HTML, XML, RDF et OWL. Néanmoins, les recommandations du W3C ne portent que sur des possibilités et restent neutres quant aux modèles sémantiques ou cognitifs utilisés. Ces derniers relèvent des choix des auteurs d'outils.

Si cette précision à propos du web est nécessaire, c'est parce qu'elle permet de caractériser la question de l'usage totalement différemment de la façon dont elle était pensée lorsqu'il s'agissait de rendre des systèmes complexes utilisables par des usagers novices. En d'autres termes, les rapports entre interfaces, bases de données et usager final sont transformés par la souplesse intuitive des langages du web. Cette intuitivité est accentuée par le développement des feuilles de styles (CSS) qui permettent de masquer l'ensemble des raisonnements derrière une interface conviviale.

Cette présentation fournit un cadre pour appréhender la question de l'usage et la façon dont on l'aborde. En effet, si le web sémantique caractérise des raisonnements, c'est d'abord de façon à assurer un accès à des données pertinentes dans le cadre d'une activité d'information. Les activités que l'on cherche à intégrer dans le web de données sont liées à l'information, au-delà des questions propres de recherche d'information. On peut intégrer la question de la recherche d'information à l'intérieur de celles des services web, mais si les services web fournissent de l'information, on ne peut les confondre avec les problématiques de la recherche d'information. Les services web peuvent être définis comme des fonctionnalités discrètes encapsulées à l'intérieur d'une interface accessible par des protocoles internes standards[42]. Cette définition est une caractérisation interne des services. Pour une présentation plus générale, on peut se référer à la définition donnée par leW3C[43], pour lequel le service web est d'abord un système supportant une interaction entre une instance proposant un service et une autre en quête de ce service. Ces instances sont au départ des humains, mais le travail de l'interaction est le fait d'agent ou automates chargés de transmettre ou de collecter les messages envoyés par les humains (qu'il s'agit de requêtes ou de ressources disponibles). Ces agents permettent l'accès aux ressources.

Caractérisation générale des services web.

On voudrait expliciter le fait que le web de données construit un contexte dans lequel s'articulent les questions d'activité et de langage. Cette activité n'est pas nécessairement intégrée à la machine, mais fonctionne de façon coopérative et finalisée, entre l'utilisateur et des données numériques hétérogènes et éloignées.

Le web de données se définit comme l'automatisation, sur le web, de certains processus, comme par exemple la recherche de services appropriés. Il s'agit de reproduire sur le réseau un raisonnement intégré dans une activité et de faire en sorte que le service puisse le proposer.

Dès l'origine, le web de données est considéré comme un cadre intégrant des outils différents, assemblés par certaines fonctionnalités. C'est par ces relations finalisées entre données que l'on construit les raisonnements par des représentations de plus haut niveau d'abstraction, utilisant des ressources appropriées. Pour une meilleure compréhension, on peut caractériser le web sémantique en considérant plusieurs niveaux par lequel il peut être appréhendé :

- tout d'abord le niveau des architectures (qu'il s'agisse des plateformes ou des infrastructures)
- le niveau des services (qui sont des fonctionnalités proposées à l'usager et insérées dans les architectures)
- le niveau des processus (caractérise les relations et inférences à l'œuvre dans les services. Il s'agit du fonctionnement interne des services)
- le niveau des technologies d'implémentation (qui intègre l'ensemble des langages de représentation et de structuration).

La question des services web implique de caractériser l'approche du langage au travers de plusieurs niveaux de pertinence ; sommes-nous dans le cadre de communications impliquant une dimension pragmatique, ou de connaissances communes, pouvons-nous nous mettre d'accord sur une terminologie ? Ces questions de niveaux mais aussi de fonctionnalités linguistiques sont liées à la façon dont on appréhende le lien et la distinction entre le travail de l'utilisateur et celui du service.
Nous proposerons de considérer le processus comme un cadre abstrait indépendant des spécialités et dont les objets peuvent être des catégories linguistiques ou cognitives. Par contre, les services sont relatifs à la mise en œuvre du processus dans le cadre d'un domaine précis, et donc utilisent le vocabulaire de ce domaine. Enfin, les interactions avec l'utilisateur sont caractérisées dans l'interface associée au portail de l'infrastructure ou de la plateforme.

Si l'on reprend notre seconde caractérisation de l'usage, il marque la relation entre l'activité et le service web. Ils impliquent une contextualisation et des cadres d'action qui peuvent être caractérisés dans le cadre des modèles d'ontologies (comme ceux définis dans les vocabulaires comme DOLCE). Ils doivent rendre compte d'états, de faits et d'événements.

Il existe des sortes très différentes de service web et les structurer constitue une tâche qui a été menée dans le domaine biomédical[166]. Les auteurs identifient 2486 services et 176 providers de ces services. Par ailleurs, les technologies associées à ce service peuvent être très différentes, depuis des outils de relation entre des bases de données jusqu'à des outils d'annotation. Le service web se caractérise donc par l'affichage d'une fonctionnalité au sein du web quelles que soient par ailleurs les technologies et les procédures qu'il implique. Le web service se distingue d'un schéma par le fait qu'il s'agit d'une fonctionnalité établie et non d'un outil mis à disposition pour des usages déterminés par l'utilisateur. En ce sens, un schéma constitue un outil dont les usages sont définis en partie par l'usager alors qu'un service possède un usage défini et stable.
Les deux possibilités peuvent être considérées comme complémentaires. Par ailleurs, à l'intérieur même des services web, on se propose de ne considérer que ceux qui reposent sur une mise en relation de ressources web et sur une caractérisation de ces relations utilisant une ontologie.

On ne peut se contenter de considérer les plateformes et les infrastructures accueillant les services d'une part, et les moteurs de recherche d'autre part. L'ambition d'un certain nombre d'outils est effectivement une meilleure accessibilité des services du web, leur découverte mais également leur utilisation[44]. L'objectif est alors de rendre compréhensibles et utilisables des services. L'ontologie OWL-S a comme rôle de connecter ces services de façon à assurer leur complémentarité[45]. Du point de vue de l'utilisateur, elles devront se présenter comme une offre cohérente et homogène. Pour cela, l'effort porte principalement sur l'interopérabilité, au

[166] http://www.biocatalogue.org/

niveau technique, et sur le contrôle des mises en relation et traduction, au niveau conceptuel. Les web services comportent une dimension d'action qui n'est pas présente dans le cadre des outils de description que l'on a vu jusqu'à présent. En ce sens, dès lors que l'on propose un raisonnement à partir de données structurées hétérogènes du web, un web service devient un outil fondamental pour produire un raisonnement régulier. Néanmoins, ce web service peut comporter ou pas une base de données et un stockage des résultats.
Le web service permet de mettre aussi en évidence le fait que les outils du web comportent un raisonnement qui s'applique à des représentations symboliques de différents niveaux d'abstraction.

Dans le cadre du web de données, les ontologies ne constituent qu'un certain type d'outil nécessaire à la mise en œuvre d'un service, important certes, mais qui ne peut être réduit à cette ontologie. Par exemple, un outil comme WSMO (Web Service Modeling Ontology (WSMO)), qui constitue un cadre conceptuel pour la description de services web, contient quatre éléments de base :

- les ontologies, qui produisent les terminologies utilisées par les autres éléments de WSMO,
- les services web, qui permettent d'accéder aux services qui, en échange, prennent une certaine valeur dans un domaine,
- les buts, qui représentent les désirs des utilisateurs,
- et les médiateurs, qui gèrent les problèmes d'interopérabilité entre les différents éléments de WSMO. Ces problèmes peuvent se poser à différents niveaux de compatibilité : protocoles (communication entre services web), processus (combinaison de services web), et données (difficultés entre terminologies).

Cet exemple illustre les différentes dimensions de l'offre d'un service et surtout l'organisation nécessaire pour rendre opérationnelle cette offre. Il laisse entrevoir comment la question linguistique est appréhendée. En effet, les dimensions lexicales et terminologiques sont aussi essentielles que les aspects conceptuels. D'où la possibilité d'envisager plusieurs niveaux de caractérisation des unités symboliques. Développer un service web sur la base de descriptions documentaires et plus globalement de document constitue une façon d'ancrer un schéma dans une structure pérenne et qui permette de recueillir et d'enrichir par l'expérience le schéma que l'on propose.

Enfin, comme le service constitue une médiation, il comporte une économie[167] et par conséquent transforme les pratiques. C'est justement l'un des enjeux essentiels de l'e-science. Par conséquent, la mise en réseau de services ne fait pas que déplacer des procédures et des usages ; les pratiques associées au document numérique sont elles-mêmes transformées par de nouvelles mises en relation qui produisent elles-mêmes de nouvelles pratiques. Les schémas ne constituent que des outils à vocation « interne », c'est-à-dire à disposition des gestionnaires de plateforme et des producteurs de services. Par contre, les services web sont dirigés vers les usagers du web et ont donc un usage prescrit bien plus établi.

Concepts et langue ; quelle est la place du langage dans les problématiques des services du web ?
Si l'on se penche maintenant sur les structures de symboles utilisées dans le cadre du web sémantique, on trouve de façon sommaire les ontologies et les lexiques, avec de nombreuses variantes, comme les ontologies terminologiques ; dans les ontologiques, la signification est

[167] Nous ne faisons que mentionner cet aspect fondamental. Nous laissons à d'autres le soin de le développer.

essentiellement considérée comme conceptuelle. Les processus (de classification, de catégorisation) sont caractérisés comme des constructions mentales, suivant en cela la tradition de la psychologie cognitive et de la sémantique formelle[46].

Dans cette dernière partie concernant une caractérisation générale des usages, nous nous penchons sur l'outil essentiel qui permet de construire les outils du web, à savoir les ontologies. Les ontologies constituent une modélisation des connaissances : une activité à l'intérieur du web de données se caractérise d'abord par le maniement de connaissances. C'est pour cette raison qu'il est utile d'introduire ici les ontologies. Elles pourront être ensuite développées dans toute leur complexité. Nous nous limitons ici à comprendre en quoi les connaissances n'ont pas à être considérées en dehors d'une sémantique plus générale (et en partie linguistique).

Les ontologies caractérisent le niveau d'organisation des connaissances sur le web. Ce niveau intensionnel permet de mettre en relation des structures de données pouvant elles-mêmes contenir des connaissances. Une ontologie est d'abord construite pour caractériser à un haut niveau d'abstraction, un lien entre des structures de données distinctes. Par exemple, la GENEONTOLOGY constitue un système complet permettant d'accéder à l'ensemble des données déposées relatives au domaine de l'ingénierie du génome.

Cette ontologie est fondée sur les principes de structuration de concepts définis par B. Smith. L'intérêt de cet auteur, c'est d'avoir élaboré et utilisé une théorie des ontologies fondée sur des principes aristotéliciens[47]. Ces principes sont nécessaires pour traiter des données complexes, notamment les documents médicaux, parce qu'ils impliquent de nombreux éléments du langage naturel. Il présente quatre contraintes que doit satisfaire une ontologie :

1. Une connaissance des termes et de la façon dont ils sont utilisés dans la langue naturelle ;
2. Une connaissance du monde et, surtout de la façon dont les référents des termes sont en interrelation dans le monde et dans certains types de contextes ;
3. Un algorithme qui soit capable de non seulement de calculer la représentation d'un usager, à propos d'une portion du monde, mais également
4. de définir les façons par lesquelles on exprime ce qui n'existe pas dans le monde.

Une ontologie caractérise ici ce qui doit exister (et non pas ce qui peut exister pour une communauté d'acteurs). En ce sens, l'ontologie se distingue de la terminologie parce qu'elle se préoccupe d'entités et non de termes. Par ailleurs, dans une ontologie, il y a un rapport d'instanciation (à savoir que les concepts universaux se réalisent dans des instances) en plus de rapport entre génériques et spécifiques, à savoir des relations complexes de termes. Enfin, les objets utilisés par l'ontologie perdurent dans le temps, alors que les processus sont ancrés sur la linéarité temporelle.

Une autre question essentielle est celle de la granularité des partitions, à savoir la façon dont les rapports partie-tout (ou relations d'ingrédience) sont caractérisés. Les partitions sont considérées par B. Smith comme des inventaires et des cartographies de certaines parties de la réalité. Les partitions donc des artifices de la cognition et servent à distinguer de façon constante des zones dans lesquelles des informations peuvent être enregistrées.

L'intérêt de la position de B. Smith est qu'elle distingue les dimensions linguistiques, les représentations de connaissances et les ontologies. L'analyse linguistique sert à déterminer des emplois de termes dans le cadre des documents sur lesquels se fonde l'ontologie. L'ontologie doit respecter les constructions grammaticales et de cette façon, résoudre les problèmes d'ambiguïté. Cette relation entre questions linguistiques et de connaissances fondent les positions théoriques et méthodologiques concernant les ontologies : les propositions de B. Smith sont fondées sur une transparence relationnelle, et donc une complémentarité, qui selon lui, n'est possible que grâce au « réalisme ». Les attaques de G. Merrill[48] sont centrées sur la simplicité de cette relation.

D'autres modèles ont été élaborés, notamment DOLCE. Il constitue un cadre dans lequel un certain nombre de concepts fédérateurs sont retenus, et peuvent être adaptés à des contextes différents. Néanmoins, N. Guarino considère essentiellement les unités langagières dans leur dimension conceptuelle, sans lien direct aux dimensions linguistiques. (Cela ne signifie pas que ces dimensions ne sont pas considérées comme pertinentes dans le cadre d'une ontologie de domaine. Elles ne sont pas considérées comme telles pour la construction de l'ontologie générique proposée par DOLCE).
A contrario, CYC propose une ontologie générique fondée sur la dimension linguistique.

En définitive, les liens entre dimensions linguistiques et conceptuelles sont une part essentielle de l'évolution du web sémantique, et plus particulièrement relativement aux questions documentaires. A ce titre, la mise en place de cadres de travail comme LEXINFO constitue un indice essentiel de l'intérêt de la communauté scientifique pour cette articulation. Celle-ci, rappelons-le, est fondée sur le fait que l'usage (à savoir la façon dont l'utilisateur appréhende le service dans le cadre de son activité) mobilise à la fois les dimensions linguistiques et conceptuelles des unités symboliques. L'articulation entre ces deux dimensions est un fondement de la préservation de la garantie d'usage du web mais également du lien entre les structures organisant l'information (et notamment la description des documents) et les documents eux-mêmes.

2.2. Caractérisation sémantique de l'usage : langage et activité.

Nous revenons maintenant sur la question de la dualité entre les dimensions linguistiques et de l'activité, que nous avons déjà évoquées plus haut, et que nous aimerions développer ici de façon un peu plus élaborée. Nous essaierons de les articuler tout au long de cette partie.

Cette partie du travail repose sur l'hypothèse qu'il y aurait tout intérêt à ce que le web de données s'intéresse aux acceptions référentielles de la sémantique, et que cette dimension référentielle doit être articulée à des représentations conceptuelles nettement plus abstraites. On suit et surtout on accentue les principes des ontologies que l'on a présenté plus haut : une structure conceptuelle est fondée sur une caractérisation lexicale et possède donc une propriété référentielle. L'hypothèse que l'on reprend est que les systèmes permettant la circulation de l'information requièrent des constructions du monde de la référence pour organiser leurs données symboliques et satisfaire le transfert et une interprétation conformes à l'attendu. Comme nous aurons encore l'occasion de le préciser, dans l'adaptation de posologie, et donc dans notre modèle, l'information se définit comme ne pouvant être ambiguë, et son interprétation ne peut être aléatoire : ici l'information sert à déterminer des stratégies thérapeutiques (à la fois en termes de quantités et de régularités). Elle est donc directement impliquée dans l'action. La circulation de l'information ne peut être envisagée sans la dimension de la construction du monde de la référence.

Caractériser la signification par le contexte d'usage fonctionnel consiste à identifier et caractériser des contraintes sur le choix et l'organisation des unités linguistiques.
Or, dans le cadre du web sémantique, comme dans nos exemples, ce contexte fonctionnel est régulier (donc modélisable). Le modèle des flux s'appuiera sur ces régularités de contexte, et donc on pourra conjoindre une représentation des flux et une représentation de la structure d'information. En effet, ce contexte de production d'expressions informationnelles (qu'il s'agisse des protocoles de circulation d'information dans les hôpitaux, ou de l'information en gare) constitue un cadre régulier d'usage du langage. Ce contexte fait partie de la modélisation de l'information puisque c'est dans ce contexte que l'information est pertinente à la fois en production et en interprétation.

Enfin, précisons que cette caractérisation ne concerne que l'assemblage de protocoles réguliers d'information et d'unités du langage. En ce sens, on ne prédit rien directement sur la nature du langage, mais sur des contraintes au discours, dues à des dispositifs techniques.
Nous éclaircissons maintenant notre position par rapport à la dimension linguistique, en spécifiant nos centres d'intérêt dans ce domaine. Comme par ailleurs l'articulation entre les dimensions d'activité et de langage est un objet d'étude de la psychologie cognitive, nous présenterons certains aspects particulièrement pertinents de cette discipline et qui permettront de considérer plus précisément le lien entre conceptualisation et référence. Surtout, ces propositions seront utiles pour caractériser des modèles prédicatifs conceptuels. Nous présenterons alors le cadre théorique général de notre analyse de l'activité et nous terminerons par une caractérisation de la façon dont le lange peut être appréhendé dans notre cadre, à savoir sur le principe de la sous-spécification.

2.2.1. Caractérisation sémantique de l'usage : notions de corpus et question de l'autonomie du langage.

En linguistique, on n'observe le langage qu'à partir de ses réalisations : le langage constitue un système mais il ne peut être appréhendé que par la façon dont ses règles se réalisent. Or, justement, les possibilités de réalisation sont nombreuses et diverses.
Face à cela, les linguistes ont élaboré des corpus de plus en plus fournis de données linguistiques numérisées, accompagnées ou pas de contextes (données audiovisuelles notamment), tels que les échantillons utilisés soient les plus représentatifs de la langue dans sa globalité. Ces échantillons peuvent être envisagés soit par une couverture exhaustive des différentes sortes de réalisations (comme le British National Corpus par exemple), soit par l'intégration dans le corpus du plus grand nombre de paramètres linguistiques ou para linguistiques (silences, gestualité, expressions faciales, onomatopées) intervenant dans la cette réalisation linguistique. Les corpus recueillis et transcrits par l'équipe ICAR[168] en constituent un bon exemple.
Ces corpus, de plus en plus fournis et précisément annotés, constituent des bases d'études qui visent à contrecarrer l'aléatoire des exemples issus des textes littéraires ou des conversations ordinaires. Pour la linguistique, ils ont fourni les moyens et des méthodes pour une analyse plus rigoureuse des réalisations.

Nous ne suivons pas ces principes méthodologiques. En d'autres termes, on ne peut prétendre à une validité linguistique au sens où notre travail ne pose pas principalement un problème de

[168] Ces corpus sont répertoriés dans la base de données CLAPI, accessible en ligne, et sont décrits en utilisant à la fois les métadonnées Dublin Core et les marqueurs TEI.

langue ou de discours, mais une question liée à un usage du langage dans un cadre fonctionnel précis, celui de la description informationnelle d'objets. (Cela dit, le propos est à nuancer parce qu'étudier les discours revient à étudier la langue dans ses différentes fonctions, notamment informationnelle ; donc, il n'est pas toujours aisé de distinguer l'étude des réalisations de la langue et celles de l'information). En d'autres termes, on n'hérite que partiellement des questions de la linguistique parce que l'on ne s'intéresse qu'à une fonction discursive. Par ailleurs, en tant que fonction sociale, elle dépasse le cadre des questions de linguistique.
En apparence même, nos vocabulaires contrôlés et nos langages normés ne laissent guère de place à la langue naturelle. Les couples attribut-valeur des métadonnées et les classifications d'unités lexicales dans des concepts apparaissent relativement éloignés du discours commun. Néanmoins, on peut penser que ces outils mettent en œuvre des processus de la langue, mais que cette mise en œuvre est solidement encadrée par un contexte technologique strict, évitant au maximum les ambigüités. L'hypothèse de la sous-détermination permet de fonder cette approche : le contexte spécifie la signification des unités linguistiques. Les vocabulaires contrôlés ne seraient pas tant des structurations abstraites et fonctionnelles utilisant des unités linguistiques que des contraintes contextuelles fortes posées sur les fonctionnements de la langue naturelle.

Dans la théorie linguistique, la langue constitue un objet d'étude abstrait et qui repose sur un postulat fondateur de la linguistique : le langage serait une dimension plus ou moins autonome de la pensée. Ce postulat fondateur est en soi un problème puisqu'une part importante de la signification provient du contexte et qu'il est difficile dans un cadre linguistique de prendre en compte cette dimension. Cette remise en cause sera initiée par H. Putnam, reprise par John Perry notamment. Chalmers[49] développe cette idée dans « l'esprit étendu ». Ce travail vise à réfléchir et à conceptualiser les perspectives ouvertes par la cognition située et distribuée, à savoir que la pensée (et le langage) ne seraient pas autonome dans leur fonctionnement par rapport à leur environnement. Toute réalisation linguistique ne saurait être expliquée que dans son contexte. Et ce contexte n'est pas statique ou décoratif ; il participe d'un processus, ou raisonnement, dans lequel le langage intervient parmi d'autres constituants. Au-delà des questions de dimension de la cognition, le problème est celui de sa nature symbolique et calculatoire. En psychologie cognitive, une telle perspective est également relativement envisagée.
Cette perspective permettrait de proposer des objets d'étude hybrides, construits autour de dimensions disciplinaires différentes. Or, ces positions sont tenues dans le domaine de la philosophie (où elles sont validées selon les méthodes propres de la philosophie). Elles sont corrélées à des études d'anthropologie cognitive et ne donnent pas encore lieu à des propositions méthodologiques concrètes si ce n'est en anthropologie cognitive.

2.2.2. Autres caractérisations sémantiques de l'usage : psychologie du langage, psychologie cognitive.

La psychologie cognitive, qui ne part pas du postulat de l'existence du langage comme objet autonome, a donné lieu à des hypothèses remettant en cause la notion de règles formelles associées au langage, et l'idée de règles propres au langage. Les tenants de la linguistique de corpus, dont nous venons de parler, partagent à peu près ce point de vue ; selon eux, il n'y aurait aucune convergence entre d'une part les résultats de l'analyse du plus grand nombre d'occurrences et de la diversité des contextes d'usages d'une entité linguistique, et d'autre part la possibilité de règles formalisables relativement simples et explicites, sensées en rendre compte.

Le problème est que les outils des perspectives scientifiques psychologiques, et plus précisément de psychologie cognitive sont expérimentaux : les données sont recueillies par le biais de protocoles très contrôlés ne peuvent être assimilés à un usage en situation naturelle. Cette objection ne nous concerne pas directement dans la mesure où ce qui nous intéresse est une caractérisation du contexte en relation au langage et non le fonctionnement du langage dans la pensée. Cette caractérisation du contexte explicite pourquoi et comment telle entité prend un sens spécifique dans ce contexte[169].
On trouve dans ce courant la problématique des « frames », telle qu'elle est développée par L.W. Barsalou[50]. Il s'intéresse à des cadres mentaux qui constituent des relations schématiques entre des objets permettant entre autre de classer des entités. Il s'agit ici d'une psychologie cognitive de la connaissance. L. W. Barsalou[51] intègre un certain nombre de concepts de la cognition « incorporée » comme de la cognition « située » (à fondement méthodologique plutôt interactionniste, donc plus globalement anthropologique), à l'intérieur de la psychologie cognitive.
L'unité de travail en psychologie cognitive n'est plus limitée à un concept, mais est étendue à une situation (qui représente le concept au travers de relations entre des propriétés et des valeurs). Celle-ci est une région de l'espace perçu, qui dépasse la durée pendant laquelle s'exerce la perception d'un acteur ; par ailleurs, la situation peut contenir des états mentaux[170] variés.
Le concept est alors caractérisé par une relation entre un prédicat et des individus ; cette relation opère par l'assignation d'attributs aux valeurs instanciées (et non, comme dans la théorie des prototypes, par des listes de traits). Les concepts sont liés dans des frames, qui représentent un niveau de structuration supérieur par rapport au concept.
En ce sens, les frames sont structurés par un rapport valeur/prédicat ou attribut. Les concepts sont ainsi structurés comme des propositions, ce qui permet d'intégrer la dimension informationnelle par l'apprentissage.

Ce qui est intéressant ici, c'est que l'on peut proposer une définition relativement stable des scènes décrites dans notre exemple présenté en annexe 1. Les situations reliant différents concepts dans la construction du frame, l'annonce en gare associe les concepts de trafic ferroviaire, de train, d'horaires et d'aléa de façon à construire une situation qui est partagée entre l'ensemble des partenaires en interaction. Cette situation permet de faire circuler l'information entre des agents éloignés sans aucune ambiguïté. Les scènes décrites constituent des instants du quotidien alors que les frames sont des constructions mentales ; ces derniers s'appliquent sur des scènes empiriques. Les concepts sont définis comme des unités accumulant en mémoire de l'expérience à propos de types de phénomènes récurrents. Ils constituent donc une catégorisation abstraite des phénomènes du monde et placent les classifications des phénomènes perçus au centre du raisonnement.

Pour en revenir à l'usage, ce que l'on retient ici, c'est le fait que l'on peut caractériser un concept relativement à son usage dans un cadre défini par un frame. L. W. Barsalou ne caractérise pas de façon fonctionnelle les situations, bien qu'elles soient analysées comme un

---

169 On pourra toujours objecter que le contexte ne peut être appréhendé que parce qu'il est construit scientifiquement, et donc, qu'on le dénomme cadre, frame ou situation, qu'il soit social, mental (sous forme d'affordance par exemple), ou caractérisé par un domaine de connaissance, il constitue une réduction de la réalité. Au moins, cette réduction est mentalement, linguistiquement ou socialement plausible.

170 Les états mentaux constituent de façon très globale des dispositions de l'esprit vis-à-vis de propositions qui peuvent être formulées. Ils sont traduits par des verbes modaux (croire, penser, estimer, etc.). Il s'agit d'une problématique essentielle de la logique propositionnelle et de la philosophie de l'esprit.

gain pour la performance (le raisonnement, l'apprentissage, etc.). C'est surtout relativement aux classifications, à l'organisation de l'expérience et donc au traitement de l'information, qu'elles apparaissent fondamentales : les frames permettent de construire des propositions qui à la fois vont s'appliquer et s'enrichir par la succession des expériences

Ainsi, quel que soit l'éloignement méthodologique entre la psychologie cognitive et les sciences de l'information, la question de l'information est associée à la classification d'une instance et un apprentissage. Un signal devient une information par son association à un attribut dans le cadre d'un raisonnement tout en conservant son statut d'instance.

La psychologie cognitive nous permet de corréler des unités du langage et des processus cognitifs. Elle permet aussi de fonder dans la cognition des phénomènes de classification et de structuration d'entités que l'on associe généralement à des opérations techniques. Ces dernières ne constituent alors que des externalisations normées de processus cognitifs nettement plus généraux.

Cette remarque nous permet de fonder également l'hypothèse selon laquelle les processus professionnels comme la classification ou le renseignement de métadonnées ne constituent pas des inventions propres de la profession, mais seulement la technicisation de processus cognitifs bien plus généraux.

2.2.3. Caractérisation des usages comme pratiques en lien à l'activité cognitive.

C'est du côté des sciences fondées sur l'observation que l'on peut se tourner pour caractériser l'usage, l'adaptation et l'appropriation afin d'établir un lien plus marqué avec les exemples que l'on utilise. En effet, dans l'acception psychologique, c'est essentiellement dans le lien à la mémoire (à laquelle sont liées en psychologie, la classification et la catégorisation) que les situations sont définies. Si l'on veut rendre compte du rôle actif des objets du monde[171], considérés au sein de l'activité cognitive, c'est du côté de l'anthropologie cognitive distribuée qu'il faut le chercher. Ainsi, on éclaircit les propos tenus précédemment sur ce courant de recherches.

Hutchins[52] a proposé un cadre conceptuel relativement complet pour caractériser le fait que l'activité cognitive serait distribuée entre les humains et les outils, d'une part, et située dans le monde d'autre part.

L'idée d'Hutchins, c'est d'utiliser les trois niveaux de la perception proposés par D. Marr[53] non plus dans un contexte de reconnaissance d'objets mais d'orientation dans l'espace, dans le cadre d'une activité de navigation - maritime dans ses premiers travaux, aérienne ensuite (et dans le cas de la navigation maritime, au travers de deux réalisations culturelles et technologiques très différentes : l'orientation des micronésiens entre leurs différentes îles et le guidage d'un porte-avion de l'armée américaine dans la baie de San Diego en Californie). L'orientation est une activité cognitive reposant sur l'externalisation des opérations mentales

171 Par exemple, en pharmacie hospitalière, patient, molécules et bactéries constituent les trois « objets du monde » construits par l'activité. La production, la circulation et la représentation de l'information sur un support consistent à caractériser l'évolution non prédictive de l'interaction entre ces différents individus : évaluation des risques d'accumulation, évolution de l'infection, l'ensemble étant appelé « rapport efficacité/toxicité ». Les individus sont des objets du monde dont le comportement ne peut être prédit avec certitude.

de reconnaissance. Donc, il distribue les différents traits pertinents nécessaires à l'orientation dans le monde et les lie dans le cadre d'un raisonnement qui permet de savoir où l'on est et où l'on va :

- les traits sont répartis dans l'univers. Ces traits sont les étoiles, la profondeur, la couleur de la mer, la position du soleil dans sa trajectoire, etc. Hutchins appelle ce niveau, le niveau computationnel (puisqu'il s'agit de calculs ne nécessitant pas de représentation) ;
- l'assemblage des traits produit une représentation en 2 1/2 dimensions[172] où le sujet qui s'oriente est au centre. De cette façon, il se situe par rapport au point antérieur et au point vers lequel il se dirige. L'assemblage des traits permet de positionner le sujet par rapport à sa position antérieure et sa position future en fonction de sa direction. Hutchins appelle ce niveau, le niveau d'organisation local (parce qu'il est relatif à la position du sujet et à la dynamique de l'opération). Il est algorithmique parce qu'il est caractérisé par des règles s'appliquant sur une constellation de traits issus des premiers calculs ;
- le passage à la troisième dimension consiste à se représenter où l'on est dans l'espace, de façon déconnectée par rapport à sa propre position. Cette troisième phase permet alors de déterminer si la direction réelle est adéquate par rapport à celle que l'on prévoyait. Hutchins appelle ce niveau le niveau de la tâche.

Ces opérations sont réalisées en continu par le navigateur (le pilote, ou l'homme de quart) quel que soit son lieu géographique, sa culture, sa langue. Les choses vont devenir un peu plus subtiles dès lors que des outils vont se développer pour aider le navigateur à effectuer ces opérations. Dans le premier niveau (le niveau de calcul) les outils que les cultures vont développer sont les outils de mesure. (Par exemple, une opération de calcul de concentration de molécule dans un prélèvement sanguin, ou une opération de pesage sont des protocoles techniques de ce premier niveau. Il s'agit d'isoler des traits pertinents et de les relier entre eux). Ce niveau est également appelé calculatoire parce que les opérations réalisées sont systématiques et constantes.

Le second niveau permet de construire des représentations en deux dimensions : ce sont les cartes et les supports qui permettent de considérer l'objet sous différentes dimensions et différents points de vue. (Ainsi, la table de travail du pharmacien s'inscrit dans ce niveau, comme la carte maritime). Ici, les outils traitent l'information afin de donner lieu à des scènes. Les scènes sont, dans un cadre constant, les représentations de phénomènes individuels.

---

172 Pour expliciter ce terme, il faut revenir à la théorie de la perception, il caractérise trois niveaux d'activité cognitive qui aboutissent à l'identification d'un objet symbolique :

- l'œil perçoit des traits locaux qu'il assemble et organise. On a par exemple des lignes, des angles, des points et des contours.
- Cette organisation produit une représentation en 2 1/2 dimensions : il s'agit de la visibilité de l'objet à partir de l'ensemble de ses points de vue. (On ne voit pas l'objet dans son entier mais à partir de l'ensemble des points de vue que l'on peut avoir sur lui). Ce traitement est préalable à l'élaboration d'une représentation globale de l'objet.
- Enfin, la représentation en 3 dimensions se situe au delà de la perception par le sujet de l'objet. Le niveau en trois dimensions caractérise l'objet dans sa totalité par un système de coordination d'axes de références qui spécifie les relations entre les parties de l'objet. (Les différents points de vue sont coordonnés de façon à élaborer une représentation détachée des points de vue du sujet). De cette façon, on construit une représentation de l'objet qui est une abstraction par rapport au point de vue du sujet. On peut alors passer au niveau représentationnel.

Le troisième niveau donne lieu à des outils comme le pilotage automatique. (Dans l'adaptation pharmaceutique, c'est le logiciel d'estimation et de contrôle des posologies). Il s'agit de l'ensemble des outils permettant de réaliser une part de l'activité parallèlement à l'acteur, en lui évitant de réaliser des opérations régulières liées à sa tâche : elles ont par exemple fonction de mémoire, de dépôt d'expérience, etc.

Les outils sont des objets matériels servant à produire des symboles : ces derniers sont les résultats des opérations réalisées par ces outils. Ces outils constituent le contexte et le caractérisent comme un processus régulier se déroulant entre humains et machines[173]. Cette coopération coordonnée construit alors un objet complexe autonome, défini comme une unité cognitive. Hutchins intègre dans cette unité un système d'information, ce qui permet aussi d'abstraire les données afin de caractériser ce système dans sa plus grande généralité. Et donc, de l'abstraire du cadre de la navigation (aérienne ou maritime) pour s'appliquer à d'autres domaines.

Le travail de G. Hutchins amène à opérer une distinction très précise entre les outils de production de l'information, les systèmes de transmission et les médias de représentation. En premier lieu, la structuration d'un univers par une culture ne donne pas lieu à symbolisation dans sa première phase : elle est marquée par des articulations de contraintes. En second lieu, une opération ne repose pas systématiquement sur un algorithme mais sur des phénomènes de connexion : le calcul du déplacement est la simple mise en relation de deux positions connues. Le résultat de l'opération est diffusé sous forme d'information ; on associe alors un système symbolique à l'opération. C'est le deuxième niveau dont on a parlé précédemment. Ce résultat symbolisé est inscrit sur un artifice de représentation. L'information n'est pas un artifice de représentation mais une de ses composantes ; il s'agit d'un certain format régulier de représentation reproduisant les unités symboliques liées aux calculs. Pour qu'elle puisse être interprétée, il faut que cette information s'inscrive sur un support (qui lui est un artifice de représentation) qui reproduise certains paramètres de l'objet sur lequel elle est indexée. Un artifice de représentation est une construction symbolique destinée à valoir pour un phénomène du monde afin de réaliser des opérations sur cet objet qui ne pourraient l'être sans ce support. (L'exemple de la carte est le plus commun). L'artifice de représentation restitue la dimension spatiale des phénomènes du monde par analogie. L'artifice permet de faire des calculs à propos de notre place dans le monde et de notre trajectoire en lui de façon plus précise et dans un plus large spectre temporel que les calculs utilisant les seuls éléments et objets naturels. Enfin l'artifice recèle des outils pour produire des opérations sur le monde. Ainsi, un artifice de représentation choisit ses traits pertinents en fonction des opérations menées dans l'activité (à savoir en fonction de la tâche).

Les artifices de représentation deviennent des outils de médiation dès lors qu'ils se définissent par trois propriétés spécifiques :

- il s'agit d'abstractions du monde (ces abstractions peuvent concilier différents modes de représentation, symbolique ou analogique, les associer, etc.)
- Ils permettent de reporter certains résultats d'opérations, mais également de réaliser des opérations ne pouvant être menées à bien qu'à un certain niveau d'abstraction par rapport au monde.

[173] On notera ici la différence de définition du contexte entre la psychologie cognitive et l'anthropologie cognitive située : dans un premier cas, il s'agit d'une construction relativement statique qui stabilise l'environnement extérieur, dans le second cas, d'un processus régulier qui construit les outils et les compétences nécessaires à sa réalisation. Le contexte est alors actif et constitue la motivation de l'ensemble de l'activité.

- Ils permettant enfin de rendre publics un certain nombre d'états de faits connus individuellement mais impactant la menée de l'activité finalité collective. Ils ont donc également une capacité mémorielle et une intégration dans le travail d'équipe.

Ce qui sera fondamental pour nous dans l'approche de G. Hutchins, c'est la distinction opérée entre l'information et le support de médiation. L'information constitue l'ensemble structuré des entités symboliques dupliquées entre l'espace d'élaboration et celui d'interprétation, le support de médiation caractérisant alors le dispositif qui permettra de structurer et distribuer les unités symboliques dupliquées lors du transfert.

Cette distinction sera essentielle surtout lorsque l'on abordera les questions de langage puisqu'elle permet d'envisager la problématique de l'information distinctement de celle des supports de médiation. En d'autres termes, on distingue très précisément les procédures de discours et les transferts d'information[174]. Dans le domaine linguistique, cela permettra d'argumenter une distinction forte entre les théories de la structure d'information fondée sur la rhétorique et les nôtres. Mais surtout, cela implique de ne pas organiser la structure d'information sur la base des seules réalisations en discours. Au contraire, dans le discours, l'information est disséminée parce que cette information est inscrite dans une activité dont le discours est constitutif. La présentation de l'information dépend d'autres intentions que celle d'informer qui influencent la structure et les enchaînements dans le discours.

Les applications des propositions de Hutchins se situent du côté de l'interaction homme-machine. En effet, il explore de quelle façon il est possible d'articuler au mieux, dans une situation et une tâche précises, les compétences cognitives de l'homme et les capacités des machines. La question du langage est alors occultée, réduite à la circulation de symboles, comme dans le modèle de Shannon[54] par exemple.

Néanmoins ces perspectives diffèrent fortement de nos objectifs sur deux autres aspects essentiels :

- l'environnement analysé (la circulation d'information dans un poste de pilotage, ou dans un aéroport, une salle de contrôle du trafic des métros) n'est jamais considéré comme un exemple d'un modèle, mais comme un objet inscrit dans une problématique plus générale de la cognition (sociale notamment) ; à l'opposé, si nous partons de situations du monde réel des publications et des ressources utilisées par celles-ci), c'est pour élaborer un modèle destiné à concevoir un outil. L'analyse de la cognition n'est pas une fin en soi mais un fondement théorique pour des propositions dans le domaine des bibliothèques numériques.
- il n'existe pas de finalité à ces travaux autre que la connaissance scientifique. En d'autres termes, les modèles obtenus caractérisent des phénomènes cognitifs et on ne déduit pas de modèles transposables de ces études (on se limite généralement à des conseils et des recommandations). Autrement dit, la dimension explicative n'est pas suivie d'une généralisation d'application en dehors d'hypothèses sur la nature et le fonctionnement de la cognition. C'est ensuite à la communauté des Sciences Cognitives de transmettre ces propositions à d'autres disciplines, qui vont les utiliser. A l'opposé, nous ne proposons Pas d'enrichir ces modèles, mais seulement de les considérer comme un fondement essentiel pour fonder le raisonnement distribué sur des dimensions culturelles (et pas seulement logiques et mathématiques). Ce fondement est pour nous essentiel si l'on veut que notre proposition soit celle d'un

[174] Un plan d'article scientifique est une procédure de discours et ce qui est décrit d'une ressource utilisée dans le cadre de cet article est de l'ordre du transfert d'information.

outil adapté à l'utilisateur. Ce raisonnement reproduirait en l'externalisant un raisonnement humain.

Sauf à considérer que ce qu'il est possible de décrire constitue une instance d'un processus beaucoup plus général (mais sans avoir le moyen de prouver quoi que ce soit), et donc comme nous le faisons, d'élaborer un modèle, il est difficile d'utiliser de tels outils.
En effet, les travaux d'anthropologie cognitive située et distribuée notamment, n'ont pas comme intérêt la conception dans le domaine et dans le contexte qu'ils traitent, mais dans d'autres champs disciplinaires, notamment la robotique. Là, l'expérience dont l'anthropologie cognitive rend compte et la modélisation qui en émerge peut donner lieu à la conception d'outils. En effet, les modèles cognitifs établis à partir d'une situation naturelle sont utilisés dans une même situation, mais avec des agents artificiels. La contrainte est que dans les deux cas, on a affaire à un même modèle de l'activité cognitive, et que le même processus est réalisé.

Enfin, le domaine des Sciences Cognitives est défini par la mise en commun de travaux d'appartenances disciplinaires très différentes. Dans ce cadre, les propositions circulent entre les différentes disciplines scientifiques et ainsi contribuent à l'enrichissement de la communauté.

2.2.4. Sémantique, contexte, sous-détermination. Quelle caractérisation du langage est permise ?

Les différentes approches que nous venons de parcourir montrent qu'il est possible d'envisager des modèles relativement stables du contexte, contraignant l'usage et la signification des unités du langage. Par conséquent, on peut se doter d'une définition de la signification sensible à ce contexte. C'est la raison pour laquelle on adopte l'axiome de sous-détermination des unités du langage par rapport à leur contexte.

Un cadre pour le concept de « sémantique ».
Dans un travail pluridisciplinaire, fondamentalement on se trouve confronté à plusieurs définitions pour un même terme. C'est bien évidemment le cas pour « sémantique ».
On distingue déjà entre une caractérisation conceptuelle, inférentielle et linguistique de la sémantique.

1. Si le point de vue consiste à représenter des concepts, et donc des connaissances, la sémantique a pour objectif de caractériser comment les unités linguistiques portent ces concepts. Cette définition est issue de l'Intelligence Artificielle et donc ne prend pas en compte la dimension linguistique.
2. Si le point de vue consiste à caractériser comment la pensée humaine se traduit au travers de relations entre des unités linguistiques, on pourra alors spécifier une dimension inférentielle à la sémantique. (Cette inférence peut être seulement intensionnelle si elle caractérise la façon dont les expressions linguistiques sont interprétées dans un univers prédéfini). Cette définition se caractérise par une analyse des catégories linguistiques (verbes, articles, prépositions, connecteurs, etc.) mais pas du système de la langue dans son ensemble.
3. Si le point de vue consiste à étudier comment les unités de la langue se structurent relativement au contenu qu'elles véhiculent, on a lors une autre définition de la sémantique. Elle est donc fondée sur les principes fondamentaux de la linguistique : substitution, inversion, grammaticalité.

Ces différentes définitions spécifient des objets d'étude distincts. A cela on doit ajouter des caractérisations de l'interprétation qui spécifient encore les points de vue.

1. Si le point de vue est celui des structures de la langue, il s'agira évidemment de caractériser la sémantique à la fois en génération et en interprétation.
2. Si le point de vue est celui des inférences, la sémantique servira alors à caractériser l'interprétation des expressions.
3. Enfin, s'il s'agit de représenter des concepts, les questions d'interprétation et donc de communication ne se posent pas.

Notre positionnement peut sembler relativement difficile à saisir dans la mesure où les langages documentaires sont des langages artificiels. Par conséquent, une définition linguistique de la sémantique n'apparait pas aisée à utiliser. Par contre, les définitions conceptuelles, qui sont en grande partie reprises par la théorisation des ontologies, et celles qui relèvent de la sémantique formelle, apparaissent nettement plus opportunes dans notre cas. On caractérise la façon dont des structures régulières d'information sont interprétées dans des contextes spécifiés. On considère donc une dimension inférentielle à la sémantique. Cette définition permet également de rendre possible une prise en compte de phénomènes linguistiques, notamment les structures d'information. Elle ne permet pas de structurer des domaines de connaissances, mais seulement la caractérisation des univers d'interprétation (donc la définition intensionnelle des unités linguistiques).

Par ailleurs, dans la littérature, il n'existe quasiment pas de liens entre la sémantique formelle telle qu'on la définit et les modèles d'ontologie. Mis à part quelques travaux à propos des données géographiques, aucun lien n'a vraiment été établi[55]. Cela peut sembler étonnant, relativement au fait que les langages du web (à commencer par RDF) se réfèrent explicitement à la théorie des prédicats, donc aux premières formalisations logiques de la langue, à savoir les travaux de B. Russell. Cette ignorance réciproque peut être expliquée historiquement, par le fait que les principaux promoteurs du web de données sont issus de la logique associée à l'IA, et non de la linguistique formelle. Pour la théorisation des ontologies, la référence est la logique aristotélicienne pour B. Smith, associée aux travaux fondateurs de l'IA pour N. Guarino[56]. Par ailleurs, la linguistique associée à l'informatique a développé bien plus des modèles de grammaire dans lesquels la dimension logique ne constitue pas une part essentielle (par exemple, les HPSG, LFG, etc.). Ces modèles de grammaire peuvent aujourd'hui être intégrés dans les outils web comme LMF. Enfin, en sémantique logique, l'attention portée à la complexité des problèmes linguistiques n'a pas permis l'émergence de propositions faciles à utiliser dans le cadre du web de données[175].

Cette distinction doit quand même être décrite plus précisément : en effet, les flux donnent lieu, en tant que modèle de contraintes entre des structures de données hétérogènes, à des travaux importants dans le cadre du web de données. Mais c'est essentiellement dans sa dimension logique mathématique. La dimension linguistique est dès lors totalement absente.

Du côté des ontologies, les ressources sémantiques mobilisées sont essentiellement de nature lexicale (des lexiques associés à LEMON aux propositions de FRAMENET). L'idée consiste à relier des types d'objets distincts (linguistiques et conceptuels), mais en gardant pour chacun une autonomie totale.

L'héritage de la sémantique formelle n'est pas repris par les travaux sur les ontologies. La distinction entre expression et connaissance constitue une barrière, bien plus étanche que celle qui distingue des modélisations conceptuelles de la signification (notamment les frames) et les

[175] Par exemple, on pourra se référer au contenu du journal « Logic, Language and Information » : http://link.springer.com/journal/volumesAndIssues/10849

structures de connaissances. Dans ce dernier cas, la complémentarité des représentations est particulièrement opportune.
Plus précisément encore, la distinction opère entre une modélisation d'un domaine de connaissances et une représentation de l'interprétation. Les structures conceptuelles caractérisent des phénomènes ayant une existence alors même que la sémantique requiert et utilise des univers sans contenu conceptuel. Par exemple, un monde possible est caractérisé de façon à ne pas être peuplé d'ontologies de domaine.
Or, justement, un trait essentiel de la théorie des situations est qu'elle peut être considérée comme un cadre interprétatif dans lequel les constructions de connaissances (et par exemple les ontologies) peuvent être activées. Le cadre de travail sémantique fourni par la théorie des situations peut être défini comme un contexte social, empirique et communicationnel dans lequel les ontologies et les structures lexicales liées peuvent être considérées comme les ressources mobilisées par les acteurs. En somme, la théorie des situations (et certains modèles sémantiques proches comme les DRT ou la théorie de la pertinence) peut être définie comme un modèle sémantique formel permettant de positionner, dans un cadre de communication linguistique, les ontologies et leurs corrélats lexicaux comme des ressources mobilisées dans l'interprétation.
Ainsi, la sémantique à laquelle nous faisons référence est définie par les principes de la théorie des situations (et donc ses sources également, comme la philosophie de F. Drestke). Considérant les univers interprétatifs, on peut y intégrer les ontologies, ce qui permettrait de caractériser la sémantique contenue dans ces dernières comme une partie des univers interprétatifs. En effet, les ontologies ne couvrent pas les questions liées aux échanges et à l'interprétation. Par conséquent, la sémantique associée aux ontologies (y compris dans l'articulation aux lexiques) ne fait pas apparaître la dimension des discours et de l'expression, ce qui fait que le projet d'une sémantique formelle (et donc la définition d'une sémantique qu'il propose), restent d'actualité.
Par contre, on pourra éviter de considérer la théorie des situations comme une seule modélisation du contexte, mais bien des structures d'information, telles qu'elles sont interprétées comme décrivant un événement[57] ayant une portée dans les ressources.

Quelques éléments de sémantique intensionnelle.
Associée à la sémantique intensionnelle figure la définition de la relation entre des expressions et des univers d'interprétation afin de rendre compte de l'interprétation comme conséquence logique (M. Chambreuil, *Sémantiques*[58], p. 45). En ce sens, le rôle d'une sémantique logique consiste à représenter le raisonnement depuis l'expression jusqu'à son interprétation.
L'interprétation entendue comme l'affectation de valeurs de vérité à des contenue sémantiques constitue à partir de Frege une activité mentale. Dès lors, le but d'une sémantique sera soit de construire des règles de traduction de structures de l'expression vers des univers sémantiques (comme la sémantique intensionnelle de Montague), soit de proposer des modèles de la construction intensionnelle du sens (reprenant par exemple la théorie des actes de langage), sans se préoccuper outre mesure des questions d'organisation syntaxique.

Montague et ses successeurs ne changeront pas de méthodologie. En distinguant d'un côté des systèmes formels syntaxiques et des règles de traduction de ces phénomènes dans des univers d'interprétation, (composés des mondes possibles, d'entités individualisables, de contextes, de temps et de conditions d'expression), on redéfinit considérablement la relation de conséquence logique. La conséquence ne se caractérise plus seulement par des références et des conditions de vérité dans l'univers "réel" mais dans des mondes possibles. (C'est de cette façon-là que l'on peut expliquer le problème posé par " le roi de France est chauve ").

La sémantique intensionnelle caractérise la traduction de toute information d'une construction syntaxique vers un espace conceptuel appelé monde possible. Un monde possible se caractérise par un ensemble de propriétés pouvant se trouver actualisées à un moment dans le monde par une expression. Les mondes possibles seraient des réserves de concepts qui se trouvent actualisés à un moment dans une expression.

En suivant B. Partee[59], on résume les principes de la sémantique définie par Montague en distinguant :

- les modèles des structures du langage, dans lesquels sont organisées les expressions, notamment en syntaxe,
- les modèles des mondes possibles, dans lesquels se réalisent les entités du langage.

Le postulat fondateur de Montague, avant même de réaliser la sémantique, est l'établissement de règles formelles permettant de mettre en relation les expressions du langage naturel et les formes logiques.
Ces règles constituent des algèbres, à la fois au niveau de la syntaxe et de la sémantique.
Une algèbre, c'est :

- des ensembles d'éléments définis
- des opérations bien définies (ayant comme opérandes valeurs des éléments de l'algèbre).

En fait, en distinguant les interprétations sémantiques directes et indirectes, on postule :

- interprétations indirectes : passage par un langage intermédiaire avec une traduction compositionnelle. (Il y a donc interprétation sur ce langage intermédiaire.
- Interprétation directe : sans passage par un langage intermédiaire.

La théorie de Montague se caractérise par un homomorphisme entre les phénomènes syntaxiques et les entités sémantiques ("what is constrained is not the "substance" of the semantic but some properties of its structure in relation to syntactic structure").
Il s'agit alors de l'homomorphisme, à savoir la compositionalité. IL s'agit d'une approche "règle par règle" de la correspondance entre syntaxe et sémantique.
(Les règles syntaxiques prennent les expressions pour former des expressions complexes ; les règles sémantiques interprètent l'ensemble comme une fonction des interprétations des parties précédentes). Elle se distingue des sémantiques fondées sur la phrase.

Les conditions de vérité et les relations d'implications sont les deux fondements de la sémantique. On peut suivre le principe de Cresswell (78) de la plus grande certitude : nous ne connaissons pas la signification, mais nous savons que si pour deux phrases, si dans un cas l'une est vraie et l'autre pas, elles n'ont pas la même signification.

L'une des conséquences essentielles de la théorie de Montague : l'interprétation uniforme de tous les NP comme quantificateurs. (La sémantiques des prédicats du premier ordre rend impossible la possibilité d'imaginer une interprétation pour "le", "un", tout", "aucun". D'où le travail sur les quantificateurs généralisés[60].
La méthode par fragment est utilisée : elle propose une analyse complète de certaines parties du langage et non pas un analyse totale des phénomènes linguistiques. Elle porte sur des parties d'expressions).

Précision sur les entités intensionnelles.

On caractérise comme entité intensionnelle les énoncés modaux, les conditionnels, et "tout ce qui aurait pu avoir lieu si les conditions auraient été autres".
L'idée, défendue notamment par Cresswell, consiste à dire que le langage naturel est fondé sur des ontologies qui fondent la dimension intensionnelle du langage.
La nature même des entités linguistiques est d'être intensionnelle. De cette façon, les états mentaux ou structure cognitive constitue en fait le centre de l'élaboration de modèles sémantiques formels. (Cf. Cresswell, Stalnecker[61], etc.)

A l'opposé, on assiste au développement des théories fondées sur la référence directe, pour lesquelles le rôle du langage est avant tout de désignation. La référence directe inverse la perspective par rapport à Frege : on considère la conceptualisation comme un recours lorsqu'une désignation directe est impossible. Ainsi, dans le cas des descriptions définies (" le chat "), on n'aurait pas un concept limité par un déterminant, mais une désignation utilisant un concept pour signifier.

Théorie des situations et entités de l'univers d'interprétation.

L'idée de départ de la théorie des situations réside dans une difficulté rencontrée par Frege dans le traitement des démonstratifs (" ceci, cela, celui-ci ", etc.). Les démonstratifs désignent quelque chose qui est un objet, qui peut par ailleurs être conceptualisé comme tel mais qui en l'occurrence n'est pas désigné par une entité d'un monde possible.

Le fondement de la théorie des situations consiste à considérer que les expressions transmettent des informations (à savoir des contenus pouvant être conceptualisés), à propos de situations dans le monde les vérifiant.
Ainsi, on voit apparaître l'importance de Dretske et de sa théorie de l'information : c'est en vertu de certaines caractéristiques des phénomènes dans le monde (de pouvoir être inscrits dans un type), que ce type et cette occurrence transmettent une information –occurrence expressive interprétée dans un type- qui peuvent être référés à cette construction.
Dans sa formulation la plus simple, la théorie des situations repose sur une distribution de l'interprétation : elle est une activité mentale fondée sur des objets construits appartenant au monde. De cette façon, **la théorie des situations constitue la plus élaborée d'une sémantique fondée sur la cognition située** (où le travail interprétatif est partagé entre l'acteur et l'univers environnant).

Stratégiquement, J. Barwise & J. Perry considèrent que les expressions constituent des phénomènes avec type et occurrence. Ces phénomènes sont des productions langagières énoncées dans un certain temps et lieu et une structure formelle informationnelle exprimée dans un langage contraint). Ils considèrent que ces expressions indiquent dans l'univers interprétatif un certain nombre de phénomènes du monde auxquels on a accès grâce aux index, phénomènes étant par ailleurs abstraits ou conceptualisables. On dira alors que les index individualisent des objets TYPES, ou classes abstraites.

Dans ce contexte, l'information est la capacité qu'ont les expressions à indiquer quelque chose qui soit interprétable et qui réfère. (Tout le problème, c'est d'arriver à comprendre comment un type peut référer à un objet particulier, qui est au fondement de l'information).

Définition du contexte.
Avant de traiter de la sous-détermination ou sous-spécification, quelques précisions s'imposent à propos du contexte. Dans un cadre linguistique, il peut être défini ainsi : on considère comme contexte à une occurrence les expressions antérieures représentant un état de l'objet et les états de l'objet qui ne sont pas représentés.
Cette définition a plusieurs intérêts :

- elle prend en compte la notion de choix en distinguant la valeur réalisée et celles qui auraient pu l'être ; l'ensemble de ces valeurs constitue un contexte limité pour un prédicat. Cette définition permet également de caractériser le contenu informationnel.
- Elle définit le contexte comme un état dans lequel une occurrence s'inscrit et prend sens. La caractérisation de cet état sera le propos de la théorie des situations, dont il sera question plus loin, mais également de l'ensemble des propositions visant à exprimer en quoi on a une information, entendue comme une transformation à l'intérieur d'un état de connaissances[176].

Nous développerons plus amplement ces deux dimensions dans les parties concernées. Cette définition constitue une base qui montre l'articulation entre les problématiques de l'information et celle des états et des cadres. L'ensemble des choix informationnels s'inscrit dans un cadre construit et un choix réalisé signifie dans ce cadre.

Définition de la sous-détermination.
La sous-détermination (ou sous-spécification selon les auteurs) considère que s'il existe des règles proprement linguistiques, elles sont génériques et permettent des réalisations locales différenciées. On distingue alors les règles d'usage des règles de système (au sens du système de la langue). Par conséquent, on doit considérer que les propriétés associées aux entités du langage sont sous-déterminées par rapport à la précision qu'apporte l'information. La sous-spécification est directement une remise en cause de la compositionnalité ; la sémantique d'un terme serait choisie par son contexte (à savoir les autres entités, celles qui donnent sens à l'entité en question). La compositionnalité considère que la signification d'une proposition est la somme de la signification de chacun de ses membres et qu'il existe un lien direct entre réalisation syntaxique et sémantique. Elle a constitué pendant de longues années le paradigme fondamental de la sémantique parce qu'elle offrait la possibilité d'articuler la sémantique lexicale à celle de la phrase et du discours. La première remise en cause de ce principe est le fait de Z. Vendler[62]. Les unités lexicales ont en elle-même des propriétés distributives et argumentales. On trouvera dans les travaux de J. Jayez[63] des illustrations particulièrement pertinentes de ces principes, dans un cadre de sémantique formelle.
La sous-spécification constitue un argument de poids de la sémantique cognitive [64]; de même évidemment les cadres théoriques fondés sur les modalités[65].

Le premier champ de recherches linguistiques dans lequel les questions de sous-détermination sont exploitées est celui de la linguistique automatique, notamment les HPSG (Pollard & Sag) et les TAG[177]. Nous illustrons notre propos en utilisant ces dernières. La sous-spécification

[176] Le concept de contexte a donné lieu à une littérature impressionnante ; s'agissant d'une dimension fondamentale d'un objet d'observation, le paradoxe est qu'il est impossible de l'étudier. Inversement, si l'on étudie le contexte, qu'advient-il de l'objet ? Un certain nombre de propositions pour résoudre ce paradoxe ont été formulées, ARC 95, et il a fait l'objet d'une théorisation relativement poussée [BREZILLON]. Le problème du contexte est qu'il est hétérogène par rapport à l'objet d'observation, et par conséquent, le contexte apparaît difficile à appréhender sans une claire caractérisation de ses liens avec l'objet d'observation. Le dernier problème du contexte est sa récursivité : dès lors que le contexte devient un constituant d'un objet d'étude, il est dans un certain contexte, etc.

[177] HPSG : Head-Driven Phrase Structure Grammar (grammaire de la structure de la phrase gouvernée par la tête).

peut être considérée comme un niveau de représentation des phénomènes sémantiques ; c'est ainsi que l'entendent A. Joshi & L. Kallmeyer[66] lorsqu'ils caractérisent le niveau sémantique à partir de la représentation des dérivations d'arbres syntaxiques. La sémantique d'une phrase est dépendante de la structure de dérivation des TAG (Grammaire d'arbres adjoints) : elle exprime des dépendances d'argument et de prédicat. On aborde la dimension grammaticale en se fondant sur les relations structurelles (au sens entendu par Z. Harris), et non les dimensions génératives.
Ainsi, une représentation sémantique consiste en un ensemble de formules interprétées en conjonction et en un ensemble de variables d'argument :

- les formules sont des labels propositionnels (L1, L2, …)
- les variables d'argument sont des positions dans l'arbre syntaxique.

Ainsi :
« Tout chien jappe » : tout (x, chien (x), jappe (x)). Se réécrit : [tout (x, P1, P2)] avec les propositions P1 et P2 et P(x) avec x comme argument de « jappe ».

Les arbres élémentaires d'un LTAG (ou TAG lexicalisé) représentent les projections étendues d'items lexicaux et encapsulent les arguments syntaxiques et sémantiques dans l'ancrage lexical. Cette proposition signifie que les règles syntaxiques et sémantiques sont inscrites dans les unités lexicales et qu'une représentation de l'unité lexicale intègre la syntaxe et la sémantique qui lui est liée. Comme on y reviendra, les structures prédicatives (avec les rôles de prédicat et d'argument) s'intègrent parfaitement à cette perspective. L'intérêt des promoteurs des grammaires TAG pour la structure d'information s'explique de cette façon.

Ces arbres sont élémentaires au sens où tous les arguments, mais seulement eux, sont encapsulés. Ainsi, on sauvegarde la compositionnalité, qui constitue alors un principe essentiellement méthodologique : en conservant un niveau d'abstraction important pour l'articulation syntaxe sémantique, on peut alors construire une représentation sémantique compositionnelle par la sous-spécification des unités lexicales qu'elle accepte.

En sémantique formelle proprement dite (où l'on ne cherche pas à construire une grammaire mais seulement à caractériser la signification ou l'interprétation des unités et des structures linguistiques), on peut caractériser la sous-spécification de trois façons différentes :

1. comme J. Bos[67], on peut considérer qu'il s'agit d'une question de quantification et de portée de cette quantification ;
2. comme M. Pinkal[68] , on estime que la sous-spécification est un fait de langue (donc relève de la sémantique lexicale). Il s'agit d'un codage des ambiguïtés pour porter un ensemble d'interprétations pendant un processus d'interprétation. Dans le domaine lexical, il s'agit d'un enrichissement lexical. Les restrictions de sélection sont des phénomènes de "coercion"[69]. Au lieu de positionner des entrées lexicales homonymes multiples, on peut préférer une structure d'information complexe, dans laquelle l'existence de sous-parties variées permet la flexibilité de l'interprétation[70]. Ces sous-parties qui sont utilisées dans le processus sont dépendantes soit des restrictions de sélection des autres items, soit des principes généraux d'organisation des discours. Ici, on peut envisager des structures de discours variées ;
3. pour M. Poesio[71], il s'agit d'une interprétation partielle (un module des facultés du langage à traiter l'information ne produit qu'une interprétation partielle) : les types

---

TAG ; Tree-Adjoint Grammar (grammaire d'arbres adjoint).

Il s'agit ici de deux grammaires formelles d'unification, traitant à la fois des dimensions morphologiques, syntaxiques et sémantiques.

> d'interprétation sont sous-spécifiés lorsque les expressions interprétées montrent un certain type d'ambiguïté. On peut définir cette acception de la sous-spécification de la façon suivante Köning & Reyle [72] p. 1 : « formalism developed to represent sentence or text meanings with that degree of specificity that is determined by the context of interpretation. As the context changes they must allow for (partial) disambiguation steps performed by a process of refinement that goes hand in hand".

Ainsi, des significations distinctes en fonction des contextes d'usage et l'identification de ces contextes permettent alors de spécifier la signification établie dans le système ; les règles de contexte sont alors des structures, des cadres ou des types.

La conséquence de cette proposition initiale est une modification de l'approche des phénomènes linguistiques : elle privilégie les questions de structure et de compatibilité entre structures spécifiées dès le niveau lexical sur des hypothèses à fondement syntaxique ou sémantique. Par ailleurs, outre la dépendance au contexte, la sous-spécification permet de considérer les dimensions cognitives du langage (en présupposant un lien entre les structures et les raisonnements), une caractérisation différente des relations entre syntaxe et sémantique (A. Joshi et M. Kallmeyer, op.cit.) et enfin, une approche des lexiques par proximité en utilisant les clusters par exemple.
Les clusters définissent des ensembles de valeurs termes qui ne sont pas substituables exactement, mais ont suffisamment de parenté pour pouvoir être associés. (C'est de cette façon, par exemple, que l'on peut mettre en relation des termes équivalents mais appartenant à des vocabulaires professionnels différents). Ces clusters sont par exemple utilisés pour élaborer les bas niveaux d'abstraction des ontologies. Cette perspective permet enfin d'envisager des ontologies pour exprimer les relations entre structures hétérogènes. Ces regroupements de termes se distinguent des classes distributionnelles dans la mesure où les clusters ne sont pas définis par un rôle identique dans la prédication. Ils construisent des ensembles de termes par similarité d'usage, mais par d'autres moyens que l'analyse prédicative, donc, en ce qui nous concerne, l'information.

Pour nous, cette approche a une double conséquence ; elle permet d'envisager l'usage au travers de formes préétablies comme les structures d'information notamment et de structurer les ensembles lexicaux par des relations de proximité d'usage entre les termes, qui composent alors une classe[178].

Enfin, parce que l'on étend les contextes aux flux réguliers de transmission d'information, la sous-détermination permet de concilier deux dimensions temporelles des expressions, jusque-là inconciliables : d'une part l'instantanéité des réalisations linguistiques, d'autre part la durée des systèmes de transmission d'information et de leurs contenus. Cela peut sembler un problème de construction de données, mais il a son importance : il caractérise le fait que l'on puisse travailler sur un objet caractérisé par des régularités de fonctionnement (comme le système d'information ferroviaire) et l'ensemble des lexiques possibles, donc des contenus de cette information. On peut ainsi considérer n'importe quelle réalisation linguistique comme une instance du système.

---

[178] Le terme « classe » pose un certain nombre de problème. Nous nous en tiendrons à la définition proposée en sémantique par G. Gross. En effet, les classes qu'il propose sont issues des principes de Z. Harris, et sont donc fondées sur des opérations informationnelles. Malgré l'ambiguïté par rapport à l'acception logique du terme, il nous semble le plus conforme à la représentation d'ensemble de termes pouvant être substitués dans le cadre d'une prédication.

2.2.5. Conclusion d'étape.

Nous pouvons maintenant reprendre nos hypothèses initiales et considérer la façon dont on peut appréhender un usage. Il s'agit d'une articulation d'observables hétérogènes (opérations de symbolisation, médias de transmission, système de représentation) coordonnés au travers d'une même fonction. Les propriétés du langage sont considérées au travers du dispositif informationnel.

L'usage que l'on étudie se définit alors par la transmission d'information, depuis un processus de symbolisation d'un état du monde, exprimé dans une structure, jusqu'à l'interprétation des contenus transmis. Cette interprétation se traduit par des décisions[179] à propos de l'état du monde symbolisé. Il contient des entités circonscrites, qu'il s'agisse des unités linguistiques ou des outils et objets matériels. Un tel processus régulier, que l'on vient de le cerner, peut contenir d'autres dimensions que celle que nous allons privilégier : il peut être l'objet d'une sociologie de la traduction[73], et notamment Dominique Vinck[74]. Nous aurions pu mener cette étude dans le cadre de la pharmacie hospitalière[180].
Nous allons insister sur la modélisation du processus d'information. En effet, l'intérêt de la pharmacie hospitalière et de l'adaptation individualisée des posologies qui s'y déroule, c'est de permettre l'observation -dans un contexte limité et en utilisant des traces matérielles non numériques (au moins partiellement)- des phénomènes au cœur des processus actuels de structuration du web : échanges, transcription et description de contenus. Cette dernière tâche demande quelques explications quant à sa possibilité même ; c'est ce que l'on va discuter maintenant.

2.3. Questions de transferts d'information et de niveau d'abstraction.

Les précédentes propositions concernent la description du processus d'information « dans le monde » au travers de différents angles d'approches possibles, considérant toujours qu'il est nécessaire d'articuler les dimensions symboliques et d'activité.
Rappelons que les principaux enjeux de cette partie sont d'abord d'articuler les différentes dimensions du web de données de façon à se donner les moyens de concevoir des outils fondés sur une analyse de l'univers dans lequel ils auront à trouver un usage.

Depuis le début de cette partie, nous présentons le fait qu'il est nécessaire de disposer d'outils théoriques permettant de fonder l'analyse et la structuration des données dans l'objectif de proposer un outil. Notre point de vue est celui des Sciences Humaines et Sociales. Ce ne sont pas les seuls appuis théoriques possibles pour fonder théoriquement des outils. Par exemple, les théories de types, les logiques de description constituent des outils théoriques logiques permettant la maîtrise de l'élaboration des outils du web. Néanmoins, ces cadres ne prennent pas en compte les dimensions culturelles et cognitives constitutives de l'usage des services du web de données. Donc, nous devons pouvoir articuler ces deux approches parce qu'elles nous apparaissent les plus productives pour une maîtrise de la construction des outils.

---

[179] La décision ici peut être une attribution de valeurs de vérité ou une décision cognitive plus élaborée ou une décision pragmatique dans le cadre de l'activité dans laquelle l'information s'insère. La seconde nécessite la première.

[180] Nous ne l'avons pas fait parce que cette étude n'apportait pas à priori de connaissances pertinentes pour la menée de notre travail, qui consistait alors à identifier des régularités informationnelles permettant ensuite, par abstraction, d'élaborer un modèle.

Nous avons cerné le cadre explicitant comment il est possible d'appréhender les scènes présentées en annexe ou la circulation de l'information pharmaceutique dans l'hôpital[181]. Nous n'avons pas encore parlé de notre démarche et de la façon dont on peut articuler différents points de vue et approches dans le cadre d'une finalité scientifique de modélisation et de production d'un outil.

Comment alors concilier cette analyse avec la réalisation d'outils, sachant que l'on ne vise surtout pas à proposer des outils pour ces situations professionnelles ou sociales, mais justement pour des contextes autres, qui partagent les mêmes situations et où la circulation de l'information constitue un enjeu constant. Il s'agit là de l'enjeu de ce modèle que de pouvoir être adapté et profiter à des situations différentes.
Le lien entre les différentes situations (annonce en gare, pharmacie hospitalière et relations entre données primaires et publications dans le cadre de la description documentaire) réside dans la similarité de processus. Cette mise en relation conditionne le niveau de généricité du modèle.
Par contre, les distinctions entre les différentes situations sont importantes : les annonces en gare sont artificielles au sens où elles ne prennent pas en compte le contexte. Elles valent uniquement comme illustrations. A l'opposé, la circulation de l'information en direction de la pharmacie hospitalière, et relativement à l'adaptation de posologies, intègre l'activité dans son contexte spatio-temporel de la façon la plus complète possible. Enfin, pour le projet, le modèle comportant un flux d'information constitue l'apport essentiel proposé.
Ce passage depuis un phénomène observable en contexte vers une proposition d'outil ne constitue pas un acte anodin, mais le transfert d'un savoir-faire depuis un domaine où il constitue une pratique établie vers un espace en construction, qui est la structuration du web documentaire.

La différence est tout de même que pour la gestion du trafic ou pour la pharmacie, les questions d'information ne sont pas centrales, à la différence des bibliothèques et de la documentation. En effet, on intègre les questions de flux à l'intérieur de la description documentaire, qui évidemment débute avec les notices bibliographiques, puis se poursuit avec les questions de métadonnées. Les ontologies s'intègrent à un autre aspect de cette science, par le biais de l'organisation des connaissances.

L'idée fondatrice des niveaux d'abstraction et des transferts d'information, c'est que conformément aux propositions de la psychologie cognitive et de l'anthropologie cognitive distribuée de G. Hutchins, il est possible de faire circuler de l'information à condition d'avoir classée, donc abstraite. Cette théorie est également essentielle aux Sciences de l'Information, parce qu'elle justifie le recours aux organisations de connaissances.

---

[181] Nous introduisons ici notre principal objet concret, beaucoup plus complexe que les exemples proposés plus hauts, et dont le but était seulement d'illustrer concrètement ce sur quoi nous allions travailler. Pendant près de cinq ans (de 1995 à 2000), nous nous sommes installé dans un service de pharmacie hospitalière afin d'observer le « lien entre l'information et l'activité ». Rapidement, l'activité d'adaptation individualisée de posologies, réalisée à l'aide d'un logiciel de calcul probabiliste bayésien, nous est apparue comme pouvant fournir le socle matériel d'un travail de grande ampleur. Nous préciserons tout au long de ce travail les différents aspects de cette activité lorsqu'ils seront pertinents.

Enfin, mentionnons que le matériau observé et recueilli au cours de ces années ne correspond plus à celui utilisé aujourd'hui : le document numérique est diffusé dans l'hôpital, et le logiciel USC*PACK, coûteux et difficile à manier, n'est plus utilisé.

Avant d'en venir plus précisément à la définition et l'utilisation de ces deux concepts dans notre cadre de travail, nous aimerions revenir sur leurs enjeux dans notre cadre scientifique.

Classification et épistémologie sociale.

Les Sciences de l'Information, comme la Logique, peuvent être considérées comme des méta-disciplines, à savoir qu'elles structurent relativement à leurs propres finalités les produits des autres sciences. (Le parallèle est souvent établi entre une science qui cherche à expliciter les raisonnements, notamment ceux d'autres sciences, et une autre, qui vise à différencier les objets et les problématiques des différentes sciences). Cette position, à laquelle s'oppose L. Floridi[75], a donné naissance à la théorie de l'épistémologie sociale. Selon cette théorie, défendue notamment par Alvin Goldman[76], la véracité d'un fait scientifique serait définie comme le produit d'un travail social répondant à des postures collectives, négociées à l'intérieur des communautés.

Celui-ci postule une dimension intrinsèquement sociale à l'épistémologie et donc à la classification des différentes sciences (et au départ, d'un ouvrage dans une classe). Par conséquent, les classifications seraient définies comme des interprétations de la production culturelle et scientifique dans un cadre social. La notion de niveau d'abstraction ne serait alors plus pertinente[77]. La dimension sociale de la catégorisation, voire des clustérisations, rendrait mieux compte du travail de l'organisation des connaissances que les principes de la classification. Les principes qui justifient cette proposition sont d'une part la partialité des connaissances de chacun et la nécessité d'une intercompréhension. Le principe distinctif (le fait qu'une classe émerge et se distingue des autres) serait dû à l'accès ou pas à la connaissance associée à cette classe en fonction de son domaine de connaissances. Ainsi les principes classificatoires de bibliothèques seraient associés à la distribution sociale de la connaissance. Sont concernés ici les travaux caractérisant avant tout les classifications documentaires universelles. Les questions ne se posent pas de la même façon lorsqu'il s'agit de thésaurus puisque la dimension parcellaire et orientée par l'usage est intégrée directement dans la conception de l'outil. Un thésaurus est par définition socialisé. C'est pour cela que la mise en réseau des thésaurus sur le web nécessite l'utilisation d'un outil particulier, SKOS, qui construit de nouvelles formes de relations entre des connaissances. Ainsi, par exemple, si les thésaurus de nouvelle génération (EUROVOC[182] par exemple) privilégient les associations de termes par rapport aux hiérarchies, c'est sous l'influence des langages du web et de la montée de la pluridisciplinarité. Certes, les termes associés constituent un concept opératoire déjà utilisé depuis longtemps ; ce qui évolue, c'est la façon dont on construit une connaissance et la façon dont elle est envisagée en lien à une autre. Ainsi, doit-on être exhaustif sur les composants d'un domaine ou pour un domaine donné, doit-on expliciter les liens entre une classe et l'ensemble de celles qui lui sont relatives ? Comment la notion de grain peut-elle être utilisée ?

Toutes ces questions essentielles ne peuvent être éclaircies que si l'on explicite comment on peut classer une instance dans un type, par quelle opération régulière un tel processus peut être mis en œuvre. Notre raisonnement ne porte pas sur les items ou instances considérés (la caractérisation des termes ou concepts, ou encore descripteurs utilisés), mais sur le processus, ou l'opération, qui permet d'opérer la classification ou plus généralement le typage d'une entité. Le typage permet de renouveler la question de l'abstraction et de la définition d'un niveau opportun pour caractériser une classification adéquate.

Rappelons que ce sont les relations entre les entités conceptuelles et leur évolution qui sont le centre de nos questions, et non le choix de tel ou tel terme ou descripteur. Ces deux questions ne sont pas totalement autonomes, mais ne relèvent pas des mêmes approches puisque dans

[182][182] http://eurovoc.europa.eu/drupal/?q=fr

un cas, il est question de dénomination et de description sémantique, et dans l'autre d'inférence.

Les niveaux d'abstraction et le typage ont un rôle fondamental dans cette étude et notre projet. Ils ne sont pas non plus secondaires dans les perspectives plus traditionnelles des Sciences de l'Information. Simplement, au travers de la caractérisation de la classification comme opération, on propose un outil permettant de structurer les représentations en explicitant le processus mis en œuvre. Cette explicitation se marque dans la portée informationnelle des classifications. A partir du moment où l'on considère les classifications comme des apports informationnels sur les documents, elles peuvent être utilisées comme descripteurs de ces documents ; c'est ce que fait le Dublin Core en proposant de les utiliser dans le cadre du renseignement des métadonnées. L'usage des classifications et des langages contrôlés change à partir du moment où ils sont inscrits dans des procédures de renseignement de métadonnées.[183]. Cette transformation des usages ne remet pas en cause les classifications elles-mêmes, ni leur usage initial. La véritable transformation est qu'elles servent à décrire des publications, et par conséquent, elles ne peuvent plus être seulement considérées comme une organisation de connaissances. Les entités symboliques qu'elles contiennent sont considérées avant tout comme porteuse d'une information à propos des publications qu'elles décrivent. Cet usage n'a aucune incidence sur les classifications elles-mêmes, notamment la façon dont elles sont construites et révisées. Néanmoins, cet usage transforme leur rôle dans le cadre de la circulation des informations.

Ce nouvel usage des classifications et plus globalement des langages documentaires s'inscrit technologiquement dans l'adoption de mêmes formats de représentation. Les capacités descriptives de ces classifications et langages documentaires sont liées aux propriétés prédicatives de RDF. Mais cette explication n'est pas suffisante. Elle concerne en effet certaines caractéristiques internes de la structuration et de la signification des unités symboliques inscrites dans ces classifications et langages documentaires. Les classifications et langages documentaires sont souvent comparés aux ontologies et sont par exemple préférés à ces dernières par le Dublin Core pour élaborer les profils[184]parce qu'elles proposent une structuration des domaines plus élaborée.

Deux questions s'inscrivent dans cette partie : d'une part il s'agit d'expliciter comment on peut passer de données empiriques à des représentations abstraites, et ensuite pourquoi ce sont seules ces représentations qui permettent de transmettre de l'information. Ces deux questions concernent un cadre relativement général et l'amorce d'une réponse (que nous développerons plus loin avec la théorie de l'information) à la question : pourquoi est-ce que l'on abstrait, quel est l'usage de cette abstraction ?

Nous commencerons donc par présenter la question de la modélisation sous cet angle. Elle nous permettra d'introduire le recours au typage et enfin nous enfin nous aborderons comment on peut définir les transferts d'information au travers d'une perception dynamique.

---

[183] La question de savoir si les grandes classifications des sciences sont elles-mêmes scientifiques ou si elles relèvent d'une épistémologie sociale constitue une question d'un autre ordre : en effet, est-ce que la Dewey ou la Classification Décimale Universelle constituent des outils pratiques, faciles d'usage et reflètent donc les connaissances des bibliothécaires et des documentalistes et non celles des communautés scientifiques ? Ou au contraire, sont-elles l'agrégation des différentes structurations internes des disciplines scientifiques ? Répondre de façon précise à ces questions, au moins maintenant, n'est pas notre propos.

[184] (http://dcevents.dublincore.org/IntConf/index/pages/view/APaltOO

2.3.1. Questionnement classique relativement aux modèles, leur généralisation et leur adaptation.

La question de l'abstraction (dont le typage constitue une partie) constitue un problème fondamental à la fois pour notre projet directement et pour les outils et objets que l'on utilise. En effet, comment peut-on représenter ces liens entre ressources et publications en utilisant l'expérience de ce que l'on a des circulations d'information ? Il faut arriver à représenter le processus (mental) à l'œuvre de façon suffisamment abstraite pour qu'il puisse être utilisable dans d'autres contextes.
Néanmoins, cette question peut être largement élargie et concerner toute activité impliquant un raisonnement où l'abstraction constitue une condition pour le transfert d'information. Les classifications et les autres langages documentaires intègrent à la fois ce raisonnement d'abstraction et son usage dans le cadre de la mise à disposition des connaissances.

Si l'on considère qu'une classification est une abstraction, et que cette abstraction constitue une condition pour relier entre elles différentes données, alors on peut proposer une méthodologie qui aurait deux domaines d'application : d'une part fonder notre propre modèle d'élaboration d'un outil, d'autre part expliciter la constitution et l'usage des outils à base de classification.

Dans ce contexte, la question se pose de deux façons :

- en quoi le modèle est-il adapté par rapport à l'outil que l'on cherche à construire ?
- en quoi le niveau d'abstraction de la description correspond-il à celui requis par le modèle que l'on cherche à établir ?

La première question n'est pas nouvelle, et parcourt les sciences articulant une dimension descriptive et des préoccupations de conception, qu'il s'agisse de la linguistique ou de la relation entre la psychologie cognitive et l'Intelligence Artificielle (et plus particulièrement les propositions de grammaire génératives ou de *General Problem Solver* de Simon et Newell[78]). La critique généralement formulée est que les modèles décrivent plus leur cohérence interne relativement à l'objectif fixé que les phénomènes du monde sur lesquels ils s'appuient (en ce sens, la grammaire générative n'explique guère les langues qu'elle décrit mais seulement son propre fonctionnement). Dans notre cas, on ne vise pas à automatiser un processus existant mais à exporter dans un espace en formation (le web de données) des éléments de culture élaborés et utilisés dans d'autres domaines (notamment celui de l'information relative à l'adaptation de posologie) ; pour cela, on doit postuler que le modèle de raisonnement que l'on construit est suffisamment général pour pouvoir être utilisé dans des contextes distincts de celui dans lequel il a été observé : notre hypothèse est que seul le typage des données permet de transférer les résultats d'un domaine d'activité vers un autre. Le processus que l'on décrit sous forme d'un modèle de raisonnement est partiel par rapport à une activité mais présent dans de nombreuses activités humaines.

La seconde question se justifie par le fait qu'il faille utiliser des résultats depuis un contexte analytique vers la conception d'un outil en explicitant l'adéquation de l'usage de ces résultats d'analyse pour la conception. En d'autres termes, comment peut-on affirmer que telle analyse sera pertinente pour la conception d'outil ?
Cette question est différente de l'adage selon lequel plus on abstrait, plus on gagne en généralité tout en perdant des propriétés : il s'agit de connaître le niveau de pertinence requis pour la description afin qu'elle puisse être utilisable dans le cadre de la conception d'un outil d'information particulier. La question renvoie donc à la spécificité de chaque outil, et au niveau d'abstraction du traitement de ses données. S'il faut abstraire une description pour

élaborer un modèle, l'espace de réception doit également être construit de façon à hériter de cette analyse au niveau d'abstraction requis.
Notre hypothèse consiste à considérer qu'un seul niveau d'abstraction ne peut être pertinent, mais que c'est bien la relation entre plusieurs niveaux qui caractérise l'approche la plus pertinente. Les niveaux requis seront celui des instances, celui des types qui permettent de classer les instances et enfin des théories permettant d'expliquer les inférences qu'il est possible de réaliser entre ces deux précédents niveaux. Cette proposition constitue une application des flux en tant qu'outil de méthodologie scientifique.

Caractérisation du phénomène du monde que l'on cherche à abstraire. Présentation générale de l'adaptation de posologies.

Le contexte de l'adaptation de posologie comme terrain d'investigation trouve ici son intérêt. L'adaptation de posologie constitue une activité de pharmacie hospitalière qui se caractérise par un ensemble de procédures de suivi thérapeutique.
Comme toute activité hospitalière, l'adaptation a à la fois une légitimité pratique et constitue un enjeu scientifique, à savoir que la façon de mener cette activité est un sujet de débat à l'intérieur de la communauté scientifique concernée.

Au départ, l'adaptation repose sur le contrôle du dosage d'un médicament, suivant en cela les paramètres de l'efficacité thérapeutique ou de l'élimination, sachant que l'accumulation de molécules dans le rein comporte d'importants risques toxiques. Le facteur temps est également important puisque l'antibiotique injecté génère des anticorps de la part du patient, ce qui le rend rapidement inopérant.

L'adaptation de posologies se caractérise par un objectif relativement simple : l'adaptation de doses de médicament en fonction des besoins du malade. Dans son versant pharmaceutique, cette acception se traduit par la proposition de doses (en quantité et en temps) en fonction des capacités d'élimination (rénale et non rénale) du patient autant que la nature de l'infection. Les informations sur cette élimination proviennent d'analyses et de mesures combinées. (Le choix du médicament est proposé par le prescripteur et l'administration est le fait des infirmières).
Le problème que doit résoudre toute adaptation est bien l'adéquation des doses pour l'efficacité thérapeutique et l'absence d'intoxication liée au surdosage de la molécule injectée. Pour cela, des données relatives autant à la clinique, la thérapeutique qu'au contrôle de la thérapie (données recueillies durant le traitement) sont nécessaires.

L'adaptation est un processus circonscrit : outils, compétences et situations sont limités. L'engagement d'une occurrence de l'activité dépend de la communication de nouvelles données, relatives au patient dont elle se préoccupe. Ainsi, les données symboliques, et leur actualisation (et ce même si l'on ne s'intéresse pas immédiatement à leur contenu linguistique), constituent des objets contraignant l'engagement et le déroulement de l'activité.
Nous pourrons noter que ce système (et donc le flux d'information qui le parcourt) est essentiellement constitué d'informations écrites sur des supports pré-imprimés, et rarement d'informations numériques[185].

L'adaptation est aussi un processus circonscrit parce que seulement trois types « d'entités du monde » (construites par des connaissances) sont les constituants fondamentaux et uniques de

[185] Nous nous référons à la période 1995-2000.

l'activité. Il s'agirait du patient, de la molécule et de la bactérie. Ces « entités du monde » sont considérées comme individus, distinctement des objets fonctionnels dont le comportement est prédictible (comme les outils d'administration par exemple). D'autres individus existent dans l'activité, comme les médecins et les personnels soignant et les médecins, mais sur ceux-ci, on ne peut pas agir, mais seulement argumenter ou requérir.
Or, pour caractériser comme individus ces « entités du monde », une seule identification par des marques linguistiques s'avérera insuffisante ; des critères spatiaux seront nécessaires pour caractériser efficacement quels sont ces ingrédients. Ils permettront de constituer le concept d'individu, emprunté à P.F. Strawson[79], et qui caractérise une entité occurrente inassimilable dont le comportement ne peut être modélisé sans un fort degré d'incertitude.

Cette activité repose sur un flux d'information circulant entre la pharmacie, les services de soin et le laboratoire. L'information sert à la fois à engager l'activité, sert d'argument à chaque décision, mais également elle sert à contrôler la validité de chaque action proposée. Les résultats d'analyse servent à réviser les dosages prescrits. Dans un tel contexte, l'information véhicule un contenu ontologiquement vrai.

L'activité d'adaptation repose sur les produits d'autres activités, comme la prescription médicale, l'analyse biologique, le soin infirmier. Ces activités produisent des informations qui sont transmises à la pharmacie. (Le système d'information de la pharmacie dépasse largement l'activité elle-même, voire même l'activité ne constitue qu'une part très réduite de ce système d'information dès lors que l'on envisage le circuit du soin).
Le système d'information[186] s'inscrit à l'intérieur de plusieurs activités, et assure leur mise en relation. En ce sens, il est difficile de considérer l'information comme un outil dédié à une activité si l'on ne prend pas en compte les entités du monde évoquées plus haut et les émetteurs/récepteurs. Ce flux d'information constitue, à l'intérieur du système, une direction régulière pour l'échange d'un certain type d'informations.

L'adaptation constitue aussi la seule activité professionnelle parmi celles impliquées dans le système d'information de l'hôpital à ne reposer que sur l'information, et non sur une observation et une intervention directes sur le patient. Cette remarque est importante : l'information ne peut être ambiguë ou imprécise, au risque de mettre en cause la santé du patient.

En ce sens, l'adaptation constitue la position la plus adaptée pour observer les phénomènes informationnels et les caractériser fonctionnellement. L'adaptation construit son système d'information en s'appuyant en grande partie sur des circuits préexistant. En effet, la plus grande partie des informations que va utiliser le pharmacien sont impliquées dans d'autres processus de soin. Seules les analyses de cinétiques sont spécifiques à l'adaptation. (Les analyses de prélèvements sanguins sont réalisées dans la pharmacie, dans un espace de travail dédié). Ainsi, l'information utilisée n'est pas essentiellement destinée à l'adaptation. C'est l'espace de réception, à savoir l'ensemble des supports (numériques et papiers) sur lesquels l'information est copiée et archivée, qui affecte à l'information un rôle dans le raisonnement d'adaptation. En même temps, cette affectation caractérise une spécialisation d'interprétation de l'information, créant ainsi un flux.

---

[186] On entend ici système au sens le plus générique et non pas informatique : il désigne une structure fonctionnelle par laquelle chaque composant est relié à un autre dans le cadre de la réalisation d'une tâche ou du maintien de la structure.

C'est donc à partir de cette activité et du flux d'information qui l'alimente que l'on cherche à élaborer un modèle général pour l'ensemble des processus informationnels. Il est destiné à s'appliquer à un contexte techniquement et technologiquement très différent de celui de l'adaptation. En effet, si l'activité étudiée utilise encore massivement l'imprimé, les infrastructures d'e-sciences sont totalement numériques. Mais justement, si le modèle est caractérisé à un niveau d'abstraction pertinent, alors le transfert de l'acquis culturel de l'adaptation vers l'organisation des infrastructures devient possible.

Caractérisation des principes pour l'abstraction.
Le niveau d'abstraction pertinent pour observer les flux est celui qui permet d'appréhender les constituants de l'adaptation :

- les sortes d'objet sur lesquels de l'information porte : les individus,
- les mécanismes de symbolisation et de transfert de l'information : les opérations de production de symboles et de communication de ces ensembles structurés de symboles,
- et enfin la représentation de cette information : une structure de symboles normée.

Ce système est régulier : les individus sont d'un même type, les opérations sont identiques, les circuits et les modèles de représentation de l'information sont normés. Ce système possède les propriétés requises pour être généralisé à des situations empiriquement différentes.

Rappelons encore qu'ici on ne cherche pas à adopter des outils disciplinaires (anthropologiques ou linguistiques par exemple), mais à construire un modèle utilisant des méthodes, des positionnements et des connaissances issues de ces approches. Ce ne sont donc pas les disciplines utilisées dans ce travail qui nous permettront de caractériser la conduite à tenir vis-à-vis de l'abstraction.

C'est à cette question que cherche à répondre la méthode des niveaux d'abstraction de L. Floridi[80]. Cette méthode ne constitue pas une méthodologie pour l'usage de résultats depuis un contexte vers un autre, mais seulement pour identifier un traitement des données adéquat pour accéder à une généralisation. Pour traiter du transfert, nous présenterons les logiques de transition. La théorie des niveaux d'abstraction fait partie de l'ensemble des travaux (comme ceux de D. Marr ou Davidson commentés d'ailleurs par L. Floridi) qui ont cherché à conceptualiser et expliquer en quoi un changement de niveau d'observation, ou d'abstraction, changeait la caractérisation des propriétés que l'on pouvait associer aux observables. Cette méthode a donné lieu à une nouvelle publication de Luciano Floridi, qui éclaircit certains points et surtout les fondations de sa méthode[81].
Au départ, il s'agit bien de questions de méthode, mais dont les implications épistémologiques sont importantes.

2.3.2. Caractérisation générale du typage et des niveaux d'abstraction.

L'information est en lien autant avec le niveau des connaissances que celui des réalisations expressives ou encore des lexiques. Elle relie également des phénomènes mesurables (que l'on peut calculer comme les résultats d'analyse biologique) et d'autres qui sont purement mentaux (comme une évaluation). En ce sens, l'information peut être abordée au travers d'approches partielles qui toutes la considèrent par des propriétés constitutives distinctes. Rien n'interdit d'enrichir l'analyse par des mises en relation entre ces approches ; néanmoins,

ces mises en relation doivent être nécessairement explicitées de façon à garantir la cohérence du propos[187].

Cette question concerne la construction des données issues de l'objet d'observation, et non le transfert du modèle obtenu. Nous étudierons cette question dans le chapitre suivant. La question des niveaux d'abstraction concerne la caractérisation de propriétés différentes et de l'observation de leurs réalisations sur des données exemplaires.

Néanmoins, nous devons rapidement caractériser ce que recouvre le processus d'abstraction en Sciences Humaines et Sociales. La méthodologie générale adoptée pour justifier et caractériser l'abstraction se caractérise ainsi (en reprenant l'exposé de M. Steedman, pertinent d'abord en linguistique formelle) :
- certaines réalisations mettent en évidence des formes régulières.
- Ensuite, ces formes régulières permettent des réécritures utilisant la coordination et les autres principes d'identification des problèmes linguistiques (distribution, co-occurrence, incompatibilité).
- Enfin, ces réécritures peuvent être exprimées dans un autre langage. C'est à ce moment-là que la question du métalangage mathématique choisi se pose.

Le deuxième principe ne peut pas s'appliquer seul parce que les données que l'on utilise ne sont pas toutes symboliques (elles peuvent être des formats, des données, des objets matériels comme les livres) ; cette hétérogénéité oblige à se doter d'un outil de contrôle de l'abstraction beaucoup plus puissant que les principes méthodologiques que l'on vient d'évoquer.

Quels sont alors les moyens scientifiques que l'on se donne pour assurer cette cohérence ? Les différentes théories identifiées plus haut fournissent des réponses relativement différentes dans la mesure où elles ne sont pas issues des mêmes disciplines et surtout que les modèles qui en dérivent n'ont pas les mêmes finalités. Elles vont utiliser des outils distincts d'abstraction.
Les propositions de Floridi ne sont pas associées à une discipline autre que la philosophie, ce qui donne à ses propositions un caractère générique qui nous est fort utile. Il utilise les théories de types, s'inspirant donc fortement des formalisations mathématiques des classifications. Cette communauté d'utilisation (avec les flux) de la théorie des types constitue un autre intérêt de cette méthode, entre autre parce qu'elle permet d'expliciter ce que signifie chaque classification (chaque classification est l'abstraction d'un individu ; elle utilise un paramètre particulier qui permet d'obtenir un type pour tel paramètre associé à cet individu).

On utilisera de façon assez systématique les théories de types de façon à disposer d'un cadre formel homogène entre des différentes parties du travail. En effet, la théorie des types renouvelle fortement la façon dont il est possible de représenter les phénomènes d'abstraction et de classification, fournissant ainsi une alternative à la théorie des ensembles.
C'est la raison pour laquelle les théories et méthodes de niveau d'abstraction ont connu un regain d'intérêt. En effet, les premières formulations de ces niveaux d'abstractions sont dues entre autre à A. Newell et H. Simon, et ont constitué un cadre pour l'intelligence artificielle.
Depuis, de nombreuses critiques ont été formulées. Néanmoins, la question reste d'actualité dès lors que l'on utilise un minimum d'analyse conceptuelle. L'enjeu consiste bien alors à

[187] On pourra s'interroger sur la présence de notations formelles dans ce travail : on les justifie par la clarté de l'exposition, à savoir qu'une expression formelle est dénuée d'ambiguïté, ce qui ne veut pas dire qu'elle soit systématiquement expressive. Par ailleurs, notre objet intègre plusieurs niveaux d'abstraction, dont un ne peut être représenté qu'en utilisant des outils formels.

expliciter ce que l'on fait lorsqu'une telle analyse est menée : que fait-on lorsque l'on caractérise un générique par rapport à un spécifique, comment on explicite cette relation, comment une relation d'ingrédience peut être caractérisée ?
Par ailleurs, il convient de s'entendre sur ce que l'on attend ici des niveaux d'abstraction : il ne s'agit ni d'une théorie, ni d'une ontologie, mais d'une méthode permettant d'expliciter les relations entre entités symboliques.
Cette proposition est une méthodologie ; en ce sens, elle sert à expliciter un travail, un processus. Elle ne peut se substituer à une science des signes, et donc ne traite que d'unités signifiantes constituées.

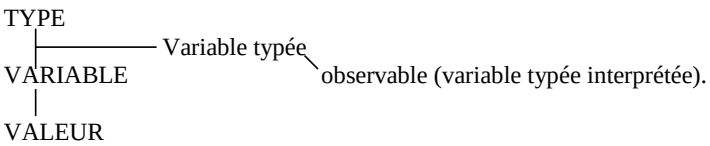
Caractérisation générale du typage.
La théorie des niveaux d'abstraction s'appuie sur une théorie de types. Une théorie de types est un outil d'abstraction permettant de construire l'univers, d'y identifier des régularités, en utilisant plusieurs niveaux d'assemblage. En ce sens, une théorie de type permet de contrôler la construction des objets en classant les occurrences dans des types explicitement formés.

Cette propriété des types permet une construction des univers et de domaines plus précise que par la théorie des ensembles par exemple[82]. Elle permet surtout une différenciation des objets relativement à des critères explicites, autrement dit à des paramètres.

Une théorie de types s'applique à la construction d'un outillage analytique. Elle peut constituer le langage d'expression d'une démarche scientifique.
Elle peut également être utilisée comme langage de description, et de classer ainsi les objets du monde.
En outre, elle permet de construire des modèles, à savoir qu'elle permet d'abstraire des régularités observées afin de proposer des programmes.

Un niveau d'abstraction se définit en premier lieu par des opérations sur des objets explicitement définis :
Une variable typée est contrainte par un certain type.
Un observable est une variable typée interprétée.
Un niveau d'abstraction est une collection d'observables.
Un comportement est en ensemble de valeurs d'observables contraintes.
Un niveau modéré d'abstraction est un niveau d'abstraction prédit sur le comportement.

Néanmoins, pour comprendre cette construction, il est nécessaire de reprendre les trois niveaux:



On obtient ainsi un niveau d'abstraction (ou collection d'observables). Cette collection d'observables permet de prédire un comportement, à savoir un ensemble de valeurs.
On caractérisera deux niveaux d'abstraction différents pour aborder la structure d'information et son contexte, et pour aborder les lexiques satisfaisant ces structures. Ce qu'apporte L. Floridi, notamment par rapport à la prédication, c'est la distinction entre l'observable et la valeur, et la dépendance de cet observable par rapport au type. Tout type est ainsi obtenu parce qu'un observable a été construit pour caractériser une variable.

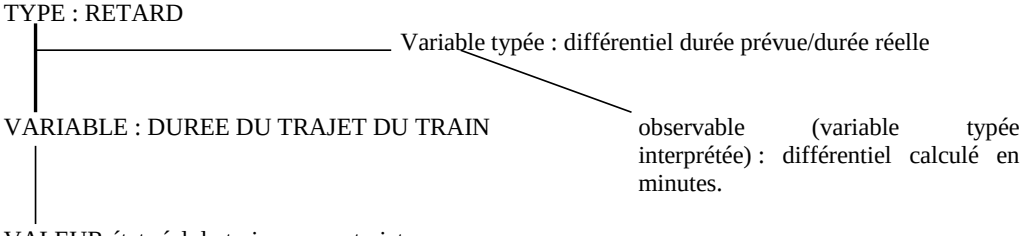
Variable typée : une variable est un symbole prenant la place d'un référent inconnu ou changeant. Donc une variable typée est une variable qualifiée pour satisfaire uniquement une certaine sorte de données.
Une variable typée est constituée d'une variable et d'un type qui est l'ensemble de toutes les valeurs que cette entité peut prendre. La variable typée s'écrit x/X (x : variable, X : le type).
Deux variables types sont égales leurs noms et leurs types sont égaux. Une variable à laquelle il n'est pas possible d'assigner des valeurs est une variable sans type.

Observable : façon dont les traits du modèle correspondent à la situation modélisée. Le modèle peut être un système, une théorie ou un domaine de discours. Les observables sont donc les phénomènes que l'on peut identifier à partir du moment où l'on a adopté une théorie, un modèle ou toute autre construction abstraite fondatrice d'une activité scientifique. Par exemple, l'annotation de corpus constitue une activité de construction d'observable à partir de données linguistiques primaires.
L'observable est une variable typée interprétée avec la caractérisation du trait du système qu'elle représente. (Le trait du système peut être empirique ou physique, il peut être également un artifice ou un modèle conceptuel). Un observable discret est d'un type fini (les valeurs possibles sont limitées, sinon, c'est un analogue).
Un observable est une attitude par rapport à une instance : c'est globalement une simplification. Un type est approprié par rapport à une instance en fonction de cette attitude, qui constitue un contexte permettant de spécifier la ou les propriétés permettant d'inscrire l'instance dans le type. Cette attitude est relative à une théorie (qui reste extérieure à la portée de la théorie des niveaux d'abstraction).

Exemple : Si l'on veut mesurer les attributs physiques de la personne : on utilise comme variable les nombres naturels et comme type l'unité de mesure. Si l'on a deux unités de mesure différentes, alors on a de mêmes variables types mais pas une égalité d'observables : les valeurs représentées sont distinctes.

En même temps, les types peuvent être basiques (regroupant des variables). Ils peuvent être complexes lorsqu'ils sont associés à des inférences, et notamment, relativement à la façon dont une variable peut être classée en fonction de l'inférence entre deux types. Ainsi, lorsque l'on caractérise la qualité d'un vin, celle-ci est déduite du vieillissement (donc de la durée entre l'année de fabrication et celle de consommation).
Par ailleurs, la fonction depuis le premier vers le second type est considérée comme un observable.

On pourra prendre un exemple issu de l'annonce en gare afin de montrer comment telle valeur de la durée de trajet d'un train est de type retard :



Mise en œuvre de la théorie des types pour caractériser les différents niveaux d'abstraction et les échanges.

Ces propositions constituent le point de départ de la proposition de L. Floridi. Elles sont complétées notamment par la notion de gradient d'abstraction. Elle cherche à expliciter le passage d'un niveau d'abstraction vers un autre, et donc l'observation de propriétés différentes caractérisées pour un objet d'observation.

On utilisera cette proposition pour justifier la construction d'un objet d'observation qui comporte :

- des phénomènes réguliers affectant des unités symboliques considérées dans le contexte de leur réalisation (qu'il s'agisse de leur émergence, de leur transfert, duplication ou encore interprétation et archivage).
- Une modélisation des régularités de comportements et d'actions qui sont d'ordre culturel et cognitif. Ces comportements sont manifestés dans la manipulation d'objets et de données comme la lecture de résultats, la communication de ces derniers, leur duplication sur des feuilles. Ainsi, une telle représentation des données permet de caractériser non seulement des comportements d'entités symboliques (en isolation et de façon structurée), mais également de les insérer dans un processus matériel typique.
- Cette caractérisation des données permet de disposer d'un cadre matériel (ou disons de données construites hétérogènes) pour observer un phénomène dans ses dimensions spatiales et temporelles. On traduit les propriétés observées dans un niveau à un autre niveau par la caractérisation de gradients.

Par exemple, si je m'intéresse, dans le contexte de la pharmacie hospitalière, aux expressions interprétées, je peux reconstituer la forme suivante : "M. Smith, kinetic, concentrations, 2,32mg/ml, T0, 12h23". Néanmoins, si je considère que cette expression est le résultat d'une duplication des entités symboliques depuis un lieu d'origine dans l'hôpital jusqu'à la pharmacie, j'introduis un paramètre spatial qui me permet de construire un nouvel observable. Celui-ci peut être schématisé de la façon suivante :

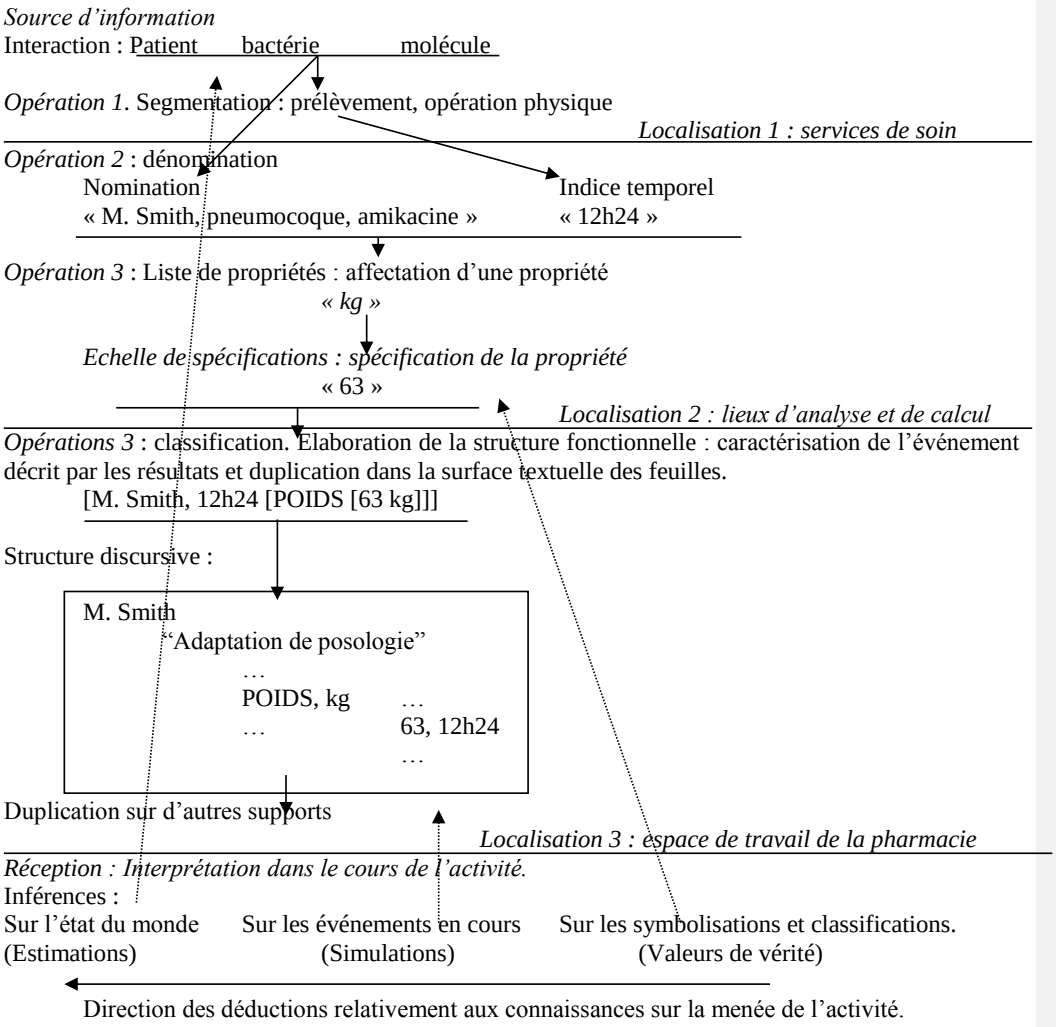


Figure 1. Schématisation de trajectoire de l'information depuis le patient jusqu'à l'espace de travail du pharmacien.

Les entités constituant l'expression interprétée sont ainsi distribuées dans l'espace. On ne dispose plus du tout du même observable que si l'on ne prend pas en compte la dimension spatiale. En effet, une entité symbolique est interprétée relativement à son sens et à sa référence mais également à son contexte de réalisation première (l'opération localisée dont elle est issue).

La conséquence immédiate de cette transformation des observables par rapport à une approche traditionnelle du langage est que l'on peut justifier de types qui usuellement ne sont pas utilisés dans l'analyse linguistique. Toute la partie relative à la structure d'information sera fondée sur cette construction des données. L'enjeu sera également essentiel lors de la caractérisation d'ontologies : on peut distinguer et structurer les différents concepts intervenant dans une ontologie en utilisant le paramètre de la situation et des paramètres de réalisation de ces concepts. Par exemple, un concept réalisé dans le cadre d'une ontologie ne

pourra pas être utilisé de la même façon qu'un concept représenté par un même mot dans un lexique ou une occurrence textuelle.

Nous venons de présenter le typage de façon relativement informelle et de façon à montrer qu'il permet de faire exister un autre domaine de réalité que le triangle sémiotique. (Notons que ce niveau supplémentaire doit beaucoup aux propositions d'E. Hutchins dont nous avons amorcé la présentation plus haut). En effet, la caractérisation d'opération, et plus encore d'opération réalisée dans l'espace pour l'obtention d'une unité symbolique, constitue un résultat initial important associé à ce typage. Il apporte une dimension supplémentaire à la construction d'observables : à la triade Signifiant-Sens-Référence, on peut adjoindre le cadre d'obtention de ce symbole. Nous reviendrons ultérieurement sur cette question ; restons pour le moment sur la présentation de la théorie des niveaux d'abstraction.

Hypothèses et contraintes centrales de la théorie des niveaux d'abstraction.
Nous présentons maintenant les concepts fondamentaux de la théorie des niveaux d'abstraction en considérant des collections d'observables et de variables typées. De cette façon on peut caractériser les niveaux et les gradients de façon plus précise. Pour cela, on présentera l'algèbre relationnelle simple qui permet de les représenter de façon plus formelle.
On peut poser la question de niveaux à la fois dans un cadre de modélisation et d'analyse : à partir des entités observables, qu'est-ce que l'on peut construire comme structure assemblant ces observables et comment caractériser ces structures assemblant des observables ?
Par exemple, il est effectivement possible de travailler parallèlement au niveau des ontologies et dans celui de l'information (considéré comme la description des données). Si les langages de représentation diffèrent, ils peuvent être traduits d'un niveau vers un autre grâce à des projections et des relations de spécialisation ou de généralisation.
Par ailleurs, on peut également spécifier les différences de construction d'observables par des degrés de granularité. Ces niveaux ouvrent la possibilité d'observer et de lier des propriétés afférentes à chacun des niveaux.
De telles propositions requièrent de connaître exactement l'information qui circule entre ces niveaux d'abstraction, autrement dit ce que l'on met entre ces niveaux. Cette question entraînera ensuite Floridi à proposer une théorie de l'information.

Ce sont sur ces questions de précision et de portée de la théorie que nous aimerions revenir.

Le niveau d'abstraction.
Un niveau s'abstraction se définit par une collection de variables typées combinée dans un vecteur observable qui est le produit cartésien des types de variables constituantes. Le résultat est un observable complexe. Le niveau d'abstraction suppose donc un observable permettant de lier les différents types construits. Cet observable vectoriel possède les mêmes propriétés que tout observable.
(L'utilisation de l'ensemble des observables d'un corpus est lourde : on se base de préférence sur des collections d'observables de façon à obtenir des données plus simples à manier).
Les niveaux d'abstraction ne contiennent pas d'ordre sur les observables. Ils sont considérés comme les blocs de construction d'une théorie caractérisée par la définition des observables.
Si tous les observables sont discrets, alors le niveau est discret (Linguistique par exemple). S'ils sont analogiques, alors on a affaire à des sciences physiques. S'ils sont hybrides, alors on a affaire à des sciences comme les mathématiques.

Les comportements.
Les comportements caractérisent les relations entre les observables. Toutes les relations ne sont pas modélisées. Ainsi, on peut identifier des relations qui ne sont pas décrite par le niveau d'abstraction dans lequel on se situe. Les comportements font apparaître des phénomènes réguliers comme des structures et des processus, comme ceux qui caractérisent un transfert d'information.
Un comportement se définit par une prédication dont les variables libres sont observables. Les substitutions de variables qui rendent le prédicat vrai sont des comportements du système.

Les gradients d'abstraction.
Dans un système donné, les différents niveaux d'abstraction sont organisés autour de gradients d'abstraction. Il s'agit d'un formalisme défini pour faciliter la conversion de systèmes d'abstraction au-delà d'un ensemble de niveaux d'abstraction. Un niveau d'abstraction formalise la portée et la granularité d'un modèle simple. Un gradient d'abstraction permet de changer de niveau d'abstraction afin de faire des observations sur les différents niveaux. Par ailleurs, les observations faites à un niveau d'analyse peuvent être reliées à celles d'un autre niveau.

Fonctionnement de la théorie des niveaux d'abstraction.
La théorie fonctionne sur la base de raffinements progressifs, comme peut l'être l'écriture d'un article, où l'on commence par le titre, le plan, les paragraphes, les notes pour chaque phrase, etc.

L'idée consiste à construire des objets typés d'une part, et de caractériser l'interprétation des objets concernés par le fait qu'une certaine situation (prévue par un domaine de discours, un système, etc.) leur attribue une certaine propriété d'autre part. Ces objets construits sont des observables. La théorie fonctionne comme une théorie logique classique, avec une représentation prédicative et une représentation propositionnelle.

Un niveau d'abstraction est le résultat de la relation qui s'établit entre différents observables. Cette relation se caractérise par un ensemble de comportements. Un comportement est décrit par un prédicat dont les variables libres sont les observables.

Néanmoins, il arrive qu'une valeur associée à une certain observable ne produise pas les comportements recherchés. C'est parce que le niveau d'abstraction ne correspond pas à celui requis pour la prédication. Ce qui alors est explicité, c'est la réduction ou raffinement de l'approximation, donc le changement de niveau d'abstraction. Cela se fait au travers de gradients d'abstraction. Ils se définissent par des relations entre prédications, où l'une spécifie celle d'un autre niveau d'abstraction. De cette façon, on peut obtenir une représentation différente de celle qui est obtenue à un autre niveau d'abstraction.

Une **algèbre relationnelle** est une relation depuis un ensemble A vers C ; c'est un sous-ensemble du produit cartésien A x C, dont relèvent les paires (a,c).
L'inverse d'une relation R est l'image miroir :
$\{(c,a) \mid (a,c) \in P\}$
Une relation R de A à C traduit n'importe quel prédicat p de A dans un produit PR(p) sur C qui satisfait c: C qui est l'image dans R de a: A si:
$PR(p)(c) = \exists a: A\ R(a,c) \wedge p(a)$.

Le gradient d'abstraction, marqué par l'ensemble $\{Li \mid 0 < i < n\}$ et la relation $Ri,j \subseteq Li \times Lj$

pour 0 < i =/= j < n, relie les observables de chaque paire Li et Lj qui sont de différents niveaux d'abstraction.
Les relations sont inverses : étant entendu que i =/= j, Ri,j est l'inverse de Rj,i ; le comportement de pj de Lj est au moins aussi fort que le comportement traduit PRi,j (pi) ; donc pj ⇒ PRi,j (pi).
Le comportement modérant chaque niveau bas d'abstraction est consistant par rapport à celui spécifié par le niveau le plus haut. Sinon, les niveaux n'auraient pas de lien.

Si un niveau d'abstraction Li étend un autre (Lj) par l'addition de nouveaux observables, alors la relation Ri,j est une inclusion. Les contraintes imposées sur les observables Li restent vraies sur Lj. Les nouveaux observables restent hors de portée de Ri,j.
Les conditions de consistance imposées par les relations Ri,j sont en général fondées : néanmoins on risque quelques phénomènes cycliques. On caractérise alors les gradients de deux façons :
Les gradients disjoints sont complémentaires.
Les gradients emboîtés sont caractérisés par l'ajout d'information.

Pour rendre compte de ces phénomènes, il est nécessaire de caractériser la notion de **fonction** : une fonction f depuis un ensemble C vers un ensemble A est une relation et produit un sous-ensemble du produit cartésien C X A :
∀c: C, ∀a, a': A, ((c,a) ∈ f ∧ (c, a')∈ f) = a=a'.
∀c: C, ∃a: A f(c) = a.
La fonction est surjective si :
∀a: A, ∃c: C f(c) = a.

On a un gradient disjoint si les relations de Li sont vides. Donc, ils n'ont pas d'observable en commun.
On a un gradient inclus si les relations non vides sont celles entre Li et Li + 1 pour chaque 0 < i < n-1. L'inverse est une relation où chaque Ri,j+1 est une fonction surjective depuis les observables de Li+1 vers ceux de Li.

Par exemple, on peut caractériser comme gradient disjoint, les représentations des différents services dans une maison. Les comportements des différents systèmes sont explicables par ceux d'autres, mais sans que jamais ils ne se superposent.
Un gradient inclus permet de représenter le fait que toute observation abstraite a au moins une contrepartie concrète. Inversement, toute observation concrète a au moins une contrepartie abstraite: toute observation à un niveau concret provient d'un niveau abstrait.

Proposons une autre sorte d'exemple. Si l'on dispose de trois couleurs (L0) et si l'on veut caractériser les variations à l'intérieur de ces couleurs (L1), on pose une variable *wl* dont le type est un nombre réel positif correspondant à la variation de couleur. On caractérise alors le prédicat suivant :
(λa < wl <a') V (λb < wl< b') V (λc < wl<c').
La séquence représente un gradient inclus où le type le plus petit, abstrait, {a, b, c} est une projection du plus large {a', b', c'}. Formellement, la relation R (couleur, wl) fonctionne uniquement ssi pour chaque C: {a, b, c} : couleur = c ssi (λc < wl<c').

Opérations possibles sur les structures d'objets.
Les gradients permettent de passer d'ensembles disjoints vers des ensembles inclus et vice-versa :

si A et B sont disjoints, alors A U B constitue un ensemble augmenté (où A est inclus dans B).
Inversement, avec les ensembles augmentés A et B et le premier inclus dans le second, on distingue alors A et le différenciant A\B : ils sont disjoints. Un ensemble inclus peut ainsi être converti dans un ensemble disjoint. Nous renvoyons pour cela aux opérations associées aux théories de types, et qui sont présentées entre autre dans les grammaires catégorielles.

L'intérêt de cette présentation formelle de la méthode est la possibilité de définir par des moyens purement logique les objets que l'on observe. Dans le cadre de systèmes distribués, cela permet de caractériser des unités et des structures observables qui ne sont pas de même nature, comme les unités symboliques, les opérations régulières et les unités spatiales. Ainsi, on peut construire un objet d'observation qui serait fondé sur des entités hétérogènes, à condition que l'on se situe à un même niveau d'observation. Ainsi, il ne serait pas pertinent de considérer les unités symboliques étudiées comme des catégories grammaticales, ou de caractériser les opérations par leurs dimensions techniques : les niveaux d'abstraction ne seraient pas identiques. Par contre en identifiant des objets matériels et des réalisations matérielles de symboles (tels qu'ils sont écrits sur des supports), on caractérise ainsi des objets de même niveau d'abstraction. Il est alors possible de les étudier corrélativement.

Conséquences de la théorie des niveaux d'abstraction sur les niveaux d'analyse.
Cette théorie explicite le fait que certaines propriétés apparaissent à un niveau et pas à un autre. Elle explicite également le fait que différentes théories (modèles, domaines de discours, etc.), appréhendent des objets typés différemment en leur affectant certains traits précis spécifiques inopérant à d'autres niveaux d'abstraction. Elle pose donc la relativité des théories et fournit quelques règles assurant leur complémentarité.

Ces niveaux et gradients d'abstraction servent définir et expliciter comment un objet quelconque est appréhendé. Comme L. Floridi l'indique, il s'agit d'une position qui permet d'expliciter les objets que l'on prendra en amorce d'analyse et la façon dont on pourra les utiliser ensuite dans un autre cadre, celui de la proposition de modèles. Elle n'indique pas non plus quel outil théorique est le plus pertinent pour les appréhender.

Pour nous, il s'agit tout d'abord d'expliciter le modèle que l'on prend au début du travail, et ensuite la façon dont on pourra exploiter cette première description dans d'autres niveaux d'abstraction :

- quels sont les observables que l'on peut considérer comme pertinents relativement à l'information ?
- comment est-il possible de déterminer les différents niveaux d'abstraction pertinents pour construire des outils d'information, et notamment comment peut-on envisager le très problématique « niveau des connaissances » ?
- Enfin, la méthode des niveaux d'abstraction est-elle suffisante pour envisager la traduction d'usage dans des outils ?

Afin d'illustrer cette méthodologie, reprenons le contexte de la pharmacie hospitalière : nous disposons d'une première abstraction qui est caractérisée au travers de la schématisation présentée plus haut. Celle-ci caractérise les unités symboliques dans l'espace et représente les transferts ou duplication de ces entités.
Un premier travail consiste alors à étendre cette schématisation à l'ensemble des unités symboliques qui dans l'espace de l'adaptation, satisfont le même traitement. Ce premier travail permet alors de constituer des catégories de termes, que l'on pourra appeler clusters

(ou catégories de termes substituables dans un même contexte). A ce niveau d'abstraction là, on pose des variables et la schématisation n'est plus simplement descriptive. La description du système laisse place à sa modélisation. On obtient ainsi la possibilité de traiter l'ensemble des possibilités de réalisation du système.

Jusqu'à présent, nous n'avons fait que caractériser le niveau d'observation des phénomènes, sans avoir déterminé de cadre théorique.
A ce moment, il est possible d'analyser l'objet construit de deux façons :

- soit en considérant qu'une théorie de sémantique cognitive explique ces comportements. Les comportements sont alors analysés à la fois dans la dimension sémantique et de transfert (ou informationnelle).

- soit en considérant que ces comportements de transfert sont l'objet à modéliser. Ce sera alors l'objectif de la caractérisation des flux d'information.

Paradoxalement, la théorie la plus abstraite n'est pas celle qui utilise des outils de représentation de type logique et mathématique. C'est bien la théorie sémantique qui explique l'objet construit à partir de postulats sur le fonctionnement de l'esprit humain. Parmi les différents paramètres qu'elle exploite, on trouve la question de la référence.
Enfin, ces deux directions possibles sont complémentaires. En effet, une explication cognitive ne restreint pas la possibilité de caractériser un modèle qui puisse être réutilisé.

Spécification des niveaux d'analyse.
Afin de rendre son modèle plus facilement exploitable, L. Floridi propose deux niveaux d'analyse : l'organisation et l'explication.

Niveau d'organisation.
Il s'agit d'une approche ontologique, par laquelle le système analysé est supposé avoir une structure hiérarchique en lui-même, dont on considère la description et la formulation objective dans un langage d'observation neutre. Les descriptions du processus informationnel de l'adaptation de posologie mais également la caractérisation des objets du web de données sont considérés dans le niveau d'organisation.
Il y deux façons de connecter le niveau d'organisation avec les niveaux d'abstraction : si la structure hiérarchique du système lui-même est considérée comme un gradient, alors chaque constituant de niveau d'abstraction a un niveau d'organisation correspondant. Donc la méthode d'abstraction repose sur un gradient dont les niveaux constituant sont les niveaux d'organisation. Le niveau d'organisation intègre les niveaux d'abstraction contenus dans l'objet lui-même.

Niveau d'explication.
Il s'agit d'une approche commune aux sciences cognitives et à l'informatique. Elle permet de distinguer différentes approches épistémiques (comme enseignant-utilisateur ou novice). Il s'agit d'une approche pragmatique qui ne prétend pas refléter le système.

La dimension socialisée de la méthodologie est inscrite dans le cadre de schémas conceptuels.
Il s'agit de réseaux de catégories pouvant être acquises et permettant d'organiser l'expérience pour une communauté. Ils ne sont donc pas impossibles à traduire et sont inévitables pour une communauté. Les niveaux d'abstraction se situent par rapport à ces schémas conceptuels, en considérant qu'il s'agit bien de réseaux d'observables mais pas de prérogatives anthropo-centrées. (Les agents empiriques ou théoriques interagissent avec le monde à certains niveaux

d'abstraction). Ce sont des modèles du monde ou de son expérience.

Les propositions de L. Floridi ne sont pas isolées. La meilleure illustration qui peut en être faite est caractérisée par les trois niveaux d'analyse de D. Marr. Nous les avons déjà abordés dans l'interprétation qu'en fait E. Hutchins. Nous reprenons ici une formulation générique.
Tout d'abord, un système complexe de n'importe quelle sorte ne peut jamais être une extrapolation à partir des propriétés de ses composants élémentaires. Pour tout système complexe, on ne caractérise pas les effets par un ensemble très complexe d'équations, mais à chacun des niveaux distinctement.
Néanmoins, les descriptions microscopiques et macroscopiques doivent être considérées comme consistantes les unes par rapport aux autres.

Pour ce qui est des systèmes complexes, D. Marr propose :
1. Niveau des calculs.
Il s'agit de la théorie abstraite des calculs dans laquelle la performance d'un mécanisme est caractérisée à partir des propriétés abstraites de sa planification, pour lesquelles on démontre l'appropriation et l'adéquation par rapport à la tâche concernée.

2. Niveau des algorithmes.
Il s'agit du choix de représentation depuis une entrée vers une sortie incluant l'algorithme utilisé pour passer de l'un à l'autre.

3. Niveau de l'implémentation.
Niveau de représentation physique de l'algorithme. Il explicite comment l'architecture est détaillée au sein d'opérations élémentaires.

D'autres tripartitions sont possibles : sémantique, syntaxe, physique, (Z. Pylyshyn[83]) ou intention, conception, réalisation physique (D. Dennett).

En fait, ces différentes théories permettent de valider le propos selon lequel les niveaux d'abstraction constituent des outils d'organisation et de représentation d'un travail de recherche. Cela dit, comme il s'agit de caractérisations de la dimension cognitive, les niveaux d'abstraction sont caractérisés par rapport aux schèmes conceptuels.

Les niveaux d'abstraction constituent, au moins dans la version proposée par L. Floridi, des outils méthodologiques. Ils constituent une armature théorique pour envisager les mécanismes cognitifs dans les autres propositions (D. Marr notamment).

Discussion.
La proposition de L. Floridi répond semble-t-il assez facilement aux questions que l'on peut se poser : comment articuler des objets d'analyse hétérogènes, comment associer plusieurs points de vue sur un objet d'analyse complexe et dont la construction requiert l'utilisation d'outils neutres par rapport à un domaine scientifique particulier. Nous présentons maintenant quelques problèmes et cas d'application de cette théorie sur nos exemples.

Quelle que soit ensuite la stratégie que l'on choisit pour déterminer ce qui sera effectivement soumis à l'abstraction, le principe du typage requiert quelques commentaires : le typage permet la mise en série de phénomènes, et donc l'observation de phénomènes réguliers, répétitifs, dans cette activité. Cela est facile à observer sur une activité reposant sur des circuits d'information comme ceux que l'on observe au sein de la pharmacie hospitalière : la

variabilité de chaque suivi thérapeutique est inscrite à l'intérieur d'un cadre régulier de production, circulation et interprétation d'information.
Comme on a pu le mentionner dans notre discussion à propos des individus, les comportements ne peuvent être prédits avec certitude. Comment explique-t-on alors l'assurance et la maîtrise du suivi par le pharmacien ? En considérant que la situation est du même type quels que soient les patients individuels, il devient possible d'établir un raisonnement régulier mais alors, comment peut-on prendre en compte la variabilité individuelle ? La seule réduction de cette variabilité a consisté pour les pharmaciens à établir des populations statistiques de patient, et à tester leurs hypothèses de doses sur ces populations, dans lesquelles le patient individuel concerné a été préalablement inscrit.

Le typage a aussi un autre intérêt : il permet l'extension d'une observation, et sa transposition, par le plus haut niveau d'abstraction, à d'autres situations.

Les unités symboliques (unités linguistiques mais également marqueurs spatiaux) illustrent plus particulièrement cette capacité à être saisies à la fois comme instance et inscrites dans un type. Ainsi, les régularités du haut niveau d'abstraction peuvent autant être observées que les variations du bas niveau. Ce ne sont pas les mêmes propriétés qui sont observées dans les différents niveaux d'abstraction.

Enfin, cette démarche n'est pas sans rappeler (voire même fournit un cadre conceptuel) pour les travaux menés dans le cadre du web de données où les unités lexicales doivent être considérées à la fois au niveau linguistique et conceptuel. En ce sens, la perspective ouverte ici constitue un outil fondamental pour traiter de la pluridisciplinarité et de la complémentarité entre différentes approches. Par exemple, à quelles conditions peut-on intégrer des métadonnées dans une ontologie, ou comment corréler une approche de lexiques sémantiques et une analyse conceptuelle ? Ces questions essentielles du web nécessitent une identification des niveaux d'abstraction de chacune des analyses par rapport à un objet construit. Comme on l'a déjà suggéré, le web de données accueille et utilise des structures de données symboliques allant des lexiques de la langue naturelle vers des langages normés et contrôlés particulièrement structurés. La façon dont on peut représenter ces langages et leurs relations requiert une caractérisation de leur niveau d'abstraction réciproque.

Niveaux d'abstraction et explication.
Si l'on définit l'abstraction par la capacité à produire des explications et des représentations à portée plus ou moins générale, et ce indépendamment du langage utilisé (mathématique, langue naturelle), alors le niveau d'abstraction le plus élevé est celui qui explique le plus large ensemble de faits.
Nous suivrons le principe général de structuration suivant :

- le niveau des réalisations caractérise des instances (qu'il s'agisse d'opérations ou d'unités lexicales).
- Les régularités observables entre ces instances (donc l'observation de variables) définissent des types.
- Les relations régulières entre ces types manifestent des théories explicatives.

Le passage d'un niveau d'abstraction à un autre consiste à élargir la portée de la proposition initiale en lui adjoignant une dimension informationnelle : l'abstraction d'une instance permet d'informer sur le type de l'instance, et la théorie permet d'informer sur le rôle du type dans le processus en question. On procède ainsi pour aboutir à une hypothèse sur la dimension cognitive, qui nous apparaitra comme étant la plus générale qui soit : l'hypothèse cognitive

constitue le niveau d'abstraction maximal de l'analyse. Elle ne pourra pas être rigoureusement démontrée, et constitue donc la limite de notre travail.

Nous appliquons ainsi les principes des niveaux d'abstraction à l'élaboration du modèle que nous proposons. Comme on l'a déjà annoncé, on utilise la même méthode en analyse et en conception. Ainsi la théorie des flux a une portée initiale relativement limitée : elle représente des relations entre des entités structurées hétérogènes, relations constituant des contraintes entre des situations. Néanmoins, ces flux reposent partiellement, comme nous le verrons sur une théorie des classifications. Cette théorie peut être intégrée dans le cadre méthodologique des niveaux d'abstraction. C'est en partie ce que nous nous emploierons à démontrer plus loin.

Extension de la proposition vers les outils d'information.
On présente maintenant la façon dont il est possible de rendre compte de cette distinction de niveaux dans le cadre des systèmes d'organisation de connaissances. On pourra ainsi expliciter les distinctions entre les systèmes d'organisation de connaissances (de type thésaurus par exemple) et les modèles de formulation de connaissances (initiés en psychologie cognitive mais généralisés ensuite en IA sous l'appellation « représentation des connaissances »), et considérer quels déplacements opère le web sémantique. En effet, tant dans leurs finalités que leurs modalités d'élaboration, ces deux types de caractérisation formelle des connaissances sont éloignés. Le premier repose sur un consensus social, le second sur une explicitation, voire une verbalisation de connaissances implicites. Or, des outils de représentation comme les recommandations SKOS permettent de mettre en relation des structures hétérogènes de représentation de connaissances, et par conséquent permettent d'envisager des utilisations complémentaires.

Les niveaux d'abstraction et les transferts constituent certes des outils permettant de caractériser une méthode, cette méthode permet de caractériser plus précisément ce qu'est l'information.
Par ailleurs cette question des niveaux d'abstraction reste implicite mais essentielle dans de nombreux points cruciaux de la structuration du web. Ainsi, les questions relatives au choix de classification ou de thésaurus, distinctement des ontologies, pour définir des profils, constitue un problème de niveau d'abstraction.

Par ailleurs, ces questions de niveaux d'abstraction sont au centre des langages documentaires (de type générique/spécifique), à la différence par exemple de l'annotation et de la description de textes où il n'est pas nécessaire d'abstraire au-delà d'un modèle de la prédication. Dans les langages documentaires, le problème est celui du lien entre les descripteurs : si les logiques booléennes permettent de rendre compte des inclusions, exclusions et autres relations, notamment les ingrédients, il n'en reste pas moins qu'il s'agit d'outils distinctif fonctionnant sur la base d'appartenance et d'exclusion et non de distinction de niveau d'abstraction. La relation entre deux unités de même niveau est toujours considérée par rapport au niveau supérieur, en ce qu'il l'inclut empiriquement ou pas.
Dans l'exemple suivant, issu du thésaurus Santépsy, la hiérarchie n'est pas fondée sur une abstraction progressive, mais sur un même niveau d'abstraction sur lequel on définit des descripteurs de thématique d'une portée plus ou moins large :

ASSURANCE MALADIE
DEPENSES DE SANTE
MAITRISE DES DEPENSES DE SANTE
OBJECTIF NATIONAL DES DEPENSES D'ASSURANCE MALADIE
REFERENCE MEDICALE OPPOSABLE
REMBOURSEMENT DES SOINS
TICKET MODERATEUR
TIERS PAYANT

Il n'y a pas de hiérarchie non plus dans les annotations, mais seulement une structure d'outils descriptifs. On peut rapprocher les annotations des métadonnées, à la différence près que les métadonnées décrivent des documents alors que les annotations constituent des indices d'appartenance dans des parties de document.
Ainsi, les questions se posent très différemment entre ces deux types d'outils descriptifs : d'un côté, les thésaurus sous-tendent une logique propositionnelle, et les annotations sont de type prédicatif.

Néanmoins, le problème se pose lorsqu'il s'agit de mettre en relations ces différentes structures. Il est nécessaire d'envisager en quoi leur approche des connaissances et de la langue diffèrent et de quelle façon des usages et pratiques différents (ou schèmes conceptuels) peuvent être conciliés.

Ce qu'apportent les outils du web, c'est l'importance des dimensions lexicales et plus généralement linguistiques. En rendant possible techniquement une mise en relation de ces structures hétérogènes par des standards adaptés, le W3C oblige à penser les conditions et les contraintes à ces mises en relation. En effet, tout autant que les métadonnées, les terminologies, notamment via le standard TBX (Term Base eXchange)[188] utilisent les standards XML. Par ailleurs, des travaux comme ceux de LEXONTO, visent à établir des corrélations entre analyse lexicale sémantique et analyse conceptuelle, via les ontologies.

Au-delà, ces questions renvoient à la caractérisation des différents niveaux d'abstraction requis par chacune des sciences, et donc dans ce cadre, de la caractérisation de l'abstraction portée par les thésaurus. Les niveaux d'abstraction permettent d'expliciter ces différentes approches de mêmes unités linguistiques.

En définitive, l'intérêt de ces niveaux réside d'abord dans la méthodologie pour approcher la conception d'outils documentaire sachant que ces outils vont être reliés à d'autres, de différents niveaux d'abstraction, et qu'il faudra gérer les relations signifiantes entre chacun des niveaux. Dans ce cadre, de nombreux travaux commencent à voir le jour afin de comparer ces différentes méthodes d'organisation et de caractérisation des connaissances[84]. S'il s'agit dans ces travaux de montrer comment les ontologies enrichissent les systèmes de classification et de structuration des connaissances, ils supposent que l'on sache à quel niveau chacun appartient.
Ainsi, par exemple, une telle clarification permet de poser la question de l'intégration de métadonnées dans une ontologie. Dans la méthodologie la plus commune, les ontologies se caractérisent par des niveaux d'abstraction successifs depuis les lexiques, taxonomies et terminologies. Ces termes constituent les fondements linguistiques de bas niveau

[188] http://www.ttt.org/tbx/,

d'abstraction des ontologies[189]. Les relations proposées par l'ontologie constituent alors le haut niveau d'abstraction. Dès lors que l'on propose d'insérer des métadonnées à l'intérieur d'une ontologie, à un niveau qui n'est pas celui des lexiques, il devient nécessaire de caractériser la nature et la légitimité de cette intégration. Par ailleurs, si l'on souhaite intégrer dans une ontologie des outils de raisonnement, comme les flux, il est nécessaire de définir à quel niveau chacun peut être décrit et ainsi spécifier comment cette intégration est possible.

Des outils de description de l'information vers les questions de transfert d'information.
Les niveaux d'abstraction caractérisent les conditions et contraintes pour structurer les relations entre différentes structures de données symboliques. Les niveaux d'abstraction servent également à justifier l'utilisation d'une représentation formelle pour caractériser des phénomènes observables empiriquement (et par ailleurs comportant des composants hétérogènes). La théorie des flux, qui constitue une représentation formelle, utilise le même outil de caractérisation des données que la théorie des niveaux d'abstraction : il s'agit de la théorie des types. Elle peut donc être utilisée à la fois pour représenter un phénomène régulier et pour construire un modèle.
Les niveaux d'abstraction permettent également d'entrevoir la dimension philosophique de ce formalisme puisqu'ils servent à montrer comment deux systèmes classificatoires distincts peuvent effectivement s'enrichir mutuellement par un raisonnement.
Cette distinction permet de spécifier la portée de notre projet ; dans un premier temps, il propose de décrire ce qu'apporte une structuration à une autre (et non l'organisation d'une structure), et dans un second temps, pose un certain nombre de principes méthodologiques pour avancer sur la question de la mutualisation et de la complémentarité des outils d'organisation et de structuration des connaissances. Cette question est particulièrement d'actualité puisque techniquement les standards du web rendent possible cette articulation.

2.3.3. Les logiques de transition comme régulation du transfert d'information.

Les niveaux d'abstraction servent à spécifier le degré d'abstraction des descriptions, et les propriétés relatives à chacun de ces niveaux. Néanmoins, ils n'explicitent pas le raisonnement permettant de faire circuler des informations d'un contexte descriptif à un autre, et dans ce cadre, d'enrichir chacune des descriptions. Qu'est-ce qu'apporte cette information ? Qu'est-ce qu'elle change ?
Cette question est différente de celle des niveaux d'abstraction parce qu'elle traite de la conséquence de la mise en relation de plusieurs niveaux d'abstraction : quelles sont les contraintes qui se posent lorsque l'on veut s'assurer d'une transition effective des résultats ?
Les logiques de transition tendent à répondre à ces questions : il s'agit de logiques dynamiques relationnelles. Leur spécificité tient dans une insistance sur le passage de données d'un domaine vers un autre[85].

Les logiques de transition sont, à la différence des niveaux d'abstraction, relativement éloignées des préoccupations des sciences de l'Information, voire même des applications web

[189] Ce modèle est le plus fréquent et est commun aux perspectives réalistes et conceptuelles. Or, à la suite des critiques de Bremer, les perspectives plus linguistiques, notamment proposées dans CyC, se retrouvent confortées.

sémantique. Ce sont pour le moment, essentiellement des outils logiques et mathématiques ayant des applications éventuelles dans le traitement automatique du langage.

Le but de cette partie consiste à expliciter les conditions et contraintes par lesquelles on pourra décider si une mise en relation de données structurées caractérise effectivement une information, considérée comme un outil servant à changer d'état. Les logiques de transition ne constituent pas des outils d'analyse ou de représentation de l'information, mais des outils de contrôle. En d'autres termes, ils servent à caractériser de quelle façon l'information a modifié l'état des connaissances.
Nous avons essentiellement parlé jusqu'à présent de classification et d'abstraction, mais pas des relations portées entre les différentes classifications. On vise maintenant à caractériser les relations qui peuvent être représentées entre les différentes structures de données. Ces relations constituent l'arrière-fond dans lequel les flux et les structures d'information prennent sens. Plus généralement même, l'information prend sens dans une dynamique.

Si l'on poursuit plus loin, on pourra considérer que ces logiques dynamiques caractérisent le contexte dans lequel l'information s'inscrit. Lorsqu'il est question de contexte, on fait référence à ces processus. Comme nous le verrons, les logiques dynamiques rendent particulièrement bien compte de la conversation et plus globalement des activités communicationnelles.

Présentation générale des logiques dynamiques.
On caractérise le concept de dynamique par sa formulation logique, qui nous semble effectivement la plus pertinence notamment du fait de son objet d'application : les programmes. (On ne caractérise pas les programmes d'un seul point de vue informatique, mais au sens d'un processus maîtrisé quant à ses finalités et ses étapes).

Les logiques dynamiques sont le produit d'un tournant important en logique, décrit par L. Floridi & P. Allo[190]. En effet, les problématiques des valeurs de vérité, de signification et de perception, qui du point de vue de l'information traitent de son rapport à la connaissance, sont insérées dans des modèles dynamiques. Ceux-ci intègrent les traits communicationnels et épistémiques. C'est dans ce cadre que nous souhaiterions situer notre travail.

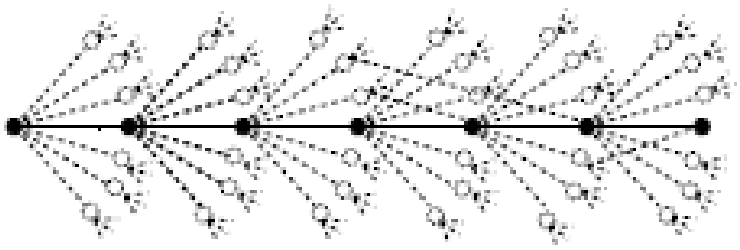
Les logiques dynamiques consistent à raisonner à propos de programmes, et non plus comme dans la logique classique des prédicats, sur une vérité statique. Alors que les logiques classiques fondaient leur propos sur des expressions (comme par exemple « le roi de France est chauve »), les logiques dynamiques changent d'objet de travail : on raisonne maintenant sur des suites d'opérations. Ces suites d'opérations sont appelées programmes parce qu'elles ont les mêmes propriétés de finitude et de régularité que les programmes informatiques.
A la base, les programmes changent les valeurs des variables, donc les valeurs de vérité des formules. Les variables sont considérées en fonction des valeurs d'entrée et de sortie des programmes.

Si l'on veut mieux caractériser plus précisément l'enjeu de ces logiques dynamiques, il convient d'en présenter la source. Le principe de base de ces représentations est formulé par S. Kripke[86], et la succession d'états qu'il propose. Ce modèle rend compte autant de la dimension des choix lexicaux que des opérations d'engendrement de structures.

[190] Allo, P. & Floridi, L., Introduction, *Logique et Analyse*, 196, 2006, pp. 343-344.

Enfin, comme il s'agit de représenter le choix d'une variable réalisée par rapport à un ensemble possible, le lien à la théorie de l'information (le choix d'une entité dans un contenu possible) n'est pas fortuit. La formulation de S. Kripke peut être considérée comme l'articulation de la caractérisation de l'information et celle des programmes.

M. Steedman[87] (p.24) propose une représentation qui explicite ce modèle :



Chaque éventail représente un ensemble de choix possibles à partir d'un état. Les choix en pointillé représentent ceux qui n'ont pas été réalisés. Inversement, un état ne peut être expliqué que par une seule cause. La causalité constitue un point essentiel sur lequel nous reviendrons ultérieurement, parce qu'il s'agit d'une explication fondamentale des inférences réalisées à l'intérieur des flux.

Le modèle de S. Kripke est une proposition générale pour caractériser les modalités d'engendrement d'une succession d'états à partir d'opérations de choix. Ce modèle rend compte autant de la dimension des choix lexicaux que des opérations d'engendrement de structures.

Un tel modèle sert notamment à caractériser des opérations compositionnelles dynamiques, telles qu'elles peuvent être caractérisées dans le cadre d'une grammaire.
Successions d'état, choix, dynamique du traitement de l'information : le modèle proposé par S. Kripke a le double intérêt de se situer aux frontières de la modélisation du langage, du temps et des systèmes. C'est l'ensemble de ces dimensions que l'on va aborder maintenant, considérant la succession des états comme une caractérisation possible de la dynamique de l'interprétation, qui trouverait dans l'information une segmentation particulièrement opportune. Si l'on définit la structure d'information comme un ensemble des constituants ordonnés ayant fonction de transmettre un contenu sans ambigüité et de permettre l'interprétation d'une information déclenchant un changement d'état, alors la structure d'information constitue une dimension pertinente de la caractérisation dynamique des discours.
En effet, l'une des principales qualités de la structure d'information, c'est de proposer une segmentation qui permettrait de se dégager de la structure de la phrase, et ainsi d'observer des relations nouvelles entre entités langagières réalisées.

Plusieurs familles de travaux vont utiliser le modèle de S. Kripke et ses perspectives dynamiques :

- les logiques de transition
- les grammaires catégorielles[88]
- les sémantiques des prédicats dynamiques.

Toutes trois ont effectivement un lien avec les questions d'information et de structure d'information, parce que les dynamiques s'appliquent à une segmentation pertinente des expressions : le cadre des dichotomies ou trichotomies de la structure d'information apparaît nettement plus pertinent que les catégories usuelles de la syntaxe pour définir ce comment une structure symbolique transforme un état dans le cadre d'un transfert.

En ce sens, un intérêt essentiel de la structure d'information repose sur sa dimension structurale (un ensemble d'entités inter-définies dans un cadre fonctionnel explicite comme par exemple la segmentation de la clause en [thème/rhème] ou en [fond(lien/grandeur)/focus]), en opposition à la compositionnalité, à savoir l'addition des propriétés des différents composants de la structure pour caractériser les propriétés de cette dernière.

Le modèle dynamique ne permet pas seulement de représenter des choix et de caractériser, pour un fait réalisé, l'ensemble des autres faits possibles à partir de l'état précédent. Il permet de caractériser comment un état est une conséquence d'un autre, et par conséquent porte son précédent. En d'autres termes, il permet d'envisager à la fois les contenus et les dualités[191] de signification.

Les logiques dynamiques récentes sont fondées sur les propositions de D. Harel[89] et permettent de rendre compte du discours comme contexte dans lequel s'inscrit une dynamique. La puissance de ce modèle permet par exemple de ne pas avoir à prendre en charge la syntaxe pour positionner une analyse sémantique.
La dynamique peut être le dialogue, ou encore le discours, voire l'événement. Simplement, ces théories présupposent que le contexte lui-même soit défini comme un ensemble d'opérations, et non par un système. (Or, c'est justement ce qui caractérise notre contexte : il est particulièrement régulier, dans le cas précis de la pharmacie hospitalière, se caractérise par un programme : le choix de la meilleure posologie pour un patient précis. Dans le cadre de notre projet, l'enrichissement d'une description documentaire est bien un programme. Il est réalisé par un raisonnement).

Les logiques dynamiques permettent l'expansion des principes des grammaires catégorielles aux questions de traitements d'informations.[90]
Ce sont des logiques qui reposent sur la distinction entre d'une part des opérateurs dynamiques d'inférence et des états. Elles requièrent des opérations (nécessaires ou possibles) et des formules (lexicales), ces dernières caractérisant des états.
On a à la fois des propositions et des programmes (c'est une logique d'instructions). Les programmes sont la part dynamique de la logique.

Représentation des formules et des opérations (in R. Muskens, J. van Benthem and A. Visser, "Dynamics", in van Benthem, J. F. A. K., & Ter Meulen, A. (Eds.). (1996). *Handbook of logic and language*. Op.cit, p. 616) :
Définition des formules : $\varphi :: = | \varphi 1\ \varphi 2 | [\pi]\varphi$.
Définition des opérations : $\pi :: = \alpha | \varphi ? | \pi 1 ; \pi 2 | \pi 1\ U\ \pi 2 | \pi *$

Caractérisation des programmes.
Les programmes constituent des outils assurant le passage d'un état vers un autre. Ils se caractérisent par des tests [φ?] qui constituent des programmes par lesquels si la réponse est

[191] Les dualités de signification renvoient à l'interprétation en extension et en intension.

positive, le programme continue (et s'arrête sinon).
La séquence $\pi 1 ; \pi 2$ caractérise le fait que si le premier programme est réalisé, le second peut l'être.
Le choix de programme $\pi 1 \cup \pi 2$ permet de caractériser l'exécution d'un programme soit par $\pi 1$ soit par $\pi 2$.
L'itération $\pi *$ caractérise le fait que le programme puisse être réalisé autant de fois que nécessaire.

Les programmes, en ce qu'ils sont choisis et itératifs, introduisent le non- déterminisme dans les processus langagiers. En effet, une opération peut être réalisée ou pas, voire une même l'opération réalisée de façon itérative. Avec de tels points de départ, il est aisé de rendre compte de la variabilité du langage ou du cours d'action dans une activité professionnelle.

Les systèmes de transition labellisés.
Ils constituent une part essentielle des logiques dynamiques.
Ils sont fondés sur des modèles polymodaux : $<S, \{R\alpha \mid \alpha \in AT\}, V>$
- états S de programmes.
- $R\alpha$: ensemble de relations dans S indexées par un ensemble de programmes atomiques AT.
- valuation V qui assigne fonctionnellement un ensemble de valeurs (sous-ensemble de S) à chaque lettre propositionnelle dans le langage.

Néanmoins, les actions atomiques sont inscrites dans un répertoire procédural d'opérations créant des actions complexes.
Les opérations procédurales ont une composition séquentielle, des principes de choix et reposent sur des itérations. Elles peuvent être comparées aux fonctions dirigées de la grammaire catégorielle :

$A\backslash B = \{<x,y> \mid \forall z(<z, x> \in A \rightarrow <z, y> \in B)\}$
$B/A = \{<x,y> \mid \forall z(<y, z> \in A \rightarrow <x, z> \in B)\}$

Sémantique des logiques dynamiques.
La signification d'une formule est identifiée avec l'ensemble des états où elle est vraie.
La signification d'un programme est l'ensemble des paires <a, b> telles que le processus, s'il démarre dans un état a, se termine dans un état b.

Néanmoins, cette caractérisation de la dynamique demeure largement modale au sens où l'interprétation d'un état est celui de $[\pi]\varphi$ (d'un programme à propos d'un état). Evidemment, le programme est modal, ce qui permet à M. Steedman d'utiliser la représentation modale de S. Kripke pour caractériser ses propositions relatives à l'événement. Un événement se caractérise par le passage d'un état à un autre, sachant que la sémantique de l'expression intègre les états et les formules.
C'est donc au travers de la succession des états réalisés que le programme prend sens. La signification est caractérisée par son élaboration au travers des opérations successives.

Logiques dynamiques et conversation. Distinction entre dynamique communicationnelle et fonction informationnelle.
J. Van Benthem[91] considère l'émergence des logiques dynamiques à partir de la multiplication des acteurs inscrits dans les processus informationnels. L'information est inscrite dans des activités quotidiennes et donc intègre les questions de communication dans celles d'information, avec notamment les actualisations de connaissances. C'est d'ailleurs l'un des problèmes de ce type d'approche : est-ce qu'effectivement on caractérise

l'information, ou certains problèmes logiques qui impliquent l'information (comme la croyance et l'actualisation) ?
L'actualisation et la révision des connaissances se substituent à la diffusion d'informations nouvelles. Elles reposent sur l'élimination d'états qui ne seraient pas validés par des prémisses. L'information ne consiste pas, et cela est en lien avec la définition probabiliste de l'information, à apporter un fait entièrement nouveau, mais à distinguer des états qui se vérifient par rapport à d'autres, à un certain moment et dans un certain contexte.
J. Van Benthem associe systématiquement cette révision à la réalisation d'une action, qui constitue la condition de sa réussite ; ainsi, les logiques dynamiques reposent en grande partie sur l'analyse des actes de langage. L'action permettant d'actualiser les états est baptisée « annonce publique ». Deux aspects essentiels de ces travaux, qui sont d'une part l'action représentée, d'autre part la contraction d'informations dans le cadre de l'interprétation (et non des procédures d'abstraction des données).
L'écriture d'une annonce publique prend la forme suivante :
"[A!]$K_j \phi$ after a true public announcement of *A*, *j* knows that ϕ."
L'écriture complète de l'annonce est la suivante :
$[A!]\ K_j \phi \leftrightarrow (A \rightarrow K_j(A \rightarrow [A!]\ \phi)$

Ces formules permettent de rendre compte de deux processus qui jusqu'à présent n'ont guère pu être intégrés dans la présentation que nous avons pu proposer de l'information. Il s'agit en l'occurrence de phénomènes dynamiques, au sens où ils ont trait non au système lui-même, mais à la façon dont le cadre de l'activité peut être envisagé.
Ils caractérisent le cadre dynamique dans lequel l'information a un usage. Elles permettent donc de caractériser très précisément un processus communicationnel d'une fonction informationnelle. On considère donc, conformément à ce que l'on a énoncé plus haut, que ces logiques traitent du contexte d'usage et donc de l'interprétation de l'information, et non de l'information elle-même.

Pertinence des logiques de transition pour l'analyse des transferts d'information.
Comme on l'a présenté plus haut, l'intérêt de ces logiques réside dans la formulation d'un cadre pour appréhender les phénomènes de transfert d'information, que l'on caractérisera plus précisément par des flux d'information. Les flux ne représentent pas cette dynamique. Ils sont seulement descriptifs des mouvements d'information qui alimentent ces dynamiques.

Ces logiques dynamiques ne servent pas à expliciter le raisonnement qui fonde les flux ni à le représenter, mais le contexte dans lequel les flux trouvent une pertinence. Le raisonnement que l'on cherche alors à caractériser est celui des flux, mais ses effets produits sont inscrits dans le cadre de dynamiques, qui peuvent être communicationnelles.
La logique que l'on vient de présenter est construite pour raisonner sur des outils d'analyse et de représentation des connaissances. Ainsi, les exemples pris par P. Blackburn[92] sont les logiques de description, les logiques de traits et les logiques temporelles[192]. Les logiques

192 Les théories du temps et de l'action proposent de caractériser le processus de prédication au travers d'opérateurs temporels. La théorie de l'action et du temps de J.F. Allen (Allen, J. F. (1984). Towards a general theory of action and time. *Artificial intelligence*, *23*(2), 123-154. constitue une modélisation des plans, des événements et des faits originaux occurrents.

Cette théorie emprunte donc à la fois à la logique prédicative associée à la représentation des expressions du langage et à la modélisation du raisonnement humain en situation d'activité (résolution de problèmes, tâches, etc.). Néanmoins, si elles décrivent le changement au travers de distinctions temporelles, elles n'expliquent pas les opérations associées aux dynamiques, comme peuvent l'être les conversations. C'est pour cela que les logiques temporelles peuvent être intégrées dans les logiques dynamiques.

dynamiques explicitent ce qu'il se passe lorsque d'autres logiques sont appliquées. Elles sont donc d'un plus haut niveau d'abstraction. Pour les flux, il s'agit d'expliciter la productivité de l'inférence entre deux structures de données hétérogènes, dans un contexte précis et en montrant en quoi ce contexte est modifié.

Les logiques de transition prennent comme objet des connaissances liées à des états. Les flux caractérisent le maniement d'objets symboliques, ces symboles étant inscrits dans ces états. De cette façon, l'articulation de la logique caractérisant le raisonnement à celle spécifiant le cadre de ce raisonnement (son contexte de validation) repose sur une distinction simple des objets de chacune.

Conclusion.
Nous avons jusqu'à présent caractérisé un cadre de travail pour appréhender les phénomènes informationnels. Nous ne développerons pas plus avant les questions logiques dans la mesure où elles ne constituent pas le centre de notre travail. Par contre, nous avons proposé des références pluridisciplinaires pour aborder les phénomènes que nous avons évoqués en introduction. Nous les développerons plus précisément après avoir explicité les concepts fondamentaux d'information, de flux d'information et de structure d'information.

Par ailleurs, nous avons posé quelques règles de façon à contrôler la pertinence, la légitimité et la scientificité des principales opérations de description puis de conception que nous pensons effectuer. Ces règles s'appliquent aussi et surtout à la mise en relation et au report des résultats depuis les descriptions vers la conception et le changement de domaine qui est relatif. Pour cela, nous avons choisi un cadre logique et philosophique limité aux phénomènes propres que nous étudions. Maintenant, après avoir construit ce cadre de travail, il convient de définir les concepts fondamentaux et donc des modèles permettant de construire les observables. Nous essayons de formuler une définition précise de l'information sur laquelle nous nous appuierons pour ensuite développer la question des flux, qui en constitue une partie.

PARTIE 3. Définitions de l'information.

On peut avancer que les définitions mathématiques de l'information ont comme avantage de permettre une certaine mesure de l'information (Van der Lubbe, op. cit.) et donc lui associe une certaine quantité. Or, dès lors que l'on cherche à expliquer l'information, le problème devient immédiatement celui de sa dimension sémantique (pas nécessairement entendue au sens linguistique).
La caractérisation de l'information devient alors beaucoup plus complexe, entre autre parce qu'elle articule des dimensions sociales, technologiques et linguistiques. Cela se traduit aussi par la décomposition de la question de l'information en plusieurs problématiques liées : flux, structure, représentation, organisation. Que l'on soit dans la perspective des ontologies, dans celle des métadonnées, de l'extraction d'information ou encore des relations entre structures interopérables, il est systématiquement question d'information. Or, ces définitions proviennent de sources différentes. Nous proposons ici un cadre conceptuel permettant d'assurer la cohérence entre les différentes parties de notre travail.

Cette question est loin d'être anodine dans le cadre des bibliothèques numériques. En effet, quelle est la nature des métadonnées : s'agit-il simplement d'indications techniques sur un document, au même titre que des notices bibliographiques, ou au contraire, s'agit-il de propositions se validant sur un objet ?
Notre choix est bien la seconde possibilité, donc l'association aux métadonnées de portées de signification et de représentation de faits du monde. Dès lors, comment appréhender ces objets signifiants ?
S'il s'agit d'abord de décrire l'information qui leur est associée, ensuite on pourra au mieux caractériser comment améliorer ces métadonnées, en exploitant au mieux leurs capacités prédicatives. On pourra étendre ce questionnement aux ontologies et aux différents langages documentaires.

On prendra comme base pour l'étude de l'information les axiomes suivants :

1. le contenu de toute expression informationnelle contient une nouveauté (qui constitue l'apport d'un fait inédit) ≠ redondance.
2. Le fait que cette nouveauté soit conforme à un univers de croyance : pour celui qui interprète, que ce fait corresponde à une vérité ≠ univers délirant. (On représente alors un fait dans le monde, probable, possible ou avéré).

Ces deux axiomes sont plus ou moins fondamentaux dans le développement les différentes théories de l'information (Voir Peter Gardenfors[93]) :

- soit on se fonde sur les théories du transfert et donc on se préoccupe d'abord de comprendre comment l'information peut être produite, transmise et interprétée,
- soit on considère que l'information est fondée sur une capacité d'interprétation, donc, on se fonde sur le contexte d'interprétation pour envisager ce qui est commun entre l'expression et l'interprétation.

Dans la théorie de la transmission, on se fonde sur l'idée d'une représentation attestée de l'information : l'information est matériellement descriptible. (Elle peut donc être, enregistrée, stockée, archivée). Cette approche requiert des outils scientifiques susceptibles de caractériser ce contenu, comme les modèles linguistiques ou de psychologie cognitive. (En effet, l'information que l'on considère est associée à une mémoire ; voir Van Der Lubbe, op. cit.).

Dans le second cas, l'information est identifiée dans un contexte, donc le contenu informationnel est dépendant du contexte d'interprétation (individuel et social). Elle accueille des démarches fondées sur l'herméneutique.

Nous reprendrons au départ les définitions les plus courantes de l'information pour expliciter la trajectoire et les lignes de force de la réflexion à propos de l'information dans le cadre que nous avons fixé (3.1). C'est dans ce cadre que l'articulation entre une définition propositionnelle et une définition prédicative de l'information émerge (partie 3.2.). Les flux constitueront alors la définition prédicative de l'information, distinctement d'une définition propositionnelle, associée pour nous à la théorie des situations (3.3).

3.1. Télécommunication vs informatique. Langage et discours.

Les définitions de l'information sont relatives au domaine dans lequel le concept apparait pertinent ; ainsi, on peut facilement distinguer des définitions philosophiques, juridiques, journalistiques ou biologiques de l'information. Dans le cadre des Sciences de l'Information et de la Communication, les approches sont également diversifiées[94].
Pour la clarté de l'exposé, nous retiendrons deux définitions initiales de l'information ; la première est issue des télécommunications et appréhende l'information comme une unité formelle calculable à l'intérieur d'un réseau. La seconde est issue du monde de l'informatique et considère l'information comme une donnée dotée de sens. (Cette seconde définition prend sa source dans l'association entre donnée et logique que propose Türing).

La théorie proposée par Shannon peut apparaître relativement inappropriée, notamment lorsqu'il s'agit de caractériser un contexte de communication. Par contre, un aspect de sa théorie constitue un « fil rouge » pour caractériser l'évolution des théories de l'information : il s'agit du principe de choix.

Ces définitions de départ seront reprises par Carnap et Bar-Hillel[95] dans le cadre d'une théorie sémantique, fondée sur la notion de choix (elle-même utilisée par Shannon) et donc la prédication. Ainsi le choix constitue une notion commune à la fois aux modèles statistiques de l'information et aux théories de la prédication. Cette notion de choix permet à la fois de calculer l'informativité (plus un choix est rare, plus il sera informatif) et de calculer la dépendance d'un mot par rapport à un autre.

Notion de choix et sémantique des unités lexicales.
Cette dépendance constitue alors l'un des principes de la sémantique structurale, telle que la signification d'un terme est donnée par l'ensemble des autres termes avec lesquels il entre en relation. Ces propositions serviront à caractériser les classes de distribution des entités lexicales, lesquelles constituent alors la signification de l'entité lexicale considérée. Il s'agit du fondement de la grammaire lexicale (voir les propositions de Maurice Gross[96]). Une telle construction est également intéressante parce qu'elle permet de mettre en évidence la dynamique des discours[97].
C'est également ce principe de choix qui servira à construire les théories discursives de l'information.

Choix et structure d'information.
La notion de choix doit être articulée à celle de structure, qui s'est imposée plus récemment dans l'analyse des discours, et qui propose de considérer l'information comme un contenu

nouveau associé à une connaissance partagée, reprenant par-là certains découpages des expressions linguistiques comme le rapport thème/rhème. (M. A.K. Halliday[98] entre autres).

Pour M.A.K. Halliday, le rapport thème/rhème construit la structure d'information et ne constitue qu'une part de la fonction d'une clause (qui constitue pour lui la dimension de référence de l'observation des réalisations discursives) ; il s'agit de la fonction informationnelle, distincte de la fonction cognitive et de la fonction communicationnelle.

Lien entre ces définitions prédicatives de l'information et les organisations de connaissances.
Aucune de ces précédentes définitions n'apparaît totalement complète et appropriée aux Sciences de l'Information : le rôle central prêté dans le cadre de ces propositions à la prédication semble éloigné des questions de classification, de description documentaire et de recherche d'information (c'est également le point de vue de Floridi[99]). En effet, les questions de production des discours et de distribution des unités symboliques ne caractérisent l'information qu'au travers de la prédication. Si effectivement celle-ci reprend de façon opportune le principe de choix formulé dans la théorie de l'information fondée sur les télécommunications, elle ne peut rendre compte aisément des questions posées par la description documentaire et les différents outils de classification.

On peut expliquer que ces précédentes définitions de l'information sont aussi divergentes parce qu'elles ont des fonctionnalités totalement différentes : soit une mesure, soit une dimension de la caractérisation sémantique des unités linguistiques, soit un outil descriptif du discours. Les entités symboliques ne sont pas considérées non plus du même point de vue.

Par ailleurs, ces définitions de l'information peuvent ne pas apparaître satisfaisantes dans le cadre de la structuration du web parce qu'elles n'intègrent pas la possibilité de structuration préexistante des connaissances. Chacune de ces propositions présuppose une certaine organisation de l'expérience, plutôt explicitement dans le cas de M.A.K. Halliday, beaucoup moins dans celui de Z. Harris. Parce que ce n'est pas dans leur programme de travail, ces propositions n'intègrent pas l'explicitation du cadre cognitif qui spécifie les contenus transmis et caractérise la structuration sémantique requise. En d'autres termes, ces définitions de l'information apparaissent insuffisantes : les choix sont de nature prédicative, et la prédication est une relation entre deux entités qui requiert un rapport type-instance (ou type-valeur) préalable. En élargissant le cadre, comme nous l'avons fait dans la partie précédente, on rend effectivement possible la prise en compte d'autres dimensions pour l'étude de l'information.
En définitive, dans le cadre des écritures du web, à partir de XML, et notamment le système d'interrogation SPARQL fondé sur une sémantique RDF, les processus sont relatifs à la prédication. Néanmoins, la prédication ne permet pas seule de répondre à la question posée de la dualité des phénomènes linguistiques et de connaissances qui se manifestent alors. Par exemple, pour l'exemple très simplifié des retards de trains, il est difficile de faire abstraction d'un schéma de connaissances qui explicite la situation et qui contraint donc la prédication.

A ces raisons externes, s'ajoutent bien évidemment les problèmes internes de la théorie proposée par Carnap et Bar-Hillel, présentés notamment par Bremer[100] et auparavant par Hintikka[101].

Conclusion.
On peut revenir sur les propositions de M.A.K. Halliday et clarifier certains aspects de l'information. La triple fonction de la clause (informationnelle, communicationnelle et

cognitive) permet de distinguer l'information de la compréhension comme de l'acquisition de connaissances. (La compréhension met en œuvre la dimension grammaticale du langage et l'acquisition la dimension cognitive).
L'information se définit par une fonction d'apport d'élément nouveau à l'intérieur d'un cadre compatible, sans que par ailleurs il y ait nécessairement une acquisition de connaissances.
L'information et l'analyse prédicative constituent un rapport entre problématique et outil d'analyse relativement productif. Les questions relatives aux réalisations du langage obligent à considérer la prédication tout au plus comme une méthode, y compris lorsque l'on considère l'information comme la fonction essentielle du langage (ce qui est le cas pour Z. Harris notamment). Il s'agit là d'une des difficultés de l'approche prédicative ; la prédication étant un concept et un outil d'analyse logique, elle ne peut être entièrement intégrée dans une seule approche (linguistique essentiellement).

On peut alors poser la question en sens inverse, à savoir que si l'information se caractérise par une prédication, est-ce que toute prédication peut se définir comme une information, ou plutôt, l'information ne serait pas une structure assemblant plusieurs prédications ?
D'un point de vue linguistique, comme on y reviendra, répondre positivement à la seconde question rend difficile la représentation de la structure d'information. C'est pour cela que la plupart des auteurs qui se sont penchés sur la question ont conclu à la prise en compte nécessaire des questions de connaissances.
Ce sera donc un pari que de vouloir caractériser la structure d'information de façon à ce qu'elle ait une précision compatible avec les exigences formelles du web sémantique. Or justement, les langages semi-structurés ont comme intérêt de simplifier les questions relatives au langage. Il en va de même pour nos exemples, qu'il s'agisse de l'annonce ou de la pharmacie. On élargit l'objet en prenant en compte la situation et on élimine une part de la complexité inhérente du langage. C'est ce que l'on tentera dans la partie 5 de ce travail.

Enfin, une question toujours ouverte de l'information est celle qui consiste à spécifier ce que l'on découvre, ou apprend dans l'information. Autrement dit, comment peut-on caractériser le fait qu'une information apporte un contenu nouveau (et cela dans un cadre théorique plus abstrait que les approches linguistiques, plus particulièrement d'analyse du discours où si l'on identifie ce qui est nouveau, on ne caractérise pas en quoi l'information donnée est nouvelle) à l'intérieur d'un ensemble de connaissances ?

3.2. Définition propositionnelle de l'information.

Les premiers travaux établissant une sémantique de l'information fondée sur des choix et des probabilités, notamment ceux de Carnap et Bar-Hillel, ont abouti au paradoxe de définir l'informativité maximale d'une expression par son impossibilité, ou son absurdité. (L'information la plus improbable étant la plus informative, alors l'informativité maximale est l'absurdité).
L'attitude qui consiste à considérer l'information par une proposition et certaines conditions de vérification trouve ainsi sa légitimité dans les paradoxes de la caractérisation prédicative de l'information. Ce sera l'angle choisi par la théorie des situations[102] et la théorie de l'information de Floridi[103]. Le choix (i.e. la prédication) réalisé est alors intégré dans un cadre propositionnel, ce qui permet alors d'instituer une dualité type/instance, et de contextualiser l'information dans le cadre de connaissances.

La définition propositionnelle de l'information décrit et explique la distinction entre l'information vraie et l'information fausse. Elle repose sur les principes suivant, proposés par Gordana Dodig- Crnkovic[104] :

- Association de l'information et la faculté d'adaptation : "In the beginning there was information. The word came later. The transition was achieved by the development of organisms with the capacity for selectively exploiting this information in order to survive and perpetuate their kind." On retrouve ici un fondement quasiment biologique à la caractérisation fonctionnelle de l'information.
- La précédente caractérisation fonctionnelle de l'information entraîne une définition intrinsèque du contenu de cette information : l'information que l'on considère est aléthique (soumise aux conditions de vérité), déclarative (positive), objective (représentant un phénomène du monde) et sémantique (dotée de sens).

Ainsi, l'information décrit un fait caractérisant une transformation dans un organisme. On rejoint ainsi le cadre dynamique dont nous nous sommes doté précédemment. Le paradoxe de l'informativité maximale et absurde sera résolu par la définition d'un cadre, celui dans lequel se réalise l'information. Cette solution reprend celle proposée au sein de la théorie des situations. Au sein d'une situation ou d'un cadre, défini par une proposition vraie, on intègre une proposition informationnelle contrainte par ce cadre. Les propositions seront limitées par le cadre, ce qui permet d'éviter les absurdités.

Par ailleurs, la caractérisation propositionnelle de l'information intègre la distinction entre valeur et type, et donc a un rôle essentiel pour associer les outils d'expression du web et les modèles de la linguistique et de la psychologie[193].

Cadre général pour l'étude de l'information en tant que proposition.
Gordana Dodig-Crnkovic (op. cit.) oppose une théorie de l'information fondée sur des données signifiantes à celle fondée sur des données associées à une valeur de vérité.
Ce changement vise à répondre à la critique de la vérité associée à l'information. Comment distinguer avec les probabilités (et l'information comme donnée signifiante) entre une information vraie et une information fausse ?
Cette vérité est définie dans un univers (ce qui permet de faire exister les licornes). Un univers est considéré comme approprié dans un certain contexte ou situation.
En ce sens, le modèle de l'information est contextuel. C'est dans un certain contexte que l'information se vérifie ou pas. Donc, on préférera « adéquation » à « vérité ». (Cette formulation des valeurs de vérité reprend les logiques modales).
Ces acceptions permettent de caractériser l'information en fonction d'un degré de pertinence dans l'univers choisi. C'est en ce sens que l'on parle d'apprentissage et d'enrichissement dans l'univers.

Le cadre dynamique a une autre conséquence ; il sous-tend le fait que l'information caractérise le monde lui-même ; voire même constitue un cadre de travail pour identifier et décrire les réalités primaires de ce que l'on comprend. Il existerait ainsi une dimension réaliste dans l'information.

[193] On pourra noter qu'une logique prédicative permet de résoudre le problème en présentant des contraintes distributives entre les prédicats et leurs arguments. La distribution permet de produire des classes de mots, qui définissent les possibilités d'emploi des mots dans un contexte défini. Voir les travaux de Maurice Gross et de son équipe.

C'est pour cela que les questions de sémiotiques sont fondamentales lorsqu'il s'agit de caractériser une ontologie de domaine. Une ontologie ne peut être alimentée qu'à partir d'informations relatives à des phénomènes du monde (Barry Smith[105]).

Ainsi, les systèmes d'information (entendus au sens le plus large, comme par exemple le système d'information de l'adaptation de posologie) ne constituent pas seulement des outils techniques, mais bien des simulations, représentations et prédictions des activités du monde réel. Par là même, ils orientent le monde réel, à savoir qu'ils modifient des pratiques.
(De cette façon, il y a une forte imbrication entre les modèles de systèmes d'information et l'évolution des pratiques).

Propositions de L. Floridi.
Nous avions déjà présenté un travail de L. Floridi, relatif aux niveaux d'abstraction et à la façon par laquelle il était possible de transmettre des contenus depuis un niveau vers un autre. Cette première problématique débouche chez L. Floridi sur la caractérisation de ces contenus, et donc la question de l'information. Avant de la présenter, nous voudrions la mettre en contexte.

Limites et perspectives.
On va s'intéresser d'abord à la façon dont les propositions peuvent être caractérisées dans le cadre de l'information, et non plus celui des expressions linguistiques seules. Bremer (op.cit.) distingue alors une solution « syntaxique » et une solution « ontologique » pour spécifier les stratégies possibles pour caractériser l'information.

3.2.1. Caractérisation « syntaxique ».

La solution « syntaxique » consiste à associer l'information à une structure de l'expression. Cette structure peut être caractérisée en termes sémantiques, mais également psychologiques. Une expansion de cette caractérisation est la proposition d'une définition algorithmique du contenu informationnel proposée par Gregory Chaitin[106] : le contenu informationnel d'une formule est la longueur du plus court programme (en bits) qui est requis pour générer la formule.

Pour G. Chaitin, la question de la vitesse n'est plus essentielle, à la différence de celle de la diminution des grandeurs. La compression des données est fondée sur la redondance. (Vous pouvez éliminer une unité de deux parce que l'une d'elles peut être portée par l'autre sans perte notable d'information). Or, justement, certains faits sont irréductibles parce qu'ils sont trop complexes et donc, qu'ils peuvent être décomposés à l'infini. Par ailleurs, s'ils ne peuvent être réduits, c'est qu'ils n'obéissent pas une loi régulière, comme une loi mathématique. Donc, ils sont irréguliers.
Vu ainsi, toute théorie est une compression de données, la compréhension est aussi une compression de données. Par contre, l'information étant incompressible, elle est irrégulière.

Inversement, on peut définir la complexité comme le contenu informationnel d'un ensemble de faits et donc la grandeur du plus petit programme permettant de les calculer. Alors, une suite d'opérations est irréductible lorsque sa complexité est égale à sa grandeur (à savoir qu'il n'y a pas de programme permettant de la calculer de façon plus petite à celle qui existe).
En d'autres termes, il n'y a pas de redondance dans le programme, il peut donc être transmis directement. C'est par exemple, ce que fait chaque DVD en lisant les instructions permettant de restituer les images, et non les images elles-mêmes.

Ainsi, on a les bases d'une théorie de l'information par laquelle il s'agirait de définir l'information par son caractère irréductible. Le calcul de ces phénomènes irréductibles constitue la base de la théorie mathématique de l'information, donc constitue le calcul de ce qui est complexe. L'information représente les phénomènes ne pouvant être compressés.

L'intérêt d'une telle proposition réside dans le fait que l'on peut représenter de façon binaire des phénomènes irréductibles (comme par exemple les images). Cette proposition pose en fait un problème par rapport au rationalisme, pour lequel tout ce qui existe est explicable par une cause.
En effet, par le biais de l'information algorithmique, on représente de façon mathématique des objets qui ne sont pas expliqués par une causalité. (La théorie de la causalité propose de caractériser la connaissance (et donc la démonstration mathématique) comme l'explication causale d'états. C'est ce qui permet de distinguer les connaissances des croyances).

Donc, en suivant le raisonnement, on peut s'interroger sur la nature de la connaissance et sur sa dimension causale. On aurait ainsi des phénomènes qui sont vrais (au sens mathématique), tout en étant vrais sans raison. (En fait, cela amène à considérer qu'il existerait une infinité de faits mathématiques, tout simplement par l'application des dualités 1/0 (vrai/faux, oui/non).

Ces faits sont non seulement irréductibles d'un point de vue des calculs, ils le sont également d'un point de vue logique. (Logiquement, ils doivent être considérés comme des axiomes, à savoir des propositions qui n'ont pas à être démontrées. Les théorèmes sont les propositions que l'on peut inférer à partir des axiomes et que l'on peut prouver).

G. Chaitin fonde son raisonnement sur la nature des nombres réels[107]. « What is a real number? It's just a measurement made with infinite precision. So if I have a straight line one unit long, and I want to find out where a point is, that corresponds to a real number. If it is all the way to the left in this unit interval, it's 0.0000. . . If the point is all the way to the right, it's 1.0000. . . If it is exactly in the middle, that's .50000. . . And every point on this line corresponds to precisely one real number. There are no gaps.

0.0 —— 0.5 —— 1.0

So, if you just tell me **exactly** where a point is, that's a real number. From the point of view of geometrical intuition a real number is something very simple: it is just a point on a line. But from an arithmetical point of view, if you want to calculate its numerical value digit by digit or bit by bit if you're using binary, it turns out that real numbers are **problematical**.
Even though to geometrical intuition points are the most natural and elementary thing you can imagine, if you want to actually calculate the value of a real number with infinite precision, you can get into big trouble. Actually, you never calculate it with infinite precision. What Turing says is that you calculate it with **arbitrary** precision.
His notion of a computable real number is a real number that you can calculate as accurately as you may wish."

Si 2^n correspond au nombre de fois qu'une opération est réalisée, à ce moment-là, on ne peut caractériser oméga Ω parce que la clôture est fixée de façon arbitraire. Par contre, on peut raisonner sur le collectif, à savoir les probabilités de clôture d'un programme.

Très concrètement, n'importe quel programme s'arrête parce qu'il termine une fonction externe pour laquelle il a été défini. Donc, en soi, le programme ne se termine pas pour des

raisons internes, justifiées par les calculs. (C'est le problème soulevé par Turing, pour lequel tout algorithme est clôt par un acte arbitraire, donc non mathématique).

Ω se caractérise collectivement par la probabilité qu'une machine fasse quelque chose. Ainsi, un programme de N-bits qui s'arrête contribue précisément à $1/2^n$ à la somme qui caractérise oméga Ω. (Autrement, dit les probabilités servent à représenter et expliquer l'infini).

L'idée consiste à passer par la somme des opérations ; elle clôt l'algorithme à $1/2^n$. Si chaque bit d'un programme est choisi en utilisant une face d'une pièce de monnaie, (1/0), alors la probabilité d'avoir n'importe quel programme à N-bit est $1/2^n$. Chaque programme à N-bit qui s'arrête ajoute un 1 au Nième bit dans l'expansion binaire de Ω.

Par conséquent, on n'a pas de calcul de Ω. Néanmoins, Ω est calculable si chaque programme obéit à la règle de l'auto-délimitation : en effet, aucune extension d'un programme ne peut être un autre programme.
Pourquoi ? En introduisant de la réitération continue (principe de récursivité illimitée), on élimine toute possibilité de calcul puisque les données ne sont pas limitées (autrement dit, que l'on n'a pas un programme fini, donc calculable).
Cela garantit que la somme de tous les programmes qui s'achèvent soit une valeur entre 0 et un chiffre qui ne soit pas l'infini.

Le calcul de Ω collectif fait disparaître toutes les redondances entre les programmes (c'est une différence avec les estimations individuelles).
Donc Ω est irréductible et incompressible. Le problème est qu'il ne s'explique pas (puisqu'on ne peut pas le prédire : par conséquent, il ne peut constituer un fait mathématique).

G. Chaitin déduit une complexité infinie à partir des phénomènes mathématiques. Pour lui, les mathématiques sont quasiment empiriques.
(Généralement, dans les sciences, on compresse des données expérimentales dans des lois, éventuellement formalisées. En mathématique, on compresse des expérimentations de calcul dans des axiomes mathématiques).
C'est pour lui l'apport essentiel de la computation -et de l'informatique comme terrain d'application- dans l'ensemble des sciences.

Parallèlement, on a donc la possibilité d'une représentation de la richesse des phénomènes, à la différence des simplifications obtenues par d'autres méthodes.

Dans le cadre des propositions de G.Chaitin, par définition l'information est infinie.
On distingue l'information comme réalisation du contenu informationnel, qui, lui, est la réalisation compte tenu de l'ensemble des choix qui ont été formulés. Cette solution privilégie la réalisation sur le contenu.

3.2.2. Caractérisation « ontologique ».

La solution ontologique consiste en l'analyse des référents des constituants de l'information pour ensuite aborder la proposition complète. C'est une perspective compositionnelle mais dont l'objectif consiste à mettre en évidence des régularités de composition d'expressions ; ces régularités d'association de propositions ayant un mode de validation spécifique forment une structure, qui possède donc des propriétés distinctes de d'association des différentes propositions composantes. Cette solution initiée par J. Barwise & J.Perry dans la théorie des

situations consiste à adopter une écriture de la proposition à grain beaucoup plus fin que la définition usuelle : chaque constituant de l'expression sera alors considéré comme une proposition. Cette écriture est celle de « l'infon » (néologisme proposé par K. Devlin[108] désignant la structure propositionnelle minimale permettant de transmettre une information), qui caractérise la réécriture informationnelle de la proposition. Cette solution est dite ontologique parce que l'analyse de la proposition requiert préalablement celle de ses constituants, qui sont des entités propositionnelles ayant une existence propre.

Ainsi, les questions relatives à la dimension informationnelle de l'expression sont contraintes par la réalisation de chacun des composants de cette information. Ces composants peuvent être caractérisés selon certains types ; ces types propositionnels sont alors considérés comme des contenus. On retrouve ici les principes de l'information comme classification, exprimés par Dretske[109] et de cette classification comme signification.

Théorie forte de l'information : les propositions de L. Floridi.
Les théories de l'information fondées sur les valeurs de vérité sont faibles parce qu'elles ne prennent pas en compte les instances d'information EN DEHORS de leurs valeurs de vérité. Dans une sémantique forte, par contre, la vérité est encapsulée dans l'information, c'est-à-dire que l'information contient d'autres éléments que la seule vérité.

L'information est alors associée à un signal et à l'élimination des possibilités associées à ce signal. L'information n'est pas seulement une question de probabilités, mais de choix entre des valeurs aléthiques et contradictoires. (Les modalités aléthiques caractérisent des valeurs existant en soi, indépendamment des univers dans lesquelles elles se réalisent : +/-, 0/1).

On pourra caractériser le contenu informationnel. On considère alors un « infon » σ qui désigne un item discret d'information (sans prendre en compte sa dimension sémiotique ni son implémentation physique).Le contenu de cet « infon » est considéré négativement avec l'ensemble des états possibles de l'univers exclu par σ.
$CONT(\sigma)$ = déf.
Déf. Désigne l'ensemble de toutes les descriptions d'état inconsistantes avec σ.
Vu ainsi, l'item d'information σ contient l'ensemble des univers impliqués dans la production, la circulation et la représentation de l'information σ. Il s'agit donc d'une perspective extensionnelle, à savoir que l'information est systématiquement associée à une dimension du monde.

De cette façon, une expression quelconque réalise un nombre de messages extrêmement important.
Dans un univers quelconque E, chaque message associé à σ dénote 1 des *n* messages possibles.

- Dès lors qu'un message σ_i sera réalisé (dans l'ensemble $\sum$ de tous les messages possibles et s'excluant mutuellement et étant jointifs), toute conjonction à un autre message sera inconsistante.
- Les tautologies (un chat est un chat) véhiculent le minimum d'information.
- Les contenus contradictoires véhiculent le maximum d'information.

De cette façon également, le degré d'informativité est directement proportionnel au contenu sémantique de σ .
Ainsi, L. Floridi propose de compléter la formule définissant le contenu de l'information :
$CONT(\sigma) = 1 - p(\sigma)$ (où p = probabilité)

Pour (ι : informativité et ∝ : proportionnalité)
$\iota(\sigma) \propto CONT(\sigma)$

Ainsi les théories fortes du contenu informationnel prennent en compte la valeur aléthique de chaque instance d'information (en dehors de sa vérité ou de sa fausseté). Ces théories fortes de l'information permettent de mesurer les qualités d'information transmise. Elles permettent enfin de caractériser comment la dimension pragmatique s'articule à la sémantique de l'information.

Deux infons σ_n et σ_m sont considérés comme co-informatifs (ils ont une quantité équivalente d'information), Ci, s'ils ont une signification équivalente :
C_i (σ_n et σ_m) $\leftrightarrow$ σ_n et σ_m

On peut prendre comme exemple : « Pierre conduit la voiture », « la voiture est conduite par Pierre ». Ces deux énoncés sont qualitativement co-informatifs.

L'analyse pragmatique permet de décider quel énoncé sera plus informatif, mais après l'accomplissement des valeurs de vérité de chacun.
C_i (σ_n et σ_m) $\leftrightarrow$ σ_n et σ_m, signifie alors qu'ils sont équivalents en termes de valeurs de vérité.

Enfin, on déduit des précédents l'équiprobabilité des deux infons équivalents.

Les trois équivalences précédentes permettent d'utiliser le concept de sémantique forte de l'information dans des contextes différents :
L'équivalence de signification permet d'orienter l'analyse vers le traitement de la signification, en introduisant la dimension contextuelle du sens.
L'équiprobabilité permet au contraire une analyse des probabilités sans respect pour l'interprétation actuelle et la valeur contextuelle.

En définitive, l'informativité intrinsèque de σ est caractérisée en extension et dans un contexte idéal, en fonction du degré positif ou négatif de « distance sémantique », ou dérivation de σ par rapport à un point fixe, ou origine, représenté par une situation donnée *w*, à laquelle σ est sensé référer. La situation est définie comme une région structurée de l'espace-temps et le contexte comme un ensemble de conditions reliées dans laquelle la situation apparaît. L'informativité est ainsi caractérisée comme le calcul de la distance entre d'une part l'infon, et d'autre part la situation.
Si l'on caractérise δ comme la valeur de la distance,
$\delta = f(\sigma)$.

La position de L. Floridi est alors relativement explicite : en considérant que l'information constitue un objet déclaratif, objectif et signifiant, l'acceptation d'une information constitue alors la modification d'un état des affaires (autrement dit de ce qui est couramment pensé comme étant le monde lui-même). On reviendra plus loin sur la notion d'état des affaires ; néanmoins, on peut considérer que le contenu informationnel est l'ensemble d'état des affaires qu'une phrase actualise (à savoir renouvelle et transforme), et l'ensemble des incertitudes que cette phrase actualise. Plus précisément, L. Floridi réintroduit les conditions de vérité, mais il les considère comme encapsulées dans l'expression, à savoir relatives à la dimension mentale de l'interprétation. La validation ne se fait pas dans le monde, mais par rapport à un état des affaires, qui constitue la façon dont on considère le monde.

Mais pour bien comprendre cela, il faut revenir à une évolution dans la théorie de l'information. En effet, au départ, l'information est d'abord considérée comme une ensemble de données structurées et ayant une certaine signification. Alors, l'information a une dimension prédicative et concerne fondamentalement la dimension linguistique. Or, à partir notamment de la théorie des situations, l'information est définie par une proposition, et se caractérise par une représentation, à savoir une construction mentale. Dans le cadre de la théorie des situations, cette représentation se valide dans le monde, notamment au travers du traitement des indexicaux (déictiques). En ce sens, la théorie des situations est une théorie sémantique, et permet le traitement de certaines unités du langage.
Par contre, pour L. floridi, c'est la représentation qui est elle-même vérifiée. En ce sens, les dimensions extensionnelles disparaissent totalement dans la définition du statut de vérité. « Si ceci est vérifiable, alors c'est informationnel ». (La proposition doit être acceptée dans la théorie si on veut qu'elle soit considérée comme informationnelle).

L'intérêt de la démarche de L. Floridi est de construire un objet permettant d'appréhender les connaissances et le langage de façon relativement neutre. La perspective de L. Floridi permet d'accroître considérablement le nombre d'objets et de relations que l'on peut assembler sous le concept d'information. En même temps, il ne s'agit que de propositions théoriques, et il est difficile de savoir exactement quel objet concret peut être construit et analysé en utilisant ces outils.

Un certain nombre de critiques, en Sciences de l'Information (Bawden, D. *Organised complexity, meaning and understanding: an approach to a unified view of information for information science,* [110] p.317), ont été formulées à l'égard des propositions de L. Floridi, et plus généralement de la perspective propositionnelle de définition de l'information. Elles concernent trois implications des propositions de L. Floridi :

- La dimension collective de la connaissance ; toutes les structures d'organisation des connaissances sont socialisées, de même donc que les connaissances qu'elles se proposent de transmettre.
- La notion même de « vérité ». D. Bawden reprend le concept de K. Popper d'une connaissance nécessairement imparfaite, et donc par conséquent, ne permettant pas d'appliquer les postulats de vérité, conçus eux pour l'interprétation des expressions en langage naturel.
- Enfin, le lien entre vérité et expérience.

Contraste : caractérisation matérielle de l'information.
Les propositions de M. Burgin[111] prennent l'exact contrepied de celles de L. Floridi puisqu'elles reposent sur le principe que l'information est un phénomène matériel.
Pour lui, c'est une certaine organisation des données qui peut être appelée information. On considère alors l'information au travers de la structure d'information.
Ce point de vue amène une autre question, celle de la représentation de l'information. En effet, s'il est possible de caractériser l'information à l'intérieur d'un cadre biologique ou physique, c'est toujours du point de vue humain. Du point de vue de la molécule, est-ce que l'information existe ? La question de la représentation de l'information comme constituant de l'information, renvoie systématiquement à la question des outils de symbolisation. Elle permet aussi de distinguer entre donnée et information. (M. Burgin, op. cit., p.5).

A la façon de L. Floridi, M. Burgin fonde sa théorie de l'Information sur les questions de vérité. Cette question est liée à l'informativité et par conséquent à la notion même

d'information. Simplement, la vérité est définie par la capacité de l'information à modifier à modifier l'état du système.

En présupposé, M. Burgin considère que l'information est le fondement d'un système et qu'elle n'existe qu'à l'intérieur de ce système. Ainsi, il n'y a pas de producteur à cette information, mais systématiquement un transfert, ce qui permet de se libérer du problème posé par l'énonciateuret donc l'intentionnalité. L'énonciateur est antérieur au transfert.

L'information se caractérise par un changement à l'intérieur du système et renvoie à la mesure de la quantité d'information de Hentley reprise par Shannon.

Le système est défini par une mémoire et contient donc des états. Ceux-ci mémorisent les résultats des transformations opérées par l'information. Les informations ne sont pas enregistrées dans le système. La mémoire est celle des changements opérés par l'information. M. Burgin compare le rapport entre donnée et information à celui entre matière et énergie. Cette comparaison s'applique aussi au rapport entre connaissance et information.

Une autre implication des propositions précédentes, et notamment celle concernant le transfert, est manifeste dans la définition de l'information elle-même. Empruntant à Roedered, elle est définie par « *the agent that mediates the correspondence between features or patterns in the source system A and changes in the structure of the recipient B* » . (M. Burgin, op. cit., P 16). Cette définition permet de rendre compte du fait que l'information peut s'intégrer dans des ensembles de connaissances très différents et produire des effets imprévisibles. Elle peut ne pas voir d'effets du tout à partir du moment où les connaissances préalables du récepteur ne permettent pas d'identifier l'information.

Le dernier principe ontologique de M. Burgin rend compte des flux. Une information étant transmise, elle est nécessairement représentée. Trois classes caractérisent cette information : la dimension matérielle (comme le document), la dimension structurelle (comme le texte), et enfin la dimension mentale. L'information se caractérise au travers de ces trois niveaux. Il s'ensuit que l'information est nécessairement contenue dans une substance.

Comme dans la théorie l'information est caractérisée par un changement dans le système, on peut donc associer une mesure à ce changement. Cela permet de comprendre ce qu'est une information fausse, nulle ou vraie pour le système. Cette mesure amène in fine à caractériser le modèle de la connaissance (considérant le système comme un modèle de connaissances). Classiquement, le modèle de la connaissance de M. Burgin est fondé sur une classification où une entité F est fonctionnellement liée à sa représentation « F » , chacune étant classée par une propriété dans une échelle W pour F et dans un langage L en vertu d'une propriété permettant sa conceptualisation.

Les propositions de M. Burgin permettent de fournir un cadre à la mesure de leur généralité. La première difficulté réside dans le lien entre activité et état. En quoi la dynamique est-elle seulement liée à l'information, doit-on considérer le cadre dans lequel elle s'inscrit comme un seul système ? Nous caractériserons l'information dans une activité professionnelle afin de clarifier ce point et surtout de donner une réponse plus nuancée parce que l'univers dans lequel l'information est inscrite n'est pas statique, mais bien structuré par des tâches insérées dans une activité. A la suite de J. Barwise & J. Seligman, nous distinguerons la succession des

états d'un système ou d'une activité des flux d'information, considérant que c'est l'interprétation d'une information qui produit le changement d'état.

3.2.3. Pragmatique, actualisation et révision.

D'autres solutions encore sont recherchées, notamment par les logiques épistémiques. Ces propositions ont été reprises partiellement dans les travaux de J. Van Benthem[112], qui essaient, grâce aux logiques modales, de caractériser l'information dans le cadre d'interaction (et notamment, de paires question/réponse).

Le fait qu'une information soit produite et circule requiert des opérations inscrites dans un système. Dans un cadre pragmatique, cette proposition est incomplète si l'on ignore la dimension intentionnelle. En effet, l'information se définit, à l'intérieur du système, comme ayant une visée sur les croyances du récepteur.

Néanmoins, comment peut-on caractériser l'intention relative à l'information ? F. Récanati[113]spécifie que dans le cadre de l'information, les intentions sont transparentes et limitées à celles de production et d'actualisation d'une représentation structurée d'un univers référentiel en vue de rendre possible une utilisation de celui-ci.
En d'autres termes, F. Récanati intègre ici pleinement la dimension pragmatique tout en maintenant la possibilité d'une dimension sémantique de l'information. Il reprend ici les principes fondateurs de la théorie des situations, que nous avons déjà mentionnés dans les propositions de L. Floridi. La sémantique de l'expression transmise est la conséquence de l'intention. Cette intention caractérise la façon dont l'univers est segmenté et les relations établies entre ces segmentations. L'univers est considéré comme étant lui-même dynamique, indépendamment des actions réalisées au travers du système d'information, mais il est structuré par l'intention d'information.
L'univers serait segmenté et structuré relativement à la possibilité et la volonté d'informer. Par exemple, le corps du patient serait symbolisé et partitionné de telle façon que cela permette d'informer ; les parties du corps qui existent dans le cadre de l'adaptation de posologie le sont parce qu'elles permettent de l'information à leur propos[194].
Le cadre du flux d'information constitue la dimension pragmatique, au sens de la réalisation pratique, matérielle de l'intention, mais en même temps, puisque cette intention serait relative à une structuration du monde, les flux se caractérisent fondamentalement par une dimension intensionnelle. C'est pour cela aussi que les flux sont essentiels dans le cadre des ontologies. Ils caractérisent un raisonnement à un haut niveau d'abstraction tout en transmettant une représentation d'un état du monde, de bas niveau d'abstraction.

Cette proposition n'est pas sans lien à la construction de représentations textuelles. L'addition d'information se traduit non par une augmentation du nombre de représentations, mais par une précision accrue des représentations antérieures. En effet, l'actualisation est considérée comme une élimination de la pluralité des interprétations possibles. Autrement dit, on caractérise ici le lien pouvant apparaître entre l'addition d'information et la compression de ces informations dans l'élimination d'incertitudes[114]. Ainsi, l'actualisation d'informations s'inscrit dans une procédure de révision de croyances.

---

194 Cette proposition ne serait excessive que si l'on ignorerait l'ensemble des connaissances des pharmaciens. Il est bien évident que leur regard professionnel entraîne des interprétations impliquant des objets du corps du patient qui ne sont pas pris en compte par l'information. L'information que l'on considère ne concerne que l'état et l'évolution du patient.

Les procédures d'actualisation concernent les représentations mentales et leur évolution alors que les implications pragmatiques de ces révisions dans le cadre d'actions sont analysées comme des révisions de stratégie dans la menée d'une activité.
L'actualisation est alors considérée comme un processus limité qui s'inscrit dans des cadres plus larges, dont le premier d'entre eux est celui des jeux. La théorie des jeux permet de représenter des processus très divers, comme la décision et la conversation, qui mobilisent des stratégies à long terme. A ce moment-là, on n'est plus dans le cadre des flux et de l'information, mais de l'activité. On parlera alors non plus d'information mais de traitement de l'information.

Or, ces stratégies constituent des processus finalisés, donc limités à la réussite d'une tâche dans une activité. Ce n'est pas le cas des flux. En effet, les flux sont limités en termes de types d'objet traités dans un certain contexte mais ils ne le sont pas dans la durée. Les sources d'information sont continues.

En somme, les propositions de J. Van Benthem rejoignent, voire même s'inspirent, des propositions de cognition située et distribuée, y compris dans les références aux dimensions évolutives des systèmes naturels. Enfin, il reste un certain nombre d'incertitude dans ces propositions, notamment dans la distinction du cadre dynamique par rapport aux dimensions pragmatiques.

Néanmoins, toutes ces propositions considèrent l'information au travers d'une structure expressive (quelle que soit sa nature). Or, la structure expressive ne résume pas la question de l'information : elle ne peut rendre compte des flux.
Ces propositions permettent de faire quelques distinctions entre information et communication. En effet, à partir du moment où une information potentielle serait inscrite dans une stratégie de communication, elle perdrait sa neutralité génétique, à savoir ce qui fait sa distinction avec des connaissances, des croyances des états mentaux transmis.

Sur quelques traits fondateurs pour une définition de l'information.
S'il y a bien un point sur lequel L. Floridi et ses détracteurs sont d'accord, c'est sur la neutralité génétique de l'information. Cette définition génétique de l'information peut être reprise de M. Burgin (op. cit., p. 5), entre autre parce qu'elle explicite la distinction avec les principes communicationnels : "At the same time, if we assume that some kind of meaningful data is information, then we have to admit that Menant is right, writing that the principle of genetic neutrality (GN) "supports the possibility of information without an informed subject, to adapt a Popperian phrase. Meaning is not (at least, not only) in the mind of the user.""

Nous avons ici présenté un certain nombre de théories de l'information qui ont en commun les principes de départ que nous avons énoncé. Nous les avons choisies pour leur généralité, à savoir le fait qu'elles ne s'inscrivent pas dans un domaine restreint par rapport au champ couvert par le web de données et les projets que l'on peut y élaborer. Le corrélat à une telle généralité est une certaine difficulté à en trouver un domaine d'application. Mais ce n'était pas là le but de notre présentation. Ce cadre général permet autant d'aborder les flux que les structures d'information d'une part que les problématiques associées aux outils du web de données (lexiques, ontologies, raisonnements). Toute caractérisation plus restreinte nous aurait fait encourir le risque d'une inadéquation par rapport au contexte.

3.3. Circulation d'informations : classifications et flux. Définition de l'information comme contraintes sur les discours.

Les flux s'intéressent aux conditions et contraintes permettant la circulation de l'information. Comment et pourquoi une information circule ? Cette question posée par les flux a donné lieu à des travaux antérieurs aux propositions de Barwise & Seligman, et qui caractérisent les classifications (et notamment le rapport instance/type) dans leurs aspects à la fois sémiotiques et cognitifs.

Avant donc d'aborder la théorie des flux en tant modèle ayant certaines propriétés pour la représentation d'inférences, il apparaît important d'aborder les fondements et les inspirations de cette théorie et les ruptures qu'elle propose par rapport aux approches précédentes.

Nous envisageons les ruptures au travers de deux questions :

- Comment caractériser les structures de données qui constituent les ressources pour l'information et comment spécifier leurs relations de façon à mettre en valeur l'hétérogénéité des composants ?
- Comment et pourquoi une entité appartenant à l'une des structures peut-elle porter une information vers une entité d'une autre ?

Enfin, et il s'agit peut-être de l'apport essentiel de la théorie des flux à la question de l'information, on reprendra l'hypothèse de J. Jayez & A. Mari (*Togetherness*, op. cit) : les flux représentent des relations causales entre des ensembles d'entités hétérogènes mais liées par le fait qu'intrinsèquement, le comportement de l'un est la cause du comportement de l'autre, à la fois en extension et dans le cadre de connaissances (en intension). Si un état β est extensionnellement causé par α, on sait que α entraîne β.
Au travers des flux, la question de l'information est directement liée à celle de l'action. Nous essaierons de montrer la validité de cette hypothèse, qui par ailleurs nous permet de passer d'un cadre d'unités symboliques vers celui d'activités dans le monde.

3.3.1. Flux, lexique et hétérogénéité des composants.

Nous aimerions avancer ici quelques arguments qui permettent de distinguer notre perspective de celle qui est généralement adoptée en matière d'analyse prédicative des expressions symboliques.

Dans une théorie classique de la prédication, comme celle de Z. Harris, les choix opèrent entre des entités opérateurs prédicatifs et des « opérandes », à savoir des classes de termes acceptées par cet opérateur. On postule ainsi une familiarité de domaine entre les opérateurs et les opérandes. L'analyse prédicative de Z. Harris, reprise ensuite par G. Gross, se fonde sur l'acceptation des arguments par le prédicat. Ce principe d'acceptabilité permet de construire des classes de mots. Il se fonde sur la détection de traits de signification communs permettant à la fois de construire les classes et de lier les classes entre elles dans le cadre de l'expression. Cette conception est fondée sur de marqueurs d'homogénéité entre des structures lexicales différentes.
Par contre, dans une théorie comme celle des flux, il existe une mise en relation des entités qui s'exerce entre des classes d'entités hétérogènes. Les prédicats sont considérés comme des unités elles-mêmes typées, ce qui a comme conséquence un traitement des expressions à un niveau plus abstrait que l'analyse prédicative.

Ainsi, les propriétés et les mesures : y a-t-il une communauté de domaine entre les concentrations d'une molécule dans le rein et une échelle de mesure (de 0 à 30 par exemple) ? Aucune si l'on observe la référence. D'un côté on a une propriété biologique et physiologique, d'un autre une échelle insérée dans un appareil de mesure. Par contre, si l'on apprend que ces concentrations s'expriment sur une échelle grâce à la médiation d'un appareil de mesure (ou d'un calcul), alors on établit un lien entre ces deux ensembles de symboles. Ce lien est marqué par une contrainte, à savoir que si l'on veut renseigner la propriété, il est nécessaire de passer par l'expression d'une instance sur une échelle.

Si l'on regarde plus précisément, on s'aperçoit que l'acception de la prédication dans la théorie de Z. Harris concerne la production d'un message dans un cadre qui n'est pas nécessairement structuré par un flux régulier (un système d'information ou un réseau par exemple), mais une production discursive qui peut ne pas répondre à des modèles normatifs de production. Dans notre cas, on a au contraire affaire à un processus régulier opérant sur des ensembles de données structurées. Par exemple, les propriétés s'expriment systématiquement par le biais d'une échelle de mesure. C'est la raison pour laquelle on a besoin, pour caractériser les flux, de les considérer dans le cadre de structures. Les classifications serviront à cela.

Les flux ne partent pas d'une expression, mais de structures (lexicales et conceptuelles) qui formeraient les expressions. Dans le cadre de la théorie des situations[195], ils rendent compte de la façon dont chacun des composants de la structure contraint et renseigne le suivant. Les flux possèdent donc une dimension prédicative, mais à un niveau d'abstraction qui correspond à des propositions déjà analysées. En effet, les flux caractérisent les structures linguistiques typées. En ce sens, il ne s'agit pas d'un concept destiné à rendre compte de la sémantique des unités du langage naturel mais de la formation de structures. C'est ce que nous expliquons maintenant.

Ils caractérisent des régularités entre des structures (lexicales, de connaissances) hétérogènes, et expliquent en quoi ces régularités constituent l'information : la première structure exerce des contraintes de choix sur la seconde, qui elle-même requiert la première pour signifier. De telles structures sont caractérisées comme des formes prédicatives, mais ne prennent pas nécessairement comme fondement la phrase. En d'autres termes, la prédication est considérée entre deux types d'entités qui sont analysées comme étant distinctes et autonomes dans leur signification (voir l'analyse propositionnelle antérieure relative à « l'infon »).

3.3.2. Classification et hétérogénéité.

La première modélisation des flux a été formulée par F. Dretske[115], en montrant l'importance des classifications dans la circulation de l'information. Le principe des flux, c'est le fait que l'on établit des liens prédicatifs (accompagnés de probabilités au niveau des instances lexicales) entre au moins deux structures de données hétérogènes : ces liens sont des contraintes mutuelles entre les deux structures (au niveau des instances inversement à celui

[195] Au départ, la théorie des flux visait simplement à apporter quelques compléments à la théorie des situations. Il s'agissait de développer le concept de contrainte d'une situation sur une autre dans la composition d'une structure d'information. Ce point était essentiel dans la théorie puisqu'il s'agissait de rendre compte d'une structure à partir des éléments la composant. Les contraintes servent à exprimer le fait que l'on ait bien une structure faite d'un assemblage contraint de propositions à grain fin.

Barwise et Seligman ont perçu l'originalité de leur proposition en prédisant un cadre d'application qui serait celui des systèmes distribués.

des types), de telle sorte que les choix informationnels réalisées dans chacune des structures évitent le paradoxe de la prédication non contrainte (et donc l'information absurde).

Ainsi, on saisit l'articulation entre classification et flux : les classifications sont des conditions pour qu'une information puisse circuler.

La théorie fonctionnelle et symbolique de F. Dretske considère l'information à partir des connaissances et de leur transmission et tente de répondre à la question suivante : comment, dans quel cadre, je peux transmettre une information et m'assurer de sa compréhension ?

Le point de départ de F. Dretske est triple et associe la dimension des télécommunications à celle de la signification :

- quel contenu transmet un système d'information tel que considéré par Shannon (à savoir un modèle de télécommunications) ?
- quel signe (et plus largement quel système symbolique) peut être associé à l'information circulant dans le système ?
- Shannon présente combien d'information un signal peut transmettre, mais pas quelle information : qu'est-ce qui donc fait information dans un signal, et comment cela peut être calculé ?

Théorie du signe.
Au départ, il faut définir ce que F. Dretske entend par signe ; pour cela, il reprend les propositions de Paul Grice[116]. P. Grice considère deux types de signes : les **signes conventionnels** sont ceux où le rapport signifié-signifiant-référent fait l'objet d'une négociation et d'une convention entre les membres de la communauté utilisant ces signes. C'est le cas du langage naturel tel qu'il est utilisé dans la conversation quotidienne. C'est aussi l'objet de la pragmatique, telle qu'elle a été initiée par P. Grice.
Inversement, les **signes naturels** sont ceux où le rapport signifié-signifiant-référent est fixé par des lois contraignantes, comme le fait que la fumée indique le feu, le roucoulement la colombe, etc.).

Le signe de l'information que transmet un canal ne peut être considéré comme conventionnel (puisque ce dernier serait relatif à un consensus social et donc ne permettrait pas d'apprendre) : il doit être de même régularité que le canal du système et transmettre un état ou un fait nouveau, inédit. Par conséquent, il sera considéré comme naturel (dans l'acception de P. Grice). Il est appelé signe naturel en fonction de son lien avec la causalité naturelle (qui constituerait la façon dont la connaissance humaine appréhende les objets naturels). La causalité est valable également pour les phénomènes liés à des productions techniques : par exemple : " le flash signifie que la photo a été prise ". Qu'est-ce qui fait que l'on puisse déduire la " photo " du " flash " : tout simplement une causalité : " le flash s'explique par la prise de la photo ". La relation est systématique : elle est supportée par une loi fonctionnelle.
Nous reviendrons plus loin sur cette question de causalité, notamment lorsqu'il s'agira de caractériser le raisonnement mis en œuvre dans les flux.

Le signe naturel n'est pas en premier lieu linguistique ; il peut être considéré hors de la symbolisation (« la fumée signifie le feu », « l'aboiement signifie le chien »). En outre, il est marqué par une dimension classificatoire qui est le propre de l'activité cognitive. La classification serait alors la classification d'un certain son perçu comme ayant la propriété d'être un aboiement, et de l'animal invisible comme étant un chien. Précisons que nous n'avons pas affaire ici à une seule classification, mais à la relation entre deux. On peut avoir

une inférence causale entre deux unités signifiantes. La différence d'interprétation entre F. Dretske et J. Barwise & J. Seligman, qui est relatée dans l'ouvrage de ces derniers, repose sur la place du raisonnement causal dans la signification. Pour F. Dretske, elle est fondamentalement liée au signe, pour J. Barwise elle se caractérise dans la relation contraignante entre deux signes.
Enfin, chez F. Dretske, la relation entre deux classifications se marque par une fonction d'indication dans laquelle une première discrimination en indique une seconde. L'indication est alors caractérisée par une fonction depuis une classification ou discrimination (« fumée » par exemple) vers une autre. Le signe est donc l'association de deux classifications par une fonction d'indication. La fonction d'indication est alors considérée comme une connaissance. (C'est du fait de la connaissance *K* que la fumée indique le feu).

Le signe naturel est neutre par rapport à une situation particulière d'énonciation. Il est alors aisé de distinguer la signification non-naturelle comme liée à une situation d'énonciation et à une intention.

Le signe non-naturel (" mange ! " : ordre de manger) constitue un indice. C'est du fait du contexte social que l'on affecte tel ou tel contenu à l'entité exprimée " mange ". (Dans la perspective pragmatique dont P. Grice est un fondateur, le langage réalise un certain nombre d'actions). C'est alors un contexte ET un contexte de communication entre humains qui précise la signification du terme.

Le signe non-naturel est dépendant du contexte et de la situation d'énonciation. On peut reprendre la citation de P. Grice[196] : " Si le locuteur L veut dire de façon non naturelle quelque chose avec X signifie que L vise à obtenir que l'énonciation de X produise un effet Y sur un récepteur R grâce à la reconnaissance de cette intention. ".
" Il fait beau aujourd'hui " : interprétation : " on peut sortir ".
Le cadre de la signification non-naturelle est la reconnaissance de l'intention déterminée par le contexte de croyances et non par le contenu de l'information.
Il permet d'identifier des intentions. (Le contenu est dans l'interaction entre les participants).

Modes d'interprétation.
L'interprétation de l'information fait intervenir deux dimensions distinctes : l'interprétation cognitive et l'interprétation contextuelle.
On propose deux exemples simples afin d'illustrer la portée de concept et les deux dimensions, cognitives et contextuelles de l'interprétation du signe :

- "On est le 5" : cette information ne prend sens que si l'on parle du jour d'un mois précis. L'information n'est interprétée que dans un contexte très précis de connaissances antérieures (disons plus générales). C'est la dimension cognitive.
- "Analyse impossible". Un événement qui signifie naturellement l'incapacité du prélèvement à fournir une valeur sur une échelle pour une certaine propriété, et conventionnellement l'absence d'information disponible, peut transporter des quantités d'information importantes : l'absence de résultats signifie à propos de l'état du patient un stade avancé de dégradation des fonctions rénales. L'interprétation sémantique (naturelle et conventionnelle) illustre la dimension contextuelle. Elle ne doit pas être confondue avec l'inférence à propos de l'état du patient, qui caractérise une dimension cognitive.

Ainsi, l'information contenue dans un signal n'a pas forcément de lien avec la signification qui

[196] Traduction proposée par P. Bange, notes de cours de DEA Sciences du Langage, Université Lumière Lyon 2, 1987-1988.

est établie conventionnellement à propos de ce signal.

La caractérisation du signe s'inscrit dans l'hypothèse naturaliste, qui constitue l'une des deux possibilités de considérer le lien entre le langage et le monde. Nous expliquons brièvement ces deux hypothèses :

- la position naturaliste considère que le symbole ne fait pas partie de l'objet ; le symbole est donc un TYPE associé à une occurrence. L'occurrence d'objet est conforme au type au sens où tel objet est classé dans ce type (ce symbole) parce qu'il possède les mêmes propriétés que tous les autres objets de ce type. Le type est une classification.
- la position nominaliste considère que l'objet existe (sauf le nom propre) parce qu'un symbole lui est attaché ; le symbole constitue un INDICE puisque si le symbole n'est pas associé à l'objet, ce dernier n'existe pas. On a un indice à partir du moment où l'objet est reconstitué à partir d'un de ses composants (qui ne constitue qu'une attribution de symbole à l'objet). Un symbole vaut pour un objet dans le cadre d'une convention sociale d'attribution.

Ces deux définitions n'offrent pas les mêmes possibilités pour caractériser l'interprétation. (P. Grice, op. cit. 1957)

Signe, signal et interprétation.
Pour éviter toute confusion, il faut distinguer SIGNAL, INFORMATION et INTERPRETATION :

- SIGNAL : item transmis (le produit du choix au sens de Shannon).
- INFORMATION : ce qui est appris du signal (ce qui en moyenne peut être caractérisé comme fait DANS un signal). La moyenne correspond à la plus ou moins forte informativité associée à une information ; F. Dretske reprend ici la définition de Carnap & Bar-Hillel de l'informativité. Cette informativité correspond à la dimension propositionnelle de l'information, par laquelle celle-ci est associée à des valeurs de vérité. La notion de « fait » renvoie à la factualité, et non au produit d'une action quelconque.
- INTERPRETATION : inférences que l'on établit à partir des informations. (C'est le travail de l'interprète à partir de l'information qu'il a apprise).

Exemple : « Le chat : GN ». L'expression complète est un signal. (C'est un signe par exemple dans le contexte d'un corpus électronique annoté où on associe à chaque mot du corpus une annotation indiquant sa catégorie. Le document contient à la fois le groupe de mot et son annotation).
Information : le fait que la suite de caractère « le chat » EST UN GN et non une autre catégorie. Pour une suite de caractères, l'information est le choix d'une catégorie (GN) parmi d'autres possibles.
Interprétation : les connaissances que je peux déduire de cela : GN = ART + N, etc.

La signification, c'est le fait que la catégorie GN s'applique à l'instance « le chat ». La catégorie GN permet à la suite de caractères « le chat » de constituer une instance linguistique.

Fondements des classifications.
Si l'on suit la proposition sémiotique de F. Dretske, l'information étant caractérisée comme un phénomène pouvant être naturel (au sens où il ne repose pas sur des conventions), la classification d'un fait nouveau dans un type permettant de la reconnaître constitue un phénomène éventuellement de l'ordre de la perception. Ainsi, l'information à propos d'un fait

du monde débute par la capacité d'extraction d'un individu à l'aide d'une propriété. (Evidemment, la propriété permet d'extraire d'autres individus, d'autres instances du même type).
La théorie proposée par F. Dretske s'inscrit dans le cadre d'une théorie du signe qui ne se limite pas à la langue naturelle, mais s'étend à l'ensemble des processus inscrits dans le cadre d'une communication.
Il ne s'agit pas d'une définition fondée sur le « codage » des faits empiriques, mais d'une activité cognitive. Elle requiert donc une théorie fonctionnelle de la cognition. (La question de la nature intentionnelle du signe n'est pas traitée dans ce cadre parce que le problème est avant tout considéré comme sémantique).

En parlant de symbolisation de phénomènes du monde, on introduit une question fondatrice de l'épistémologie : comment un fait existe.
Ici, le fait est occurrent et est transmis par une propriété typique qui le caractérise. La propriété est une connaissance, et l'on n'apprend d'un fait occurrent que par la reconnaissance d'une connaissance établie.
L'information à propos d'un fait du monde nécessite la classification d'un individu à l'aide d'une propriété. Les propriétés sont des outils de classification : elles se valident ou pas sur un individu. La reconnaissance d'un individu par une propriété ne signifie pas que l'on limite la connaissance de cet individu à cette propriété. Simplement, cet individu existe et peut être l'objet de n'importe quelle expérimentation, observation, de façon à produire de la connaissance sur lui.

On peut proposer la représentation suivante de la classification, et de sa propriété fondamentale, qui est l'extraction d'individus de leur contexte :
Les individus extraits sont (a, b, c)
Les propriétés qui servent à extraire sont (A, B, C)
Les propriétés sont des outils de classification : elles se valident ou pas sur un individu.
Ainsi, les objets a_1, …a_n sont validés par la relation qui les associe à la propriété P. "La propriété P valide par relation l'objet a".
Les objets ne constituent pas un ensemble a priori mais sont appropriés pour cette propriété. En ce sens ils ne constituent pas une partie de la propriété, ni un élément de l'ensemble constitué par des éléments de la propriété. La propriété permet seulement de les typer.
On considère donc que la relation constitue un item associant par relation un objet et un opérateur conceptuel qui est la propriété. (Ainsi, l'objet peut être classé par des propriétés différentes à partir du moment où plusieurs niveaux d'abstraction peuvent l'identifier).

L'information transmise ne peut se satisfaire d'une seule caractérisation prédicative parce que cette dernière ne caractérise les arguments d'un prédicat que sur la base de son acceptabilité (autrement dit sur un fondement structural : la prédication sert à identifier une structure) ; en effet, l'information utilise une classification, mais cette dernière ne peut être réduite à l'ensemble d'acceptables d'un prédicat. F. Dretske propose donc un calcul du contenu de l'information : c'est la moyenne du contenu, à savoir de ce que l'on peut apprendre de l'information. (Ce n'est pas le signal, qui définit la quantité moyenne d'information mais qui ne caractérise pas le contenu). Autrement dit, le contenu de l'information, ce n'est pas la probabilité que l'on peut avoir de classer telle occurrence dans tel ou tel type plutôt qu'un autre : c'est la possibilité de caractériser fonctionnellement des types à partir d'autres types.
Par exemple, « retard de 30mns » m'apprend sur le « durée » de ma « situation d'attente ». J'apprends un certain nombre de choses sur les événements qui vont se dérouler dans un futur relativement proche.

On vient de parler de ce qu'il est possible d'apprendre d'une information précise. Néanmoins, on ne peut calculer cet apprentissage que par rapport au contenu. Dans le cadre de la théorie de F. Dretske, le contenu, c'est l'ensemble des effets produits sur la connaissance par la fonction d'indication. Si la signification est externe, parce que l'on a affaire à une signification naturelle qui n'est pas intégrée dans les états mentaux, par contre, ce que l'on apprend de cette signification est de l'ordre des états mentaux. Dans notre exemple, le contenu est l'ensemble des connaissances que l'on acquiert sur la phrase à partir du moment où l'on obtient l'information selon laquelle cette entité est une instance de GN. On peut donc faire certains postulats sur ce que sont les autres entités composant cette expression. Dans ce cadre, les contenus sont systématiquement des états mentaux. Il s'agit de l'ensemble des possibilités ouvertes sachant que l'on sait cela.
(Le contenu ne doit pas être confondu avec l'interprétation : cette dernière met en œuvre le contexte et les connaissances de celui qui interprète, alors que le contenu ne caractérise que l'étendue des choix possibles, à la fois en contexte de production et en contexte d'interprétation).

La quantité moyenne du contenu que l'on peut apprendre, ce n'est pas le contenu individuel (à savoir la classification) de l'expression informationnelle puisque le signe est naturel. On apprend en fonction de ce que l'on connait du type et de ce que l'on apprend du type dans un certain état mental. Ainsi, une information particulière permet d'apprendre à propos d'un type. « Le chat » permet d'apprendre sur ce qu'est la phrase que l'on analyse.

On peut alors reprendre la question des classifications et caractériser en quoi les classifications caractérisent la signification des expressions symboliques. Les classifications sont directement associées au concept de signe naturel (fumée/feu, roucoulement/colombe) et sont effectivement distinctes de ce qu'il est possible d'apprendre de l'information. Pour poursuivre cette réflexion, nous renvoyons à D. Laurier[117].

3.3.3. Interprétation, inférence et information.

L'information que l'on peut acquérir d'une occurrence est fondamentalement liée au type qui sert à la classer. La classification consiste à diriger l'interprétation, et donc à apprendre à propos de l'instance grâce à ce type. Ce processus peut alors être caractérisé comme un raisonnement causal, à savoir que l'on peut inférer des chaînes de causalité relativement aux connaissances et au contexte. De cette façon, on caractérise un comportement.

On peut tirer toutes sortes d'inférences à partir d'une instance classée : dans l'information selon laquelle un objet est un carré, on traite également l'information comme quoi c'est une figure planaire, qu'elle possède un contour, une surface, etc.
Ce processus inférentiel est l'information incrémentale : c'est le type plus général que l'on peut établir à partir d'un type spécifique.

Ainsi, on peut définir l'interprétation comme une suite d'inférences fondées sur des propriétés déterminées par la classification de l'instance. L'interprétation est donc relationnelle, puisqu'elle concerne à la fois l'instance et la propriété (parmi d'autres) qui sert à la classer.

Modélisation de l'information. Conditions à l'information.
On vient de voir les capacités heuristiques lié à la notion de classification et ce qu'elle permet d'apprendre de l'information. On peut maintenant préciser ce qui vient d'être dit en présentant

la modélisation de l'information proposée par F. Dretske.

Le postulat de départ de cette théorie de l'information est le maintien de l'intégralité des contenus de l'information dans la communication, donc y compris au cours de l'interprétation. (L'interprétation d'un contenu n'est pas dépendante des ressources du contexte, comme le proposent par exemple, les théories de la pertinence. En effet, pour celles-ci, ce sont les ressources interprétatives contextuelles qui déterminent le contenu de l'information, et plus généralement, sa valeur informationnelle). Dans les propositions de F. Dretske, l'interprétation est certes contrainte par le contexte considéré au travers de connaissances préalables, mais la signification, considérée au travers du principe de classification, est antérieure à l'interprétation. (Notons que la modélisation de ce contexte et de ces connaissances sera l'objectif de la théorie des situations).
Ce postulat est justifié parce que la signification naturelle (opposée à la signification non naturelle ou conventionnelle) est fondée sur les régularités d'indication entre les différentes localisations (émission et réception) d'une classification.

Les régularités d'indication fondent l'ensemble de la théorie de F. Dretske, et sont constitutives du signe naturel. On caractérise de la façon suivante les trois règles constitutives de l'information :

- principe du transfert d'information : c'est une fonction *f* qui permet à un signal *s* étant F (F étant une propriété quelconque), le fait que *s* transmet l'information que *a* est P. (" Le tintement de la cloche signifie que la récréation est finie "). C'est ce principe qui catégorise la théorie de F. Dretske du côté des théories de flux.
- Principe de duplication : pour que le principe précédent fonctionne, il faut systématiquement un lien entre les deux classifications. Si A transmet l'information que B et B transmet l'information que C, alors A transmet l'information que C. (Il faut effectivement que le tintement signifie que la récréation est terminée pour qu'il puisse transmettre l'information selon laquelle les cours vont reprendre).
- Principe d'adéquation : si le signal transmet l'information que *a* est P, alors le signal transmet autant d'information à propos de *a* dans « *a* est P » que n'en contient *a*. (Nous parlons ici d'information transmise et non de contenu ; le contenu que transmet le signal est nettement plus vaste). Ce principe se complète en considérant que le signal transmettant l'information que *a* est P transmet n'importe quelle information contenue dans « *a* est P ». (Ainsi, l'information selon laquelle " cette feuille est de format A4 " transmet l'information selon laquelle elle est rectangulaire, qu'elle est une surface planaire, etc.). (Il s'agit d'un principe lié à l'information incrémentale).

Modélisation de l'information. Caractérisation du cadre de communication.
Le raisonnement qui préside à la théorie de Dretske est probabiliste. Elle reprend les propositions de Carnap & Bar-Hillel elles-mêmes issues des principes de Shannon. La différence est qu'elle ne les traite pas au niveau des instances mais des types de la classification. Elle consiste à dire que le degré d'informativité d'une information se caractérise par le degré de probabilité de réalisation d'un signe par rapport à l'ensemble des autres possibles et vis-à-vis des unités précédemment réalisées. Cette proposition fonde la théorie prédicative de l'information ; elle est caractérisée ici dans le cadre de l'enchaînement des états mentaux. On le verra, cette proposition sera contestée notamment par L. Floridi et ne sera pas utilisée par J. Barwise & J. Seligman.
Le contenu de l'information, ce n'est pas la probabilité que l'on peut avoir de classer telle occurrence dans tel ou tel type plutôt qu'un autre : pour un patient, ce n'est ni qu'il puisse avoir la propriété d'avoir des « kg », ni qu'une certaine mesure de son poids, à un moment

donné, ait permis d'obtenir « 70 ». C'est la possibilité de caractériser la probabilité d'apparition d'une des unités parmi les autres possibles sachant que l'on a le type « kg » pour un certain type de représentation du poids à un moment donné : " 25-125 "). (Que ce soit « 70 » est la surprise ; c'est que l'on apprend de l'information et qui permet de modifier un certain nombre de croyances relatives à ce patient, en lien à d'autres informations).
Cette formulation ne rend pas compte qualitativement des contenus transférés, mais simplement du maintien de la quantité informationnelle entre les deux pôles de la communication. En effet, l'échelle de mesure est commune à l'outil de pesage et aux connaissances du récepteur.
La modélisation proposée n'est pas fondée sur une syntaxe ou sur une sémantique. Elle vise d'abord à représenter la transmission de ce contenu informationnel.

La première formulation de l'information concerne la quantité transmise. Les principales formules que F. Dretske adopte sont les suivantes :
I(information) transmise par le signal *s* pour le récepteur *r* avec un degré d'incertitude *E* s'écrit de la façon suivante :
$I_s(r) = I_s - E$

Dès lors que l'on transmet une information sur un état des affaires ou un événement, on peut écrire la formule suivante, caractérisant le fait que l'on considère le contenu particulier *p* relatif à l'état des affaires S ou l'événement *a* (transmis par le signal s) :
$I(S_{s)}) = \log 1/p(S_{s)})$.

Alors, pour le récepteur, on pourra écrire la formule suivante :
$I_s (r_s) = I(s_a) - E(r_s)$

Le problème reste celui de la détermination de la quantité d'information transmise, à savoir que les formules précédentes ne caractérisent que des principes, et sont nettement moins précises que les mesures de Shannon. (Rappelons que les mesures de Shannon ne caractérisaient que des signaux (ce qui est plus facile que des contenus)).

Afin de résoudre ce problème, deux méthodes ont semblé pertinentes :

- les méthodes statistiques reposant sur des calculs de fréquence (à l'exemple de la bibliométrie),
- une méthode sémantique qui partirait d'une définition du signe. C'est comme on le sait cette dernière qui a été retenue par F. Dretske, puis par J. Barwise.

F. Dretske introduit l'idée selon laquelle une information ne transmet qu'à propos de ce qu'elle indique dans le monde. La fonction d'indication caractérise toute classification considérée comme une prédication. (Cette indication ne transmet qu'une relation entre une entité instance et une entité type). Dès lors que l'indication est réalisée, il suffit de caractériser ce qui est indiqué du monde pour comprendre la quantité d'information transmise.
(Une telle proposition peut sembler totalement utopique à propos du texte littéraire, voire journalistique, mais relativement pertinente dès lors que l'on pense à l'information médicale et pharmaceutique).

Néanmoins, la fonction d'indication pose un problème important : on n'a aucun moyen de connaître les probabilités conditionnelles d'un événement dès que l'on sort de l'information symbolique, où les choix peuvent être représentés ; F. Dretske réfléchit en effet sur des faits et non des mots, donc les calculs sont impossibles. Cela permet à L. Floridi de caractériser

l'information définie par F. Dretske comme environnementale parce qu'elle est fondée sur la base de l'indication naturelle mais cette information ne peut être caractérisée que par sa représentation symbolique.

Sachant que l'on raisonne à partir de symboles mais considérant ce qui est indiqué par ces symboles, on peut approcher une définition plus précise de l'information.

Les formules qui suivent n'ont pas de valeur absolue, mais seulement de comparaison, notamment entre l'occurrence de l'événement et la quantité d'information que transmet le signal.
(1) "Je vis en Californie". On ne peut pas calculer l'information contenue dans l'expression.
(2) "Je vis en Californie, à Lafayette " : on ne peut pas calculer non plus.

Néanmoins, on distingue l'information pure (1) de l'information incrémentale (2).
(1) : information générée à partir de la source *s*.
(2) : information marquée par le signal *r* ("Lafayette") associé à la source *s*.

La quantité d'information générée par l'information la plus précise (2) est plus faible que celle générée par l'information la plus générale (1).
Alors même que la quantité générée par le signal est plus faible pour (1) que pour (2).

En effet, l'information générée par (1) permet nettement plus d'interprétations que celle générée par (2). Du coup, la quantité d'information transmise par (1) est nettement plus importante.

Le type CALIFORNIE symbolise nettement plus d'occurrences que LAFAYETTE, donc transmet plus d'informations. LAFAYETTE réduit l'interprétation possible de CALIFORNIE (qui est donc une information incrémentale : « Lafayette est une partie de Californie »).
Si à propos d'un patient, je ne dispose que de peu de données, alors je dispose d'une quantité d'information telle que je peux proposer des scénarios thérapeutiques très variés, voire contradictoires. Par contre, au fur et à mesure que je vais disposer de données plus importantes, le nombre de scénarios possibles diminuera d'autant. Un faible nombre de données génère une information très importante, qui me permet d'interpréter une situation dans laquelle un nombre important de phénomènes peuvent se produire. Par contre, plus les données sont nombreuses, moins je peux interpréter des situations dans lesquelles se produisent des phénomènes variés. Donc l'information est plus limitée. Précisons que cette information concerne une seule et même situation : elle peut être appliquée à la description d'une publication par une classification, mais pas à la description utilisant un thésaurus : la multiplication des entrées correspond à la multiplication des situations dans lesquelles la publication serait pertinente.

Cette caractérisation de l'information explique que moins une expression transmet de contenu informationnel, moins elle prête à des interprétations ambigües. A contrario, pour qu'une information ait une forte capacité informationnelle, qui se prête donc à ambigüités, il faut que l'expression qui sert à la véhiculer diminue les contenus possibles par un ensemble d'informations complémentaires. Cette proposition n'est pas sans rappeler celles de G. Chaitin, pour lequel la compression des données amène à une précision accrue de l'information. Néanmoins, la compression suppose que celui qui interprète sache que Lafayette est une ville de Californie, ce qui n'est pas le cas de l'expression de F. Dretske. Plus

il y a d'information, plus celle-ci est compressée, et donc moins elle accessible : il faut être doté de connaissances importantes ou spécialisées pour l'interpréter. D'où alors l'importance des « frames » et autres cadres, situations, pour caractériser ces limitations. Par ailleurs, ce principe amène à caractériser avec précision les connaissances préalables permettant justement d'interpréter de telles informations.

3.3.4. Fonction d'indication.

On peut maintenant préciser ce qui est entendu par fonction d'indication. Les exemples précédents dissocient la quantité d'informations générée par l'occurrence de l'événement à la quantité d'information transportée par le signal. Pour illustrer cette distinction, on peut prendre l'exemple suivant : on peut mesurer le fait que la corde A n'est pas plus grande que la corde B sans avoir à déterminer à la fois la longueur de A et celle de B. (Le signal indique que la corde B n'est pas plus grande que la corde A, alors que l'événement contient la longueur de chaque corde. Cela s'explique par le fait que le signal ne dit rien de l'échelle à laquelle les deux cordes sont représentées. En ce sens, l'événement contient nettement plus d'informations que le signal).

Néanmoins, F. Dretske rejette l'idée de toute augmentation de la quantité informationnelle entre l'information et sa réception en considérant qu'il s'agit alors de croyances et d'attitudes propositionnelles (ce qui permet d'éviter d'entrer dans une perspective pragmatique -p.55). Il rejette également toute dimension seulement intensionnelle à l'information[197] (ce qui permet de dissocier information et connaissance), et donc limite l'information à la symbolisation d'événements et d'états des affaires. Un état des affaires correspond à un état du monde auquel on réfère par le biais d'une information. Un fait correspond au format abstrait (ou cognitif) associé à l'état des affaires, et l'événement à un phénomène intermédiaire, tel qu'il soit à la fois possible de lier deux faits et deux états des affaires. Un état des affaires est extensionnel, à la différence d'un fait, qui caractérise l'interprétation de cet état.
Le niveau d'abstraction caractérisant l'information est plus bas que celui qui caractérise les connaissances ; cela permet notamment l'actualisation et la révision des connaissances par l'information. On retrouve cette idée dans la structuration du web où RDF constitue un niveau d'abstraction moins élevé qu'OWL.

On peut maintenant présenter quelques conséquences du principe d'indication dans l'ensemble des démarches relatives à la fois au signe et à la connaissance. Prenons un exemple canonique :
Le signal lumineux (x) étant dans un état F, il indique que ma batterie (y) est dans un état G (vide), cela par une loi nomique) ou fonctionnalité[118].

Se servant de cet exemple, L. Floridi associe ce type d'information caractérisée par un signe à un contenu sémantique référentiel. Or, il nous semble bien plus pertinent de considérer cette définition du signe intégrée dans le processus d'information comme une extension de la théorie peircienne, et cela sur plusieurs points :

- le principe de discrimination, associé à une démarche de classification. Il s'agit d'un principe sémiotique de base. (La fonction d'indication est fondée sur la discrimination d'une entité du monde par le fait qu'une propriété circonscrit une instance).
- A partir de là, on pourra proposer des fonctions d'indication différentes pour une

[197] Bien évidemment, le type servant à classer est de nature intensionnelle. Par contre, l'ensemble de l'information ne l'est pas.

même instance, telles que l'on puisse associer pour de mêmes objets des classes distinctes, respectant en cela la diversité du signe proposée par C.S. Pierce.
- Enfin, on pourra discuter de ces propositions relativement aux questions de connaissances en comparant les propositions de F. Drestke d'une connaissance préalable, et celles de J. Sowa à propos de l'utilisation de cette problématique du signe dans l'élaboration de connaissances, et surtout, de différents niveaux de représentation associés à un signe linguistique.

Globalement, cette caractérisation du signe est compatible à la fois avec les questions de perception et de construction de la connaissance à partir de la vision D. Marr (op. cit.), à la diversité des points de vue et donc des niveaux d'abstraction, et enfin avec la caractérisation dynamique des connaissances.

Néanmoins, on ne peut comprendre l'intérêt du principe d'indication que s'il est couplé aux deux autres conditions de la théorie : transfert et duplication.

Transfert d'information et duplication.
F. Dretske considère que la condition pour qu'il y ait effectivement transmission de l'information réside dans une cognition non-épistémique, proche en cela des théories d'une cognition fondée sur les processus de perception, notamment D. Marr. F. Dretske entend par cognition non-épistémique une forme de raisonnement qui n'est pas fondée sur des modalités et des représentations mais sur des expériences dans le cadre d'actions.
Le principe de duplication (ou principe « xerox ») constitue un point de départ du transfert d'information. Il s'articule au signe, (s est F), et permet de fonder les flux d'information. On peut prendre un exemple simple : si dans un contexte telle entité instance est de tel type, alors dans un autre contexte, ce même type permettra de caractériser la même instance : si « M. Smith » désigne un patient dans un certain service de l'hôpital, le fait que je transmette cette information vers la pharmacie entraîne le fait que « M. Smith » sera toujours bien ce même patient. La condition étant que je mentionne alors le service. Autrement dit, la relation entre les deux termes n'est pas altérée par le changement de contexte.

Après avoir caractérisé le transfert, on peut entrer plus précisément dans la définition de la condition de l'information est appelée duplication, qui caractérise le lien fonctionnel entre deux classifications. Deux classifications sont en lien fonctionnel dès lors que la classification inférée contient la classification source. A ce moment-là, la classification source est dupliquée. Depuis une certaine classification, on en infère une seconde, le lien entre les deux entités caractérisant une inférence. La fonction présente une signification inférée, qui est l'interprétation à partir du second terme de la fonction.
Ce principe complète celui de signe puisqu'il introduit une réponse à la question de l'annulation de la distance (temporelle ou géographique) dans l'interprétation et donc rend compte de l'intercompréhension. Il en fournit également les limites, puisqu'il indique les conditions (celles d'un rapport entre type et instance, donc une classification) pour que cette duplication soit efficiente.

Cette caractérisation du lien entre les classifications présuppose une connaissance. On abordera cette question plus loin, en parlant de la notion de connaissance préalable posée par F. Dretske, et de tous les problèmes que pose cette notion.

On se contentera pour le moment de la formulation simple de la duplication :

Si A transporte l'information que B, et B transporte l'information que C, A transporte l'information que C.

En disant que la copie transmet la même information que l'original, mais avec en plus l'information relative au fait que la classification dupliquée contient des propriétés supplémentaires, on rend compte d'un raisonnement par généralisation.

Duplication et flux d'information.
Chaque fonction dans la chaîne de l'information transporte une classification appropriée à propos de son prédécesseur, mais pas suffisamment pour que le signal soit original et informationnel pour le récepteur. De cette façon, on n'apprend rien à partir de l'information tant que l'on ne considère pas le signal transmis distinctement du signal produit. Le signal produit caractérise ce qui existe considérant la classification, alors que le signal transmis caractérise une instance perçue dans le cadre d'une classification.

A partir du moment où l'on va considérer l'information comme extérieure à la vérification, on doit caractériser le fait que si le signal transmet l'information que *s* est F (signal transmis), alors ce doit être le fait que : *s* est F (signal produit). (Pas de doute possible sur l'information).

Cette distinction a quelque conséquence sur la mesure de l'information : la quantité d'information qu'un signal transporte à propos de *s* est la quantité générée par *s*' s étant F. (*s'* désigne le signal tel qu'il est transmis).
La quantité d'information transmise est nécessairement la quantité d'information générée par *s*', *s* étant F. Le signal transmet autant d'information à propos de *s* que ce qui est généré par *s', s* étant F.

Ces principes définissent les contraintes de la sémantique. Dès lors, l'information qui est transmise dépend aussi de ce que le récepteur connaît de la source : on ne peut interpréter d'information à propos d'une source que si l'on connaît la classification (une relation entre une instance et une propriété).
Ainsi, un signal transmet l'information pour un événement ou un état des affaires *r* si l'instance dépend de *s*', *s* étant F. Cette dernière formule est toujours une condition à l'interprétation d'une information. Autrement dit, qu'il y ait effectivement une reconnaissance de l'instance.

Si *s* est un indexical ou un démonstratif, *s* est F est un contenu informationnel *de re*[198] : le signal *r* transporte l'information à propos de s qui est F.
Dès lors on a une phrase ouverte, à savoir (… est F) et un individu *a*, le signal d'un contenu informationnel *de re* est déterminé par deux choses :

- l'individu *a* à propos duquel le signal transmet l'information, (à savoir sa dénomination)
- l'information (déterminée par la phrase ouverte) qu'il transmet à propos de cet individu.

(Ainsi, *s* peut également être G, et *t* être F). (Le patient a un " poids ", mais également une " taille ", et d'autres patients peuvent être classés de la même façon).
A l'inverse, une proposition *de dicto* est de la forme S est F, où S ne renvoie à un individu que

198 De re et de dicto distinguent deux sortes de report d'attitudes, autrement dit de croyance. De re caractérise un report de croyance circonstanciel alors que de dicto caractérise un report de croyances non soumis à une circonstance. Par exemple, « je crois que j'ai fait une bêtise » est un report d'attitude *de re* dans la mesure où c'est dans une certaine circonstance que « j'ai fait une bêtise ». Par contre, « je crois que les hommes font des bêtises » sera une croyance *de dicto*. Toute l'importance de cette distinction est liée aux conditions par lesquelles on passe de l'une à l'autre, mais également au rapport à l'action.

par un quantificateur. (C'est de l'interprétation qui est transmise).[119]

Ainsi, la sémantique que développe Dretske à propos de l'information est fondée sur l'identification stricte du contenu du signal. En reprenant le principe de duplication, on peut caractériser le transfert de la façon suivante :
Si un signal transmet l'information que *s* (item de départ) est *F*, alors la quantité d'information que le signal transporte à propos de *s* doit être égale à la quantité d'information que le signal transporte à propos de *s étant F*. A ce moment-là, on peut inférer des propriétés plus larges.

Pour prendre un exemple très simple : *s* = *, * est F "étoile", l'information transférée par le signal "étoile" est alors bien *, * étant "étoile".
L'idée, contraire à une sémantique purement cognitive, consiste à dire que les symboles transmettent ce qu'ils représentent dans le cadre de la classification, parce que l'information fonctionne sur le principe de la duplication. C'est pour cette raison que la théorie de F. Dretske est considérée comme une théorie réaliste. La transmission du symbole contient la référence à l'instance symbolisée.

De cette façon, on ne peut pas dire plus d'une instance que ce que son type permet d'en dire mais inversement le type n'épuise jamais l'instance puisque le changement de type permet de dire autre chose (une propriété autre) de cette instance. Cette proposition a comme conséquence le fait que la signification demeure un principe externe par rapport à l'esprit.

Enfin, une telle théorie oblige à dire que les concepts (au sens cognitif) ne peuvent être considérés comme des canaux d'information qu'à condition qu'ils classent des instances et que ces classifications satisfassent au principe de duplication. Toute entité symbolique qui ne pourrait classer des instances serait définie comme un état mental.

Principe des connaissances préalables.
Les connaissances, dans la théorie de F. Dretske, sont systématiquement caractérisées comme prérequis :
"c'est en vertu de la connaissance k, qu'une fonction f permet à un signal *s* étant F (F étant une propriété quelconque), le fait que *s* transmet l'information que *a* est P". (" Le tintement de la cloche signifie que la récréation est finie : j'identifie la signification parce que je connais la contrainte : SIGNAL SONORE → AVERTISSEMENT ").

Cette connaissance se vérifie systématiquement : en ce sens elle n'est pas épistémique. (La connaissance se distingue ainsi des croyances et autres états mentaux).
Dans le projet de F. Dretske, l'information constitue la dimension observable de la cognition, et il est possible de caractériser les connaissances à partir du raisonnement classificatoire, en considérant alors des processus de causalité entre les différentes classes, lesquels se traduisent par des comportements. L'information constitue alors l'élément observable fondamental pour une théorie de la cognition dès lors qu'il s'intègre dans le cadre de processus.
Les connaissances et leur caractérisation comme niveau intensionnel de structuration des propriétés a souvent été contesté (Y.M. Visetti,[120] p.303). Une information ne pourrait circuler et enrichir une connaissance qu'à condition que la connaissance préexiste et rende son enrichissement possible. Nous verrons, entre autre grâce aux affordances, frames et autres propositions retenues par la théorie des situations que l'on peut éviter cet écueil. La cognition située et distribuée offrira d'autres développements.

Conclusion : propriétés des flux et distinction par rapport aux approches habituelles de l'information.

Ces propositions spécifient une approche informationnelle par rapport à une sémantique formelle ou linguistique parce qu'elles proposent de modéliser ce qui fait la fonction informationnelle, et qui plus est, en intégrant les dimensions cognitives et sémiotiques. Par ailleurs, cette théorie établit un lien entre classification et signification, en conformité avec la théorie des types. Cette possibilité permettra de développer une formalisation fondée sur ces principes mathématiques, ce qui constitue un avantage de cette théorie de la signification par rapport aux propositions sémiotiques saussuriennes ou peirciennes, qui elles, n'offrent pas cette possibilité.

Les flux constituent une approche qui reprend le programme de F. Dretske, mais en insistant sur la dimension sémantique. Ils prennent comme fondements des propositions au sens entendu par la théorie des situations et expriment de quelle façon chaque construction (classification instance/type considérée comme signifiante) contraint la suivante et ainsi établit une information. La prédication se situe donc entre les différentes classifications et permet d'expliciter la structure. Cette construction possède une dimension interprétative, dans laquelle la compétence sémantique des sujets contient l'ensemble des types et leurs instances possibles, de façon à produire des inférences valides.

Le programme des flux sera celui de la caractérisation de l'information par l'assemblage de composants hétérogènes dont la caractéristique essentielle consiste à abstraire distinctement des entités du monde. Ce sont les règles sémantiques de cet assemblage que nous essaierons donc d'expliquer au cours de notre présentation de la théorie des flux.
Nous avons voulu montrer dans cette caractérisation sémiotique des fondements des flux la dimension philosophique d'un modèle, en l'occurrence ici une part de la capacité de symbolisation de la langue et de la fonction d'information. Notre objectif consistera à construire à partir de cette compétence humaine un outil qui par ailleurs possède de nombreuses fonctionnalités dans le cadre du web sémantique. Comme tout outil, il externalise une pensée. Les représentations formelles qui vont suivre constituent les moyens d'expression des régularités, ce qui permet de modéliser le processus en question et d'en expliciter les propriétés. Elles n'ont pas d'autre enjeu.

Un autre enjeu explique ce passage par une sémiotique : comme l'ont montré G. Guizzardi & T. Halpin[121], les modèles sémiotiques et notamment le triangle de OGDEN, permettent de comprendre aisément les questions qui se posent dans les relations entre les ontologies et les lexiques. Par exemple, la distinction entre les niveaux lexicaux et ontologiques s'éclaire dans le cadre d'une perspective sémiotique. Les approches du web de données ne font pas l'économie d'un cadre sémiotique, que l'on soit dans la perspective de DOLCE ou de BFO. Nous traiterons de cette question dans la partie 6. Néanmoins, nous pouvons affirmer que notre fondement sémiotique se distingue de celui qui est adopté dans le cadre des ontologies. Notre position permet d'envisager une structuration plus complète et précise du domaine, et surtout intègre les processus.

Nous avons donc ici présenté un cadre comportant des propositions ayant une certaine parenté et illustrant une convergence que l'on peut résumer autour de la question de la classification, de la notion de situation et de la caractérisation de transferts. Il est également suffisamment large pour caractériser les phénomènes que l'on souhaite étudier, à la fois dans leur dimension

empirique que dans le cadre de conception d'outil documentaires et bibliographiques au sein du web de données.

PARTIE 4. Que sont les flux ?

Nous allons maintenant proposer notre propre approche des flux, qui reprend très largement les propositions de J. Barwise et J. Seligman[122].
Notre présentation des flux sera relativement exhaustive ; ses différents raffinements nous semblent pertinents pour un usage dans le cadre du web de données. Cette présentation sera suivie d'une présentation des enjeux généraux de la théorie dans le cadre général du web de données.

On fait l'hypothèse que les flux constituent des contraintes de la fonction d'information sur les objets linguistiques (lexiques, structures syntaxiques et composition sémantique). Ces contraintes sont des conditions à l'information : cela signifie qu'il ne peut y avoir d'information que parce qu'il existe un ensemble de règles limitant l'ensemble des entités pouvant être utilisées dans le cadre d'un certain transfert d'information. Cela signifie que tout canal d'information est considéré comme relativement spécialisé parce qu'il dépend des structures de données antérieurement construites.
.
La théorie des flux est un modèle logique destiné à rendre compte de régularités observable dans différents contextes : situations naturelles, phénomènes physiques, etc. J. Barwise et J. Seligman donnent de nombreux exemples de ce phénomène. Ils proposent par ailleurs de possibles pistes d'applications dans le cadre des systèmes distribués. Cela dit, il s'agit d'un travail essentiellement théorique. Il faudra attendre quelques années et la montée en puissance de la question des ontologies pour que l'on puisse avoir quelques amorces d'application de la théorie des flux[123]. Elles concernent essentiellement d'alignement d'ontologies (nous reviendrons sur cette question lorsqu'il s'agira d'évaluer l'impact des flux sur les Sciences et Technologies de l'Information).

Les flux tels qu'on les présente constituent une interprétation de la théorie présentée par Barwise et Seligman de façon à ce que la théorie puisse apparaître pertinente à la fois pour :

- rendre compte de certaines régularités sémantiques associées à la représentation d'états du monde,
- de constituer un modèle permettant de rendre compte d'inférences entre des structures de données structurées et signifiantes hétérogènes.
- caractériser un modèle d'analyse des contraintes à la formation et au transfert des expressions informationnelles.

En somme, les flux permettent de faire le lien entre les structures de données, la structuration et le transfert de l'information.

D'un point de vue plus technologique, l'apport des flux se situe dans la caractérisation des processus d'agrégation, et dans les promesses qu'offrent ces processus de produire de la signification en reliant des systèmes hétérogènes. Les thésaurus multilingues, les alignements de thésaurus, les compatibilités de classifications, apparaissent comme les outils les plus immédiatement bénéficiaires des flux. De cette façon, il s'agit de relier des structures de connaissances de façon à enrichir des descriptions de documents. (Il s'agit des connaissances telles que considérées dans le cadre des sciences de l'Information, et non des sciences cognitives).
Pour nous, le plus important se situe dans le cadre de la représentation du raisonnement : s'il est possible techniquement de caractériser le lien entre les structures de données, les flux explicitent ce lien comme étant un raisonnement.

Nous commencerons par présenter les différentes pièces construisant l'édifice (les classifications, les paires contre-variantes de fonctions entre les classifications, les canaux) et enfin comment cette théorie peut être utilisée.
Nous justifierons une expression des flux utilisant une théorie de types. Nous rappellerons donc les principales propriétés du typage comme théorie mathématique des classifications.
Ensuite, nous présenterons les flux en considérant les problématiques actuelles du web de données. La partie qui suit a comme unique objectif que de situer globalement la problématique des flux dans l'architecture globale du web telle que proposée par le W3C[199].Ce positionnement nous offre un cadre général dans lequel les flux sont positionnés. Ce cadre est plus général que notre projet, ce qui permet d'envisager d'autres développements et d'autres usages des flux que celui que nous avons choisi.

Propos liminaire : choix d'un exemple, mise en série et typification.
Nous avons travaillé essentiellement à partir d'une situation réelle, à savoir la circulation d'informations dans le contexte de la pharmacie hospitalière. Nous continuerons de faire référence à cette activité (notamment dans la partie sémantique) ; néanmoins, par commodité, nous proposons de travailler à partir d'un exemple plus simple. Il est présenté en annexe.

4.1. Dynamique des instances et des classes.

Le principe fondateur de la théorie des flux consiste à dire que l'information requiert des classifications, à savoir que tout objet est désigné par une instance et classé dans un type. La désignation et la classification constituent des propriétés constitutives de toute représentation symbolique. Ce postulat fonde les propositions de F. Dretske, puis la théorie des situations et enfin les flux. Le type permet d'interpréter l'instance, et l'instance d'ancrer le type. En d'autres termes, on caractérise ici le lien entre connaissance et monde, antérieurement aux inférences qui peuvent être produites entre classifications. C'est ce fondement que l'on aimerait préciser maintenant, parce qu'il nous apparait fondamental à la fois pour explorer ce que représentent les flux et comment les outils du web peuvent l'adopter pour concevoir dans ce cadre de nouveaux services.

Comme nous l'avons mentionné plus haut, les classifications constituent les fondements de la signification. Ce postulat explicité et validé, il sera alors possible de caractériser l'information. Celle-ci est effectivement considérée dans le cadre de la sémantique développée à partir des classifications. Ce principe de F. Dretske constitue l'axiome fondateur des flux, notamment parce qu'il permet de fonder la sémantique développée sur une sémiotique explicite. Il s'appuie sur la distinction <token-type> formulée une première fois par C. S. Peirce[124]. Dans ce cadre, un token est une réalisation physique ou donnée. Il peut s'agir d'une réalisation graphique, picturale, numérique. Un type est l'abstraction de la réalisation physique du signe. Il s'agit alors de catégories linguistiques, de schémas ou de frames réalisés au travers des tokens.

[199] Le choix de l'architecture proposée par le W3C repose sur des critères de généralité. En effet, d'autres propositions existent, notamment au travers du projet alternatif de P. Lévy, mais la puissance économique du consortium fait que ses propositions ont valeur de standard effectif. Par ailleurs, ils offrent suffisamment de liberté de maniement pour qu'il soit possible d'innover à l'intérieur du cadre de ces standards.

Préalable : utilisation de la théorie des types comme langage de représentation.
Le formalisme mathématique qui sera le plus adapté pour représenter les flux sera la théorie des types. Néanmoins, J. Barwise & J. Seligman utilisent encore une théorie ensembliste (ce qui leur est reproché par J. Van Benthem & D. Israel[125]). Les propositions de R.E. Kent[126] montrent comment la théorie des types peut être utilisée dans le cadre des flux.
Par la suite, les travaux qui utiliseront les flux se réfèreront à des théories de types. Avant de parler des classifications, nous allons donc introduire la théorie des types.

Une première approximation de la théorie a été présentée précédemment, lorsque les niveaux d'abstraction ont été définis. Nous y reviendrons encore plus précisément lorsqu'il faudra caractériser la sémantique de la structure d'information.
L'intérêt majeur des théories de types consiste à présenter un ensemble de concepts opératoires permettant d'organiser des données de façon alternative par rapport à la théorie des ensembles.

La théorie des types est utilisée particulièrement en linguistique formelle et en représentation des connaissances.
Elle ne repose pas sur le principe d'ensembles d'objets équivalents (comme la théorie des ensembles), mais sur les arguments qu'accepte une certaine fonction. A ce moment-là, on dit qu'une fonction accepte des arguments qui sont donc d'un certain type.
Les arguments par lesquels une fonction prend une certaine valeur constituent son domaine de signification. Le domaine forme l'espace d'application d'un type.
Les types constituent une catégorie logique. Lorsque les valeurs d'une fonction peuvent servir d'arguments pour une autre on postule une hiérarchie de types (fonctionnant par un système de treillis et non un principe vertical).

On peut décrire les fondements de la théorie des types d'une autre façon : on définit un type par la possibilité d'établir une égalité entre deux entités inscrites dans un même contexte. Une égalité caractérise le lien entre deux entités relativement à un certain type : $t =_T s$. (Par exemple, deux arguments qui seraient inscrits dans un même contexte –un même opérateur-, sont d'un même type parce que l'on peut introduire une relation d'égalité entre eux).

La théorie des types est d'abord un outil formel et non un appareillage descriptif. En ce sens, présenter une théorie de type consiste plus à expliquer des mécanismes d'abstraction qu'à représenter des règles du langage ou encore un raisonnement. C'est pour cela qu'il s'agit d'un outil permettant d'aider à la formulation et à la définition des objets d'application d'une théorie.

Quelques fondements de la théorie des types.
Ces fondements de la théorie des types servent pour préciser le vocabulaire utilisé et les bases mathématiques à la modélisation à la fois des flux et de la structure d'information que l'on étudie plus loin. Cette présentation peut sembler sommaire, mais développer à propos des types n'est pas un enjeu essentiel de ce travail.

L'objectif premier des grammaires de types et des grammaires catégorielles est d'arriver à se passer du concept de variable, qui pose l'ensemble des problèmes associés à la théorie des ensembles : ensembles infinis, singleton, notamment. L'idée consiste donc à partir non des entités elles-mêmes, mais des opérations ou des combinaisons que ces entités permettent.
Si les grammaires catégorielles et autres théories de types apparaissent aujourd'hui comme les alternatives les plus développées, néanmoins on pourra considérer qu'il ne s'agit pas des

seules tentatives, notamment dans l'objectif d'établir une sémantique : les logiques booléennes et les logiques de trait constituent d'autres approches.

On ne peut associer un type à un token qu'au travers de la réalisation d'une opération simple. En linguistique, ce principe a permis de mettre en place les grammaires catégorielles, dont le principe consiste à classer des tokens en observant les régularités d'opérations qu'ils réalisent. Ces opérations sont associées en syntaxe à la génération des expressions langagières.
Les opérations peuvent se réaliser à gauche ou à droite : a/b ou b\a. C'est l'apport d'Y. Bar-Hillel au calcul de K. Adjukiewicz.
/ : fonction à droite : diviseur à gauche.
\ : fonction à gauche : diviseur à droite.
Ces implications sont caractérisées par des choix d'arguments, ce qui donne les calculs de quotients (\, /).

La loi principale qui les caractérise est l'application :

$$\frac{X \qquad Y}{XY}$$

L'application peut se caractériser par des lambda expressions ; pour une token *x*, un type *Y* est une lambda expression telle que $\lambda x.Y$, qui est une lambda expression.
Les expressions peuvent être typées à partir du moment où les applications se doublent d'une abstraction. A partir du moment où une application est régulière, les tokens réguliers peuvent être abstraits dans un type. Cette abstraction peut se caractériser par l'attribution d'une propriété associée à ce type.

Sur ces opérations, on établit des raisonnements qui permettent de prouver la validité de l'attribution de propriétés. En effet, c'est par un raisonnement logique qu'il est possible de valider l'attribution de propriétés aux entités lexicales.
$X/Y\ Y \rightarrow X$ Forward (application ou modus ponens ou Elimination)
$Y\ X\backslash Y \rightarrow X$ Backward (abstraction ou raisonnement hypothétique ou Introduction).
Ces déductions naturelles ont été formalisées par G. Gentzen de façon à caractériser le type de raisonnement en cours.
Les règles de G. Gentzen sont une représentation de la déduction naturelle[127]. Les règles **intro** et **elim** permettent d'éviter la représentation de certaines complexités dans le raisonnement : ce qui est introduit peut être éliminé de la même façon dans le raisonnement.

Ces règles sont fondées sur des types et des formules. Les formules caractérisent des règles fonctionnelles. Quelles sortes de phénomènes représentent les formules ?

La présentation par les règles de G. Gentzen constitue une autre axiomatisation. (Elle remplace A -> B par "la structure R => B"). L'antécédent R est extrait de formules considérant les opérations Γ et Δ comme les contreparties des opérations de possibilité et de nécessité.
Toute règle logique d'inférence dans le calcul de G. Gentzen introduit un connecteur à la fois dans la prémisse et dans la conclusion.

A partir de là, on va caractériser les procédures de décision proposées par G. Gentzen. Elles sont utilisées dans les grammaires catégorielles pour considérer de quelle façon une phrase ($x \rightarrow y$) est déduite de règles.

G. Gentzen pose au départ des séquents, x_1, x_2 ,…, $x_n \rightarrow$ y
Où x et y sont des types.
G. Gentzen propose un ensemble de déductions permettant de contrôler l'attribution de tel ou tel type et d'inférer telle ou telle structure. Ainsi, un séquent est une expression à partir de laquelle il est possible de dégager un type. Nous préciserons plus loin ces règles, dans la mesure où elles sont fondamentales au niveau des « théories » de la théorie des flux.
Précisons que les règles de G. Gentzen constituent des outils destinés à valider le choix de type pour une construction précise. Les séquents constituent des formules du langage naturel, auxquelles on associe des types pour une conclusion qui constitue soit l'ajout soit la suppression d'un élément[128].

Cette présentation du cadre général de la théorie des types permet de comprendre l'enjeu pour les classifications, mais également pour toute représentation de séquences et d'explicitation de la façon dont les séquences sont obtenues.

Application de la théorie des types au langage naturel[200].
Parce qu'il s'agit du domaine de recherche dans laquelle la théorie des types a trouvé de nombreuses applications, nous l'expliciterons dans le cadre de la linguistique formelle. La problématique de l'utilisation de la théorie des types y est la suivante : à partir du moment où on peut transmettre des contenus occurrents, et qu'ils sont considérés comme tels, on a affaire à des structures types dans lesquelles les occurrences sont reconnues comme telles. Ces contenus occurrents sont des instances d'un certain type.
Dans ce cadre, le problème d'une grammaire, c'est de spécifier comment les occurrences reconnaissent certaines propriétés (qui constituent des types) et comment celles-ci sont structurés (entre autre pour composer des structures comme les phrases). Néanmoins, la problématique est loin d'être limitée à ces questions de grammaire, y compris dans le cadre du langage. Par exemple, des travaux comme ceux d'Urszula Wybraniec-Skardowska (op. cit.) montrent l'intérêt de la théorie des types pour clarifier les phénomènes de signification, interprétations sémantique et pragmatique.
Néanmoins, un des principes de base de la théorie est que le type d'une instance quelconque est caractérisé en fonction d'un de ses traits distributionnels, qu'il partage avec d'autres instances.

Pour un peu plus de précision, on suit T. Fernando[129] qui utilise la théorie des types dans le cadre d'une sémantique. Les types sont obtenus à partir de formules de prédication et de propositions.
La formule suivante permet de caractériser la construction de types :
$x_1 : A_1, \ldots, x_n : A_n \Rightarrow$ A type

Cette formule exprime la règle suivante : pour une instance, le fait qu'elle ait une propriété qui soit commune à un ensemble d'autres entités permet d'inférer un type. La construction d'un type est une démarche d'abstraction.
Par exemple, pour une occurrence x_1 d'un mot ("canari"), on actualise toutes les propriétés de tout "canari" (A_1) et le fait que ces propriétés soient valables sur l'ensemble des réalisations de "canari" possibles. Cette règle implique que l'ensemble des propriétés régulières de A en font un type. De façon plus usuelle, si l'on classe "canari" parmi les [noms], on identifie des propriétés valables pour tous les noms et permettant de caractériser "canari" dans ce type : par

[200] Nous présenterons plus loin les grammaires catégorielles qui constituent la principale application de la théorie des types au langage naturel. Nous nous focalisons ici sur des travaux de sémantiques formelles qui ne se fondent pas sur les grammaires catégorielles.

exemple, on peut associer comme propriété le fait que le [nom] requiert un déterminant, qu'il est d'un certain genre et nombre, etc.

L'application de cette règle permet à la fois de caractériser le fait qu'une entité est bien d'un certain type tout en restant une entité occurrente spécifique dans un certain contexte.

Maintenant, comment fonctionne une théorie de type ? Une théorie de types repose sur un nombre limité de types de base (ou catégories) et de fonctions de base. Ces fonctions de base servent à construire des types :
$f : B_1 \rightarrow ... \rightarrow B_n \rightarrow A$
f est une fonction prenant *n* arguments de type B_1, ..., B_n résultant dans un terme de type A. (Ainsi, notre occurrence "canari" de type B_2 ou [nom] et suit "le" de type B_1 ou [déterminant] pour établir un type A ou [prédicat nominal]).

Une fonction sans arguments est une constante : a : A. Elle est valable pour l'occurrence de "canari" de type [canari].

En appliquant les fonctions de base les unes aux autres, des termes complexes peuvent être formés :
$f: b_1,..., b_n : A$
Ainsi, on pourra traiter directement des unités occurrentes comme "le canari" comme étant d'un certain type [groupe nominal].

Donc, la théorie des types repose sur un certain nombre d'égalités (si ces égalités sont satisfaites, alors les deux instances sont du même type). Ce sont donc des équivalences permettant de construire une grammaire sur la base de types simples (nom, nom propre, verbe intransitif, verbe transitif, déterminant). Chaque règle grammaticale se caractérise par la combinaison de ces types simples. (Par exemple, un adjectif constitue un type complexe élaboré à partir du type simple [groupe nominal] : "le canari vert" s'écrira : ((GN) GN) et s'explique ainsi : pour le groupe nominal identifié "le canari", l'adjectif "vert" consiste à produire un second groupe nominal incluant le précédent et se caractérisant par l'adjonction d'un adjectif).

Il existe plusieurs théories de types destinées à contrôler des grammaires (et non à structurer des lexiques). Toutes ont quand même comme caractéristique de rendre compte de régularités grammaticales et non de classifications lexicales. (Il ne faut pas confondre les types avec des phénomènes comme les classifications arborescentes). (A partir d'un [NOM] quelconque, il sera possible de construire une représentation d'autres entités, en utilisant des structures comme la phrase).

Les types peuvent être fonctionnels ou relationnels. Cela signifie la chose suivante :

- type fonctionnel : projection de l'instance dans le type correspondant. (Il s'agit, comme on l'a vu précédemment, d'un processus simple de caractérisation de propriétés valant pour une instance, à l'intérieur d'une certaine algèbre).
- relationnel : il existe un contexte reliant l'instance et le type. Mais tous deux font partie d'une algèbre différente (à savoir qu'ils procèdent de différents vocabulaires de symbolisation).

La théorie des types constitue d'abord un outil pour la caractérisation syntaxique mais en fonction de catégories qui servent aussi à établir une sémantique référentielle (phrase nominale prédicative, verbe, distinction nom et nom propre). Ainsi, on caractérise de façon syntaxique des catégories qui par ailleurs auront une interprétation spécifique : par exemple les nominaux (et certains pronoms), les verbes, la phrase.

On peut prendre l'exemple de grammaire anglaise suivant :
S1 : NP -> VP -> S
VP1 : V -> NP -> VP
NP1 : D -> N -> NP
NP2 : N -> NP
D1 : D (comptable : "a")
D2 : D (non dénombrable : "many")
N1 : N ("lions")
N2 : N ("fish")
V1 : V
Ces règles s'appliquent à la phrase suivante : "many lions eat fish".
« De nombreux lions mangent le poisson ». Ce poisson peut être intensionnel ou extensionnel. Par contre, si j'ai un quantificateur « un » (« a » en anglais), à ce moment-là j'ai deux NP obéissant à la première règle du NP (NP1). La règle sera alors D1 -> N2 -> NP. De cette façon, ce type de règles syntaxiques permet de mettre en évidence les questions sémantiques qui se posent à la lecture de cette phrase.

Les grammaires catégorielles, fondées sur des types, permettent de représenter des phénomènes grammaticaux à l'aide de types très peu nombreux mais construisant systématiquement des catégories à l'aide d'opérations sur ces types.
Nous présentons maintenant un exemple issu de ces grammaires. En considérant deux types de base, N et P, on peut rendre compte de la production de la phrase "tout chien a une laisse". Ainsi :

"Tout" : tout N + N -> (N, NP) ou "une" : une N + N -> (N, NP)
"chien", "laisse" : N
"avec" : N + avec + N et cela produit un NP : (NP, (N, N))

Cette écriture permet une déduction :

tout	chien	avec	une	laisse
(N, NP)	N	(NP, (N, N))	(N, NP)	N

(une + laisse) : NP
(avec + NP) : (N, N)
(chien + (avec + NP)) : N
(tout + (chien + (avec + NP))) : NP

Il s'agit ici d'une illustration de grammaire fonctionnant avec des types. Il s'agit plus particulièrement de grammaires catégorielles. L'usage de la théorie des types proposé par R. Montague est relativement différent.

Une représentation typée sert dans le cadre des grammaires catégorielles à associer une représentation grammaticale à un raisonnement. On entend ici le fait que l'on peut déduire d'un type d'autres types et que cette déduction caractérise la production de l'expression. De cette façon, on associe grammaire et logique. Ce modèle de raisonnement sera utile pour nous dans la caractérisation des inférences associées aux flux.

Néanmoins, la caractérisation de types dans le cadre d'une représentation mathématique n'est pas le seul niveau d'utilisation de la théorie. On peut utiliser la théorie pour représenter plus généralement des classes d'entités ; ainsi, par exemple, les "restrictions sélectives" (des limitations à l'argumentation) permettent de construire des classes distributionnelles, qui classent des unités en fonction du fait qu'ils satisfont ou pas certaines opérations[130].

Le développement précédent avait comme objectif de montrer de quelle façon la théorie des types pouvait être utilisée dans le cadre de l'analyse linguistique. Nous reviendrons sur des applications plus précises à propos des structures d'information et pour fournir un cadre à la caractérisation des situations dans lesquelles les flux s'inscrivent.
Cette utilisation de la théorie des types a aussi un autre enjeu en ce qui concerne la théorie de l'information elle-même. J. Barwise & J. Seligman abandonnent la notion de probabilités pour lui préférer le concept de situation. (Cf. Kalfoglou, Y., & Schorlemmer, M. Using Formal Concept Analysis and Information Flow for modelling and sharing common semantics: lessons learnt and emergent issues, op. cit. p.3). Le concept de situation a le double intérêt d'autoriser un modèle sémantique en substituant une logique prédicative à un calcul de fréquences, et de centrer l'explication de l'information sur des modèles cognitifs. A partir du moment où le contexte est explicité, les questions relatives à l'informativité des expressions ne se posent plus dans l'absolu mais relativement à un cadre défini. Ainsi, la théorie des types a un impact sur la théorie de l'information, comme nous allons le voir dans la présentation des flux proprement dite.

Quelles classifications sont considérées ?
Les classifications relient des ensembles de tokens à des ensembles de types. Il ne s'agit pas d'un rapport générique/spécifique, mais d'une relation entre une réalisation dans le monde et un type. Un objet peut avoir différentes propriétés, et une propriété s'appliquer à différents objets. Les ensembles de tokens ou de types sont donc relatifs à des situations précises.

Les classifications sont exprimées par J. Barwise & J. Seligman dans la théorie des ensembles. Une classification est considérée comme une relation entre des tokens et des types et opère dans un certain type de situation.

On représente de la façon suivante la classification **A**.
Σ_A : Types (objets utilisés pour classer les occurrences de A)
tok(A) : token[201] : occurrences ou objets pouvant être classés.
$\vDash_A$: relation binaire entre tok(A) et Σ_A.

Un exemple extrêmement simple serait la classification des mots dans des catégories

[201] Un token caractérise une donnée identifiée minimale qui circule à l'intérieur d'un réseau. En linguistique, il s'agit de l'occurrence individuelle d'un symbole, d'un type. Il s'agit d'une unité réalisée de parole ou d'écriture. En informatique, il s'agit de la plus petite unité d'information dans une séquence de données. Voir http://oxforddictionaries.com/definition/token?region=us

grammaticales. Une grammaire travaille alors sur les types.

Deux ensembles et une relation servent à classer. Le système de classification est donc ternaire. L'approche spécifique à la théorie consiste à considérer que le type n'annule pas le token. C'est en fonction de la relation que tel token est d'un certain type.
L'association de tokens et de types est ponctuelle. Elle est fondée sur une relation particulière, qui est fondée sur une situation. (Si l'on change de situation, par définition la classification n'est plus opératoire).

Cela dit, il est possible de structurer plus précisément les ensembles qui servent de token et de type. En faisant cela, on sort de la seule théorie des flux pour envisager une sémantique associée aux entités. Le modèle perd un peu en généralité puisque l'on ne peut plus traiter de phénomènes physiques, mais gagne en clarté pour le traitement des unités symboliques.

Classification et description.
Les classifications constituent des prédications minimales à propos des objets. Ce sont donc des structures permettant de décrire certains parties du monde, et tout au moins des objets. Elles fonctionnent sur le principe de l'affectation d'un type relativement une propriété à chaque token. Cette affection est la classification. Cette capacité est toujours un héritage des principes de Carnap et Bar-Hillel (*op. cit.*), mais cela ne définit pas pour autant une information.

Il s'ensuit que la classification constitue un certain point de vue sur un token, et un flux consiste à inférer des descriptions depuis d'autres classifications à propos de mêmes instance (ou de certaines de leurs parties pouvant avoir une existence autonome dans l'univers.

En somme, une classification caractérise un état à propos d'une certaine entité du monde ou de l'une de ses parties. Une telle proposition peut sembler inattendue parce que les flux servent à caractériser des relations entre des entités purement intensionnelles. Les relations sont comme nous le verrons élaborées et justifiées à un haut niveau d'abstraction, mais les constituants de ces raisonnements sont fondés sur ces classifications.

On peut proposer quelques règles simples qui permettent d'illustrer comment fonctionnent ces classifications.

Deux types sont **co- extensifs** si leurs tokens sont égaux: $\alpha_1 \sim_A \alpha_2$ si tok (α_1) = tok (α_2)

Deux tokens sont **indistincts** si leurs types sont égaux. $a_1 \sim_A a_2$ si typ (α_1) = typ (α_2)

Ces deux règles de structuration permettent d'envisager quatre sortes de classifications.
Une classification est **séparée** si deux tokens étant indistincts, ils sont égaux dans la classification. $a_1 \sim_A a_2$ alors $a_1 = a_2$ dans la classification A.
Une classification est **extensionnelle** si tous les types co- extensifs sont égaux dans la classification. $\alpha_1 \sim_A \alpha_2$ alors $\alpha_1 = \alpha_2$ dans la classification A.
La combinaison de ces règles (classification séparée et extensionnelle) produit une **classification exponentielle**. La classification exponentielle caractérise le fait que pour un ensemble A, un token est classé dans un type ssi l'occurrence appartient à l'ensemble des types et l'ensemble des types est un sous-ensemble de l'ensemble global formant la classification.

$a \vDash \alpha$ ssi $a \in \alpha$.

La **classification par plan** associe une fonction entre deux ensembles A et B, telle que la fonction soit une classification depuis les occurrences de A vers les types de B. La relation de classification associée à la fonction est marquée par a ⊨ b ssi b = f(a).

4.2. Relations internes structurant les flux. L'infomorphisme.

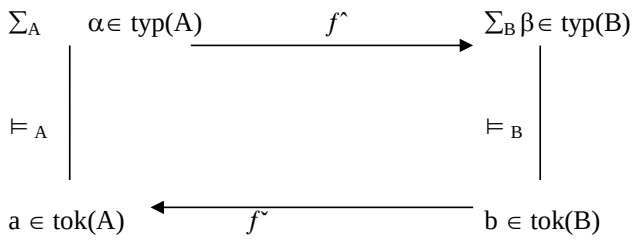
Nous entrons maintenant dans le cœur de la théorie des flux et donc dans la modélisation de la relation entre deux ensembles de classifications distinctes. La seconde étape de la structuration des flux consiste à comprendre de quelle façon ces différentes classifications sont liées entre elles. En d'autres termes, comment on peut inférer d'une classification une autre. Cette inférence constitue le cœur de la théorie de flux et constitue le socle de la modélisation de l'information. (Nous reviendrons sur ce second aspect après avoir caractérisé comment l'infomorphisme est circonscrit dans un canal. Nous verrons alors pourquoi et comment cette relation entre deux structures hétérogènes de données constitue une information).
On peut poser la question autrement : comment peut-on lier des propriétés comme la « cinétique », qui caractérise la faculté d'élimination, à des mesures de rapport entre des quantités d'une molécule donnée dans le plasma ? Ou entre cette mesure et une unité numérique sur une échelle de grandeurs ?

On peut poursuivre et étendre les exemples. L'important est de voir que ces données sont hétérogènes parce qu'elles appartiennent à de systèmes de symbolisation différents, mais elles nous apparaissent totalement liées les unes aux autres. La représentation de ce lien constitue une première étape vers la caractérisation d'une structure d'information. Rappelons que la structure d'information constitue la notion expliquant pourquoi et comment une information est interprétée. C'est une notion sémantique qui repose sur les relations entre des catégories définies d'objets symboliques. La structure d'information permet de passer des processus informationnels vers les processus linguistiques. Nous traiterons plus loin ces aspects ; pour le moment, l'objectif consiste à montrer quelles sont les relations qui fondent le flux d'information. Dans un second temps, nous montrerons en quoi cette régularité permet de caractériser une structure.

Le second principe des flux est donc l'infomorphisme. Il consiste à mettre en relation deux classifications par des paires contre-variantes de fonctions entre instances et types de chacune des classifications. Il s'agit d'exprimer une relation structurelle entre deux classifications par une fonction au niveau des types et une autre, contre-variante, au niveau des instances. Cette relation entre deux structures est définie par B & S comme constituant l'information.

L'infomorphisme se définit par une paire de fonctions : $f = \langle f^{\wedge}, f^{\vee} \rangle$, entre deux classifications, A et B.
Du coup, on a une double condition : on a un infomorphisme à partir du moment où à une fonction du type A vers le type B correspond une fonction de l'occurrence b vers l'occurrence a.

On représente l'informorphisme schématiquement de la façon suivante :



Cette schématisation représente le bi conditionnel suivante :

$f^\vee(b) \vDash_A \alpha$ ssi $b \vDash_B f^\wedge(\alpha)$
et

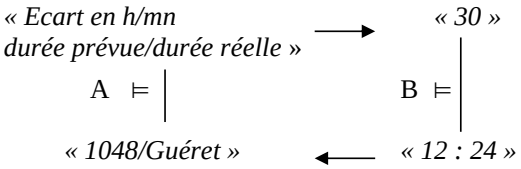
$f(b) \vDash_A \alpha$ ssi $b \vDash_B f(\alpha)$.

Cette structure peut être interprétée autant au niveau des types que des instances (Jayez & Mari, op. cit.). Dans le cadre d'une sémantique, ces auteurs adoptent un cadre logique modal.

Ces fonctions représentent un double raisonnement, inverse entre le niveau intensionnel et extensionnel. Les infomorphismes ne peuvent pas simplement caractériser des inférences abstraites mais également des fonctions (inverses) au niveau des tokens. Conformément aux propositions de L. Floridi, l'information exprime quelque chose à propos du monde.

En quoi cette structure constitue le fondement de l'information ?
Une première réponse réside dans le fait qu'il s'agit d'une traduction et que toute traduction, ou réécriture (et qui ne serait pas une tautologie), constitue une information. L'inférence d'une classification à partir d'une autre constitue un apport informationnel. (Il s'agit également de la position de J. Speaks[131]). Une telle proposition entraîne une définition de l'information fondée sur l'apport informationnel qui augmente la connaissance que l'on a du monde. Cette spécification ne peut être comparée à une succession d'états ou un enchaînement puisque le token de la classification cible est la source d'une inférence : le token de la classification finale permet d'inférer celui de la classification source.
Par exemple, j'ai un « train 1048/Guéret », qui constitue une proposition, qui peut être classé parmi les durées de trajet et les écarts entre durée prévue et durée réelle.
Cette classification n'est néanmoins possible que si on relie ce train à une unité dans le temps (« 12 : 24 »). Considérant alors la relation entre ce point dans le temps et la classification de « 1048/Guéret » (et en rappelant que ce token contient la feuille de route du train concerné), il sera possible d'inférer une unité permettant de classer l'unité temporelle par rapport à l'écart entre durée réelle et durée effective du trajet du train classé.
On peut prendre le raisonnement à partir d'un autre point de départ : l'écart entre durée prévue et réelle permet d'inférer « 30 » si l'on sait que l'on parle d'un certain point dans le temps et d'un objet qui peut être inféré à ce point dans le temps.

On peut représenter l'exemple ainsi :



Les classifications sont valides à l'intérieur de certaines situations, celles par lesquelles les lignes de trains sont considérées par leur durée (A) et les unités de temps se classent dans des durées (B).

Cette présentation explicite les relations entre les différentes unités constitutives de l'information. Elle ne comporte pas encore l'explication du lien entre ces deux classifications, c'est-à-dire ce qui explique pourquoi l'infomorphisme est possible et fondé : ce sera le rôle du canal d'information.

Nous posons maintenant un certain nombre de pistes pour la suite de la réflexion : une telle formulation présuppose des opérations qui ne sont pas représentées ; quelle est la conséquence de ces relations sur la construction d'une expression informationnelle, et enfin, quelle définition de l'information peut être déduite de cette formulation. Nous posons maintenant un certain nombre d'hypothèses qui seront ensuite développées dans le cours de ce travail. Nous posons ainsi des jalons pour la suite de notre travail, avant de développer d'autres aspects de la théorie des flux.

4.2.1 Quelques compléments sur les informations : les opérations présupposées.

En reprenant nos exemples, on pourra émettre l'hypothèse que tout infomorphisme requiert ou présuppose une opération extensionnelle, distincte et conditionnée par la classification. La classification constitue une activité mentale, à la différence de l'opération, qui se passerait quelque part dans le monde et dont les flux montrent certaines traces. Il s'agit d'un dispositif de symbolisation antérieur à la transmission de l'information. Par exemple, une information qui est un résultat d'analyse médicale reprend l'ensemble des paramètres de cette opération (l'unité de mesure, le temps, l'échelle et la valeur réalisée pour un objet identifié).
Inversement, cette opération extensionnelle requiert des classifications, comme notamment le fait que le patient possède une certaine propriété physiologique ou biologique, ou que la valeur affectée à cette propriété est d'un certain type, et est valide relativement à un certain temps. L'opération est donc dépendante des classifications.
Cette opération peut être réalisée à l'aide d'outils externes à l'esprit humain : ces outils sont des compétences humaines externalisées, pour reprendre le vocabulaire de G. Hutchins. L'exemple des opérations d'analyse biopharmaceutique peut être généralisé : une propriété existe parce qu'elle peut être mesurée donc instanciée chez les individus. Cette mesure requiert l'externalisation des capacités humaines de discrimination des quantités, ou des rapports entre deux quantités. L'hypothèse d'attribution de valeurs par des opérations externes à l'esprit humain permet d'introduire des questions de cognition distribuée (cf. G. Hutchins). Ces opérations ne sont pas exprimées ; seul leur résultat est reporté. (Les opérations ne sont pas des symboles mais des connaissances externalisées dans des outils. Les langages documentaires ont la même propriété).
L'externalisation des compétences humaines est considérée en extension. En effet, si les outils utilisent des symboles, c'est sans aucune mémoire ni représentation conceptuelle. Néanmoins,

les outils contiennent les connaissances ayant permis de les élaborer : les situations qu'ils caractérisent ont régulières et peuvent donc être abstraites.

Pour F. Dretske, la signification est associée à une classification. La classification donne sens aux tokens impliqués dans l'opération qui produit la signification. L'opération permet alors non de donner un sens aux choses du monde, mais de proposer une déduction à l'aide de paramètres à propos des tokens. Or cette opération (comme dans notre cas une opération de mesure) ne produit rien de plus qu'une symbolisation : l'association d'une entité mesurée (ou token) à une unité lexicale. C'est cette association qui constitue la classification relationnelle. L'ordonnancement (ou syntaxe) des unités symboliques réalisées et classées est fournie par les flux. Cet ordonnancement est structuré par les canaux, à un niveau supérieur d'organisation.

Qu'est-ce qui est caractérisé par l'information : ni l'opération réalisée dans le monde, ni les classifications, mais le fait que le résultat d'une opération présupposée et ponctuelle symbolisant un état dans le monde est traduit dans un autre état qui le spécifie. La classification permettant de typer le « train 1048 » dans les écarts des durées prévues et réelles conditionne l'opération qui spécifie ce retard : « 30 mn ». Inversement, parmi l'ensemble des « écarts de 30 mn » on trouve le « train 1048 » : l'opération consiste alors à lier les deux tokens (« train 1048 » et « 12 : 24 ») par une concomitance.

La caractérisation de ces opérations leur rôle dans les classifications et leur distinction avec les canaux d'information constitue un point fondamental pour explorer les dimensions cognitives des flux.

4.2.2. Comment le principe peut être interprété : notion de structure d'information.

Les flux nous amènent à définir les entités symboliques composant la structure d'information par leur hétérogénéité. Ce point constitue un aspect important de la distinction entre une approche par les flux et une approche de linguistique discursive où si l'on retrouve l'idée d'une bipolarité (entre thème et rhème par exemple) pour définir la structure d'information, il n'est pas question d'envisager l'hétérogénéité entre les deux parties de la structure (et encore moins leur composition interne).
Chez M.A. K. Halliday, mais pas seulement[132], la structure d'information est la traduction de la fonction d'information dans le discours. Sur un thème, on apporte un contenu nouveau (ce qui peut se représenter en reprenant des principes de représentation des connaissances comme chez E. Valduvi[133]). Systématiquement, une information nouvelle complète une représentation, ou une connaissance. Dans notre cas, le processus est différent puisque l'on considère d'abord la dimension lexicale et l'hétérogénéité des structures lexicales (à la fois dans leur organisation et dans la façon dont elles réfèrent à des objets du monde) qui sont mises en relation. L'assemblage se fait par la traduction d'un état dans un autre : l'information n'est pas liée à l'énonciation et l'interprétation d'un élément nouveau, mais à l'accolement de deux représentations d'états, dont l'un spécifie l'autre.

Une autre distinction est essentielle : les flux désorganisent la linéarité de l'expression parce qu'ils instaurent une proximité référentielle entre des termes qui ne sont pas nécessairement en relation de dépendance immédiate dans la structure de la phrase. Ils ne sont pas fondés sur une syntaxe, ce qui n'en fait pas un outil de linguistique formelle.

Les flux n'obéissent pas aux principes de compositionnalité : si chaque unité peut être décrite propositionnellement, seule la structure reposant sur des opérations régulières permet de rendre compte de la façon dont ces propositions singulières sont agencées.
On explique cela parce que les flux considèrent les unités linguistiques comme des marqueurs d'opérations, et non des unités du langage naturel.

Le second jalon essentiel de la caractérisation de l'information est le canal ; il rend compte de cet ensemble solidaire d'unités symboliques que l'on rencontre systématiquement en pharmacie hospitalière dès lors qu'il s'agit de transmettre une information relative à un patient. Le canal intègre le transfert de l'information ; on pourra alors caractériser une structure signifiante et proposer une analyse sémantique. L'objectif consistera alors à spécifier les limites de la quantité d'information transportée.
Néanmoins, il convient de préciser les conséquences de ce que l'on vient de présenter sur la définition de l'information, et notamment de le positionner par rapport aux propositions de L. Floridi.

4.2.3. Conséquences sur la définition de l'information.

Rappelons que ce qui nous importe, et qui aura des conséquences sur l'application du modèle, c'est la définition de la structure de données qui permet l'information. En partant des flux, et pour le moment de l'infomorphisme, on établit une structure minimale de l'information. Celle-ci permet effectivement de donner une définition de l'information et une explication de ses différentes règles constitutives qui permettront ensuite de dégager des contraintes et des possibilités d'application fondées. Rappelons que la structure d'information constitue une représentation linguistique d'un phénomène qui n'implique pas nécessairement ce type de réalisation. Rappelons aussi que l'application que nous proposerons consiste à opérer des mises en relation signifiantes fondées entre des structures hétérogènes de données sans que par ailleurs une structure d'information soit observable. Ces mises en relation ou inférences ne peuvent obéir à de mêmes critère de validité que par exemple les ontologies. Ces critères sont fondés sur des principes empiriques, à partir de l'activité observée. C'est ce que voulons spécifier maintenant à partir du cas de la pharmacie hospitalière.

On peut reprendre plus précisément la dualité entre les opérations réalisées dans le monde et une certaine organisation de la circulation des symboles que nous appelons flux. On aboutira à l'idée que l'information ne se caractérise pas seulement par les opérations des flux, mais par la façon dont se construit le rapport entre les types intensionnels et leur référence (ou tokens et ensembles de tokens) dans le monde. Ces opérations ont comme résultat un symbole et ce symbole classe un token, qui peut être identifié par un point dans le temps ou un nom propre d'individu. En ce sens, trois strates d'analyse distinctes permettent de caractériser l'information :

- la première caractérise l'appareillage technique (autrement dit les opérations) ;
- la seconde considère plus particulièrement les structurations sémantiques permettant de lier les différentes entités symboliques issues des opérations réalisées précédemment ;
- la troisième caractérise la structure d'information proprement dite, à savoir la relation entre les unités lexicales organisée de façon à assurer une interprétation conforme et adéquate à la situation décrite.

Ces trois strates sont descriptives ; la dimension cognitive permettra de proposer quelques hypothèses explicatives.

En ce sens également, les flux peuvent être analysés comme des contraintes à la représentation des résultats d'opérations réalisées dans le monde. Précisons encore qu'il s'agit ici d'un cadre d'observation : il nous sert à spécifier le modèle et non à l'utiliser. En effet, les usages de ce modèle seront nettement plus amples que la représentation et le transfert d'informations reposant sur des opérations techniques effectuées dans le monde.

Limites de la théorie des types et application des flux.

L'utilisation de la seule théorie des types bénéficierait ainsi des passerelles existant entre théories de types et grammaires catégorielles, lesquelles sont déjà largement utilisées pour décrire les structures d'information[134].

Or le fait que l'on ne s'appuie pas sur une syntaxe mais sur des opérations réalisées par des outils d'une part, et le fait que l'on caractérise un raisonnement que l'on peut utiliser pour réaliser des opérations concrètes sur des structures de données d'autre part, requièrent la caractérisation de la distribution des phénomènes que l'on cherche à décrire. En effet, une théorie purement fondée sur des opérations entre des symboles ne permet pas de fournir quelques indices sur les conditions de validité de la mise en relation de données structurées hétérogènes. La théorie des situations permettra de comprendre pourquoi on ne peut pas relier n'importe quelle structure avec n'importe quelle autre.

Cette perspective permet l'application des flux aux outils de description des documents, aux organisations de connaissances comme les classifications. On sait que les flux permettent de caractériser non pas la seule relation entre deux entités, mais la portée de l'une sur l'autre.

Nous avons insisté sur le fait que les classifications s'appuyaient sur des situations ; dans la présentation de J. Barwise & J. Seligman, ce point est très peu développé dans la mesure où il renvoyait aux propositions de la théorie des situations.

La théorie des situations rend compte entre autre de la relation entre des entités qui possèdent des modes de référence différents, et de contraintes liant les uns et les autres. De par son fondement informationnel et sa capacité à prendre en compte les dimensions à la fois linguistiques et cognitives, elle apparaît néanmoins la plus apte à caractériser les différentes dimensions qui caractérisent l'information.

Dans le cadre d'une application de ces propositions, une traduction du modèle écrit en utilisant la logique des prédicats du premier ordre (qui constitue le mode d'expression de la théorie des situations) vers une formulation logique utilisant les théories de types et les grammaires catégorielles pourrait s'imposer. Par ailleurs, ces écritures s'articulent facilement aux formalismes utilisés dans le cadre du web de données.

L'intérêt de flux, ce n'est pas seulement de mettre en relation des données, mais de porter des contenus depuis une structure de données vers une autre. Nous avons insisté sur la structure d'information alors qu'elle ne constitue pas le centre de notre projet : au travers elle, il s'agit de montrer, au travers des structures linguistiques, comment certains traits de la première structure sont transmis vers la seconde.

Ces questions s'inscrivent la formulation minimale de la théorie des flux, que nous avons présentée au travers des classifications et des infomorphismes. Maintenant, nous pouvons présenter les niveaux de structuration plus élevés.

4.3. Les canaux et les systèmes d'information.

Les canaux et les systèmes constituent les dimensions les plus abstraites de la structure d'information ; ils caractérisent la façon dont la structure s'intègre dans un cadre qui permet sa duplication et une interprétation commune dans des environnements distincts. (On parle alors d'interprétation à distance). En même temps, canaux et systèmes limitent la quantité d'information pouvant être transmise dans un seul message. Le message est donc une structure d'information transmise complète dans le cadre du canal. Néanmoins, en parlant de quantité d'information et de message, on postule un format : une structure d'information est un format caractérisé par une prédication limitée associée au transfert d'un contenu nouveau.

4.3 .1. Présentation des canaux.

Les canaux représentent le lien entre les classifications utilisées par une classe et considèrent les types utilisés par les classifications comme des instances. En d'autres termes, les canaux structurent les classifications et explicitent les liens fonctionnels entre les entités réalisées à la fois au niveau des types et à celui des tokens: les canaux caractérisent le lien entre deux classifications par une classification où l'instance est le produit cartésien des tokens des premières classifications et la somme individuelle des types de cette première classification. (La somme individuelle des types caractérise la façon dont une propriété et sa spécification fusionnent dans leur classification).
Considérons l'exemple précédent : le « retard » classe deux entités qui sont le paramètre des heures et minutes d'une part, et la spécification réalisée « 30 » d'autre part. « 30 mn » ou « 1/2h » constitue la somme du paramètre et d'un niveau sur l'échelle de mesure. Cette opération au niveau des types se double d'une autre au niveau des instances, où une entité individuelle *« 1048/Guéret »* est associé à un marqueur temporel *« 12 : 24 »*. Ces deux instances ne peuvent être confondues et leur lien fonctionnel est caractérisé par une paire associant un individu à une temporalité. Si l'on peut fusionner un paramètre et une échelle, on accole simplement l'individu et la temporalité dans laquelle on le saisit.

Enfin, le canal, comme classification des deux classifications composant l'information, explicite la dimension ontologique (entendus comme un ensemble de relations entre concepts de haut niveau d'abstraction) du modèle. En effet, les canaux assurent une relation entre deux classifications et permettent de prédire leurs relations réciproques. On peut également établir des similarités entre les canaux et les structures prédicatives verbales (notamment les verbes à un, deux ou trois arguments[135]).

De façon à expliciter l'ensemble des mécanismes dont nous venons de parler et y inscrire les canaux, nous allons reprendre l'ensemble des concepts du modèle précédemment décrits et y inscrire les canaux.

On caractérise trois sortes d'entités:
[a, b, c ,…] : ensemble des tokens [A, B, X, E] : ensembles de tokens.
[α, β, δ, κ, ο, υ] : ensemble des types [Γ, Δ, Φ, Ψ] : ensembles de types.

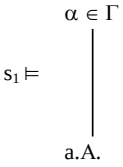
La relation ⊨ : opérations de classifications entre tokens et types

Chaque relation est donc un triple comme par exemple : <⊨, a, α>.

Le triple associe deux entités et une situation particulière dans le monde. (caractérisée par les symboles : [s_1, …, s_n]).

Les opérations produisent des expressions, définies comme des successions d'entités sélectionnées. Ces expressions peuvent être transmises et constituent donc des messages. La relation n'est pas dirigée : chaque token et chaque type sont caractérisés à l'intérieur de la situation qui sert à classer.

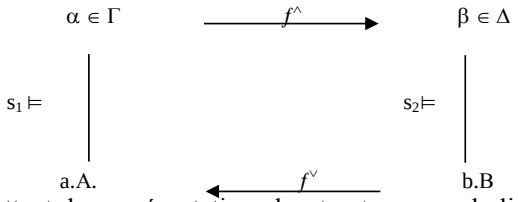
Ex. "M. Smith, kg,"



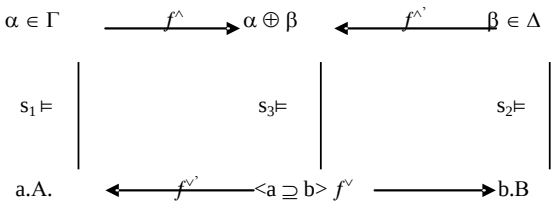
Les fonctions $f^{\wedge}$ et$f^{\vee}$, $g^{\wedge}$ et $g^{\vee}$, $h^{\wedge}$ et $h^{\vee}$ caractérisent des constituants d'opérations de même niveau. Les fonctions sont caractérisées par des paires contra-variantes entre deux classifications : chaque fonction relie deux entités de même niveau. Les formes b ←a et α→χ représentent un infomorphisme.
b ←a représente une fonction liant deux tokens hétérogènes à condition qu'une fonction inverse se réalise (par le biais d'une opération) au niveau des types α→χ.
Les triples sont: < s_1⊨, a, α>, < s_2⊨, b, β>,

$f^{\wedge}$: Γ→Δ

$f^{\vee}$: A←B
Ex. "M. Smith, kg,", " 2,58 mg/ml "



Les fonctions permettent la représentation de structures symboliques solidaires : " 2,58 mg/ml " et " M. Smith 14:54 ". Si les fonctions représentent des structures solidaires, alors on peut traduire cette solidarité au niveau des situations permettent les classifications. De ce fait, les fonctions représentent plus que le lien entre deux classifications :α→ α ⊕ β ← βcaractérisent le canal transfert à la fois les conditions et les résultats au niveau des types. a → <a ⊇ b> ←b représentent chaque entité et son inclusion au niveau des tokens.
La distinction des opérateurs est due à la différence de structuration des domaines de tokens et de types.



Chaque information nouvelle est un ensemble de fonctions depuis une classification vers une autre, où un type est traduit dans un autre, qui lui est spécifique, et inversement au niveau des tokens. L'ensemble, inscrit dans un canal, constitue un contenu informationnel.

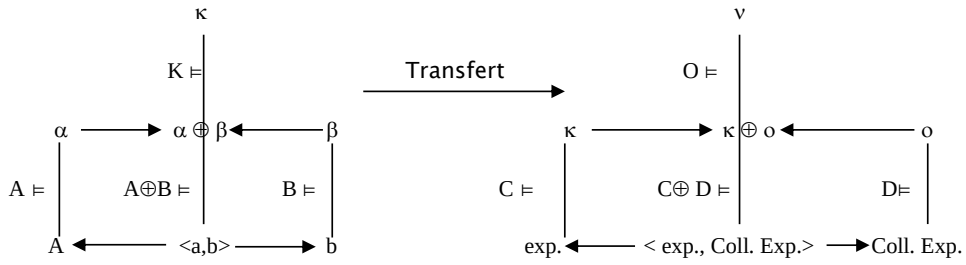
Néanmoins, cette structure ne peut être transmise que parce que l'infomorphisme est lui-même classé dans un type. Ce type est un canal et est commun au cadre de la production et à celui de l'interprétation de l'information. Le canal représente ainsi la condition de l'intercompréhension. Chaque structure transmise représente un fait. Le canal caractérise l'événement dans lequel le changement d'état se déroule à l'intérieur de cet événement. Une collection d'informations classées dans un événement peut éventuellement rendre compte de l'événement dans sa globalité.

Un canal est la classification des deux classifications initiales et de leur somme. Il représente donc une autre situation. La règle fonctionnelle qui traduit un type dans un autre plus précis se vérifie ici. Comme dans un cadre conditionnel, le choix du lexique de précision (Δ) est associé à la fois à Γ, Φ et Ψ (où Φ représente le lexique associé aux canaux et Ψ le lexique du canal lui-même). Dans cette représentation, $\kappa \in \Gamma$, $o \in \Phi$ et $\upsilon \in \Psi$.

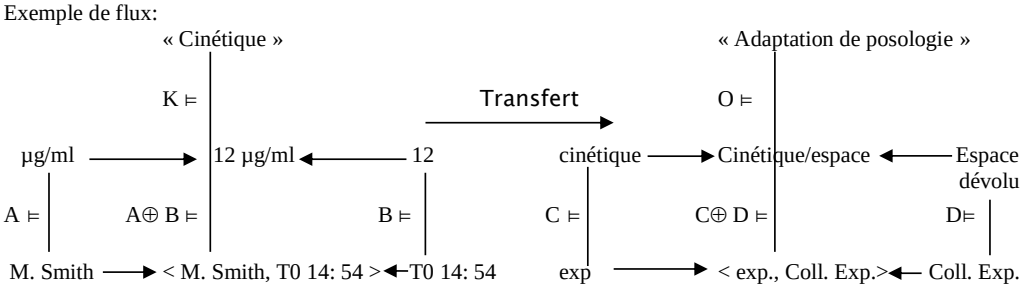
La distinction entre les deux schématisations peut être résumée par la distinction entre le processus de symbolisation et les opérations textuelles de manipulation de symboles.

Nous pouvons schématiser les canaux de la façon suivante :

Opérations de production d'information Classification dans un modèle de texte



Exemple de flux:



"Exp.": expression ou la structure globale de symboles transmis.
" Coll. Exp" : collection d'expressions sur une surface planaire.

Le flux d'information caractérise (1) la symbolisation dans un premier temps, et (2) l'identification du fait nouveau symbolisé. Le modèle de texte est expliqué par le principe de duplication. Le système d'information permet la duplication de l'ensemble des symboles structurés. Nous reviendrons plus loin sur les questions liées au texte.

Cette présentation peut être illustrée par un exemple simple : « Atos Origin a signé un accord de partenariat avec Inkra Networks en France », s'interprète comme le fait que pour les deux individus « Atos Origin » et « Inkra Ntework », leur propriété de « signer » spécifiées par « un accord de partenariat » se spécifie sur « en (localisation) » instanciée en « France ». (« En » est une locution de localisation et caractérise la portée de l'événement). Il s'agit ici du canal général. L'infomorphime de départ, qui concerne les deux acteurs de l'accord de partenariat est verrouillé par « avec ». « Avec » constitue alors le canal permettant de relier la propriété de signer avec la notion de partenariat. Cette structure classe les deux individus.
Cette réécriture intuitive requiert la sélection, pour les deux individus de départ, de l'ensemble des types d'action possibles pour eux, puis comment l'action choisie peut être spécifiée (on spécifie la « signature » par le « partenariat »). Cet ensemble est spécifié par un espace de validité, classant une localisation occurrente « France ». L'information n'est complète et ne s'interprète sans ambiguïté que lorsque le premier informophisme est classé dans un canal contenant une relation à l'identification de portée d'application.

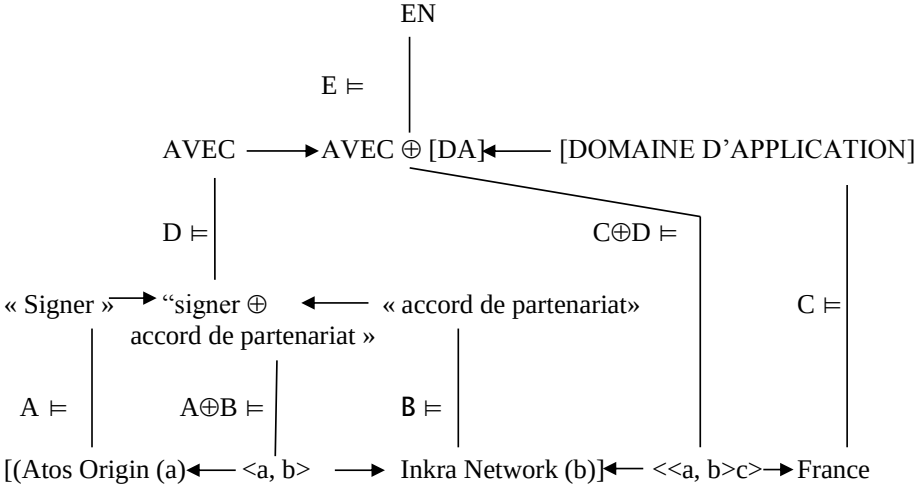


Figure 2. Schématisation de l'expression « Atos Origin a signé un accord de partenariat avec

Inkra Networks en France » en utilisant les flux.

Le canal (utilisant « avec et « en »), tel qu'on vient de le présenter, ne caractérise que la structure d'information. Il ne permet pas de représenter la transmission de l'information. Pour rendre compte de la médiation, et donc, d'un niveau de structuration représentant le transfert de l'information, on doit proposer un autre niveau.

Le canal, qu'il s'agisse de l'opération de mesure (comme « poids » ou d'un protocole plus technique comme « cinétique ») ou de la convergence entre deux entités comme dans le cas de « avec », se situe à un niveau d'abstraction plus élevé que les unités instances et types composant la structure. Le canal assure la convergence entre les deux classifications.

La représentation suivante explicite la prise en compte de la médiation dans le cadre des flux. Nous reprenons l'exemple en gare en simplifiant les opérations de structuration des flux. Les deux fonctions, APPLICATION et SELECTION, correspondent à un traitement de l'information destiné à satisfaire le principe d'économie (par SELECTION) et celui de référence (en caractérisant comment l'information a été obtenue) : APPLICATION.

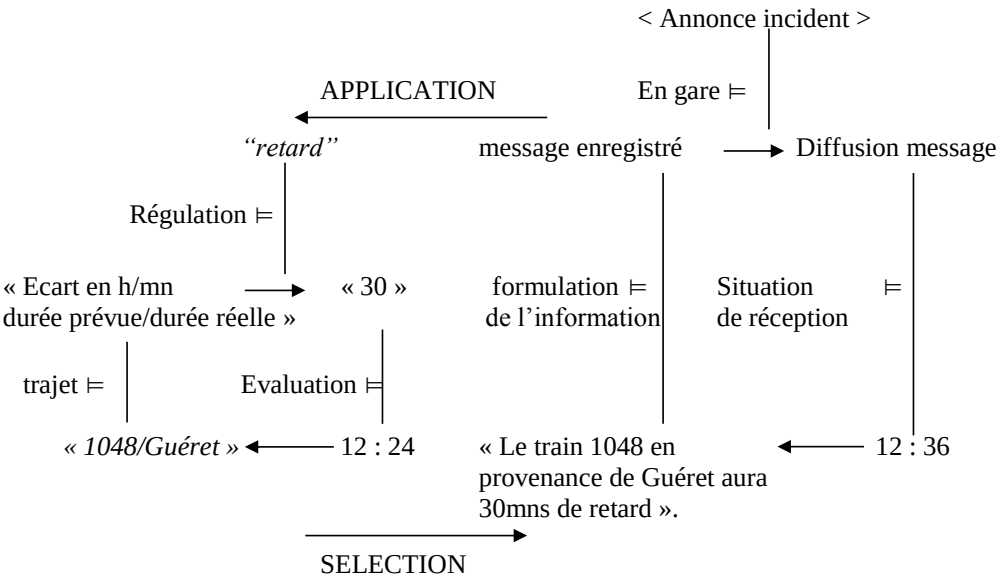


Figure 3. Schmatisation globale de la production et de la communication d'une information en utilisant les flux.

La représentation précédente ne concerne pas la structure d'information. On s'intéressera à la structure d'information à partir du moment où l'expression formulée pourra être décrite de façon à rendre compte de la façon dont elle réfère à la fois dans le monde et de l'univers des opérations associées à la production de la représentation symbolique des phénomènes du monde.
On pourra donc considérer que les représentations que l'on vient de proposer sont descriptives d'un fonctionnement empirique des flux. On n'a pas encore expliqué comment les flux

peuvent être utilisés, ni l'ensemble des phénomènes qu'ils peuvent caractériser, notamment ceux qui sont relatifs à l'interprétation et plus particulièrement à la structure d'information.

4.3.2. Les systèmes d'information.

On aborde maintenant la dimension ultime des flux, celle qui permet de clore la récursivité des classifications. En effet, dans notre présentation précédente, rien n'empêchait que l'on poursuive à l'infini les classifications et les canaux. En effet, si un canal limite le nombre et les types de classification, il n'y a aucune limite à la multiplication des canaux. Les classifications et les canaux sont clos à partir du moment où les connaissances préalables à la circulation de l'information sont est explicitées.

Le système d'information caractérise l'assemblage matériel et symbolique permettant la circulation régulière d'un certain type d'information, ayant une vocation particulière dans une activité. On entend par système d'information, de façon très classique, le système comme un ensemble organisé de ressources (personnel, données, procédures, matériel, logiciel,…) permettant d'acquérir, de stocker, de structurer et de communiquer des informations sous forme de textes, images, sons, ou de données codées dans des organisations. Cette définition du système apparaît peu éloignée de celle qui a cours en informatique. Pour une définition plus formelle, on peut se référer à Z. Pawlak pour lequel un système d'information se caractérise par la description d'un objet quelconque. L'appareillage des attributs et des valeurs possibles affectées à un objet constitue le fondement d'un système d'information[136].

Néanmoins, nous aimerions insister sur deux aspects, qui auront leur valeur dans le cadre du web de données et des capacités des grilles : d'une part, les capacités de transfert, d'autre part les dimensions symboliques. Pour nous, le système d'information possède trois règles minimales qui reformulent les principes de Z. Pawlak :

- il s'agit d'un dispositif technique et technologique régulier. Que le système soit numérique ou pas importe peu ; l'essentiel est que les opérations d'élaboration, de circulation et de stockage de l'information y soient régulières.
- Le système d'information assemble l'ensemble des lexiques et les règles de construction des expressions informationnelles. En ce sens, il construit le domaine à propos duquel l'information est produite.
- Enfin, le système d'information s'inscrit dans un cadre d'émission et de réception réguliers.

On présente maintenant, en reprenant le vocabulaire formel présenté précédemment, comment il est possible de rendre compte des systèmes :

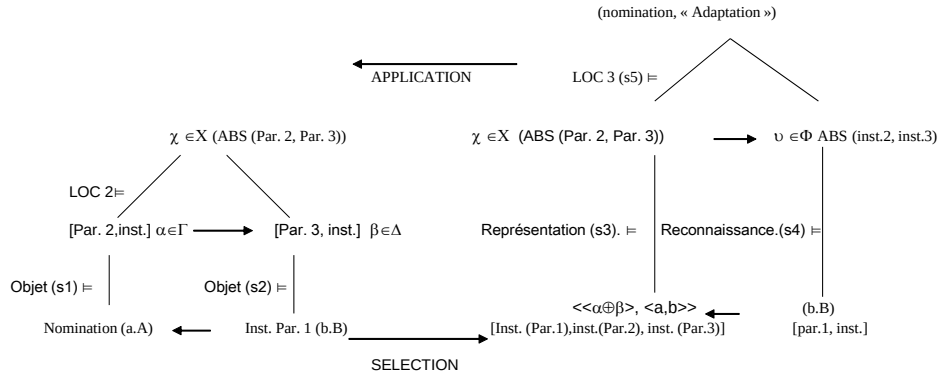


Figure 4. Schématisation des flux dans le cadre des informations relatives aux patients.

« Par. » indique qu'une opération paramétrée a été réalisée de façon à obtenir le type classificateur ou plus simplement la caractérisation de l'objet (comme dans le cas de [b]). ABS désigne une opération caractérisée par la seule abstraction d'une entité de niveau inférieur.
« S » désigne une situation et LOC un espace dans lequel les opérations et les situations prennent place, et qui constitue un fondement de la duplication de l'information.
SELECTION caractérise la fonction de duplication (assortie d'une réduction, d'où le terme de « sélection »). APPLICATION désigne la reconstitution des situations d'opération dans l'espace de formulation de l'information, à partir des représentations formulées dans le cadre de réception et reconnues.
La représentation décrit la situation dans laquelle l'information est dupliquée ; ce peut être une feuille, une base de données, etc. La reconnaissance caractérise la situation de celui qui interprète l'information représentée.

Est modélisé ici le cadre d'interprétation de la structure. Cette représentation est peu utilisée par ceux qui ont emprunté des éléments de la théorie des flux, notamment parce qu'il n'est pas toujours nécessaire de prendre en compte la médiation dans les mises en relation de structures de données hétérogènes.
Elle n'est pas non plus totalement satisfaisante parce qu'elle ne rend pas suffisamment compte de la notion de domaine et de sa caractérisation par les flux.

Cela dit, on ne rend pas compte ici de la façon dont les informations transmises sont interprétées. On ne modélise que le cadre de l'interprétation sémantique et non pragmatique. Ce sera le rôle des théories et des logiques locales.

4.3.3. Interprétation des classifications : les théories et les logiques locales.

On vient de considérer les contraintes à la production et au transfert des structures d'information. On caractérise maintenant comment cette structure est interprétée. On présente tout d'abord comment les problèmes de sémantique se posent à propos de l'information : comment les agents peuvent raisonner à distance à propos du monde, et avec une information partielle ?

On retrouve ici les questions relatives au réalisme des ontologies (opposé par exemple aux représentations de connaissances). Ce réalisme est aussi une propriété de flux, mais elle sera précisée par la théorie des situations. Les théories et logiques locales ne concernent que l'interprétation des flux.

Néanmoins, les théories et les logiques locales constituent des modèles de l'interprétation : les logiques locales restaurent les états extensionnels dans l'interprétation (ce qui est acquis par les logiques locales) et les théories structurent les inférences possibles dans la partie du monde représentée par l'information.
Comment, à partir de la situation d'arrivée, peut-on déduire la situation initiale ? Comment peut-on être sûr de l'interprétation d'une information, sachant que cette information provient d'un lieu distant, et que je n'ai aucune possibilité de la vérifier ? Comment puis-je être sûr que l'interprétation de cette situation correspond bien à un état extensionnel du monde en plus d'une représentation mentale ? C'est à ces questions que répondent les théories et les logiques locales.
Le principe est que l'interprétation réaliste requiert la connaissance des modes d'élaboration des contenus, donc des opérations. En ce sens, l'interprétation constitue la reconstitution des états successifs de l'élaboration de l'information. Ainsi, on peut considérer l'information transmise comme partielle par rapport à l'ensemble des entités impliquées dans les opérations.

La partialité s'explique ainsi : dans toute information, le contenu est présupposé, et seules les entités réalisées sont transmises et sous la forme la plus économique possible. Ce principe d'économie est à l'origine linguistique. Il est illustré en particulier par les grammairiens fonctionnalistes, notamment A. Martinet, mais a également été exploré dans le cadre de sémantiques fondées sur l'économie des moyens. En théories de l'Information, il est également expliqué par G. Chaitin (op. cit.). Nous reviendrons sur ce point dans la partie relative à la sémantique.

L'interprétation à distance caractérise la distinction de localisation et définit l'absence matérielle des instances sur lesquelles s'exerce le raisonnement interprétatif et sert à construire deux formes d'inférences :

- les théories, qui consistent à inférer des types à partir des types réalisés (causalité, connaissances préalables)
- les logiques locales, qui consistent, à partir des conclusions des théories, à inférer un nouvel état du monde, ou instance, dans le lieu de la production de l'information (reconnaissance de la situation initiale dans le monde).

On distingue les théories comme des inférences purement intensionnelles alors que les logiques locales valident les raisonnements tenus en les instanciant.

4.3.4. Présentation et utilisation des théories.

Les théories rendent compte du fait que l'on infère un ensemble d'objets du monde à partir d'une représentation restreinte (celle donnée par le type). Ainsi, l'inférence est le fait qu'à partir de l'entité réalisée faisant partie de l'ensemble des types de la première classification, il soit possible d'inférer l'ensemble des types de la seconde (ce qui est justifié par le fait que les types sont classés par les canaux). On se fonde sur le principe selon lequel le monde ne peut être atteint, donc que l'on peut interpréter un objet d'un certain type qu'avec un type plus large. Il s'agit du principe de la « feuille A4 », que l'on interprète avec l'ensemble de l'information incrémentale qu'elle contient (« rectangle », « surface planaire », etc.). Cette interprétation est la condition réaliste pour que l'on dispose d'un véritable cadre pour le

raisonnement à distance (et des possibilités d'inférence pour l'action).

On rend compte des théories en considérant que l'ensemble des types est structuré par des séquences, et que le premier terme de la séquence introduit le second, qui lui englobe l'ensemble de la classification. La représentation de cette inférence utilise les fondements de la théorie des types. Par exemple, à partir du moment où un indice type de quelque chose est réalisé, on infère la réalisation du type d'objet.
En effet, la théorie des flux reprend, pour caractériser les raisonnements à un niveau intensionnel, les règles de raisonnement établies dans le cadre de la théorie des types. L'inférence est caractérisée au niveau le plus abstrait, entre des connaissances qui peuvent être structurées par ailleurs par des relations [part_of] ou [is_a]. Ainsi, on peut établir une relation entre la structuration des ensembles de types par les théories et les ontologies.

On représente une théorie de la façon suivante :
Th(A) : ensemble des contraintes sur une classification A. Ces contraintes sont marquées par la paire <Γ, Δ> d'ensembles et la relation de conséquence $\vdash$.

La classification A est une relation entre l'ensemble des types classant les occurrences. ($\models_A$)
$\sum_A$ est l'ensemble des objets utilisés pour classer les tokens : ce sont les types de la classification A (où α est un type).
a : est un token. Il fait partie de l'ensemble des tokens de A.

Th(A) : $<\Gamma \vdash_A \Delta>$.
A :

$\alpha \in \sum_A$ où $<\Gamma,\Delta>$ constitue une séquence.

$\models_A$ |

a

La paire <Γ, Δ> est ainsi une contrainte de la classification. $\Gamma \vdash_A \Delta$ est une inférence qui signifie que Γ entraîne Δ dans la classification A. Il s'agit de l'interprétation qui est faite de la classification. (L'interprétation fonctionne ainsi parce que les types sont considérés comme des entités intensionnelles, donc des entités ayant un sens en dehors des réalisations empiriques).
Les contraintes constituent les règles de conséquence fournissant un cadre limitatif à l'interprétation. Les contraintes produisent des univers limités dans lesquels l'occurrence est interprétée.

Le trait fondamental de la théorie est une contrainte par laquelle, en considérant les règles présentées précédemment, on obtient la formulation suivante de l'inférence :
Pour une occurrence *a* de la classification A,

$(\forall \alpha \in \Gamma)[\alpha \models a] \Rightarrow (\exists \alpha \in \Delta)[\alpha \models a]$

L'interprétation faite dans le cadre des théories est intensionnelle : elle est marquée par une généralisation.
Cette interprétation est d'abord caractérisée comme une **inférence régulière**, qui, appliquée à

un type de la première classification, entraîne systématiquement un autre de la seconde.

Il s'agit d'une règle monotone. Elle est représentée comme suit :

$$\frac{\alpha \vdash \gamma}{\alpha, \beta \vdash \gamma}$$

La contrainte stipule que toute occurrence classée dans un type de la paire est interprétée dans le second ensemble de types de la paire.
Ce dernier interprète aussi des types n'appartenant pas au premier ensemble de la paire, mais d'autres ensembles de types de la classification.

Ainsi, on considère que les théories, qui constituent des outils purement intensionnels, montrent qu'un type et le token qu'il classe s'interprètent dans des catégories plus générales que le type servant à classer.

Le cadre limitatif des théories est représenté par trois règles :

- Identité : $\alpha \vdash \alpha$,
- Monotonie $\Gamma \vdash \Delta$,, donc $\Gamma, \Gamma' \vdash \Delta, \Delta'$
- limite globale : $\Gamma, \sum_0 \vdash \Delta, \sum_1$, pour chaque partition $(\sum_0, \sum_1)$ de $\sum'$, alors $\Gamma \vdash \Delta$.

Le fondement d'une théorie est sa dimension causale, à savoir que la conséquence se manifeste au travers de partitions d'un ensemble plus général. (Rappelons qu'une partition d'un ensemble se caractérise par l'absence d'intersection entre les deux sous-ensembles, sachant que l'union de ceux-ci constitue l'ensemble). La conséquence est la relation entre les deux ensembles disjoints. Les sous-ensembles qui encadrent la conséquence en constituent la contrainte.
(La causalité se caractérise ainsi : le type de la classification inférée est la conséquence de la celle de la source. La causalité définie ici n'est pas identique à celle qui relie deux classifications dans l'informorphisme). Elle sert à l'expliciter.

La théorie consiste donc à placer sur la classification une conséquence entre les deux sous-ensembles formant une séquence. Appliquée à une classification, une théorie amène à considérer que l'occurrence est de tout type d'un sous-ensemble, est de certains types du sous ensemble de conséquence. (Ainsi, on préserve la partialité). On peut à ce titre reprendre la définition de J. Barwise & J. Selgman (p. 118) :

" la théorie Th(A) = <typ(A), $\vdash$A>, générée par une classification A est la théorie dont les types sont les types de A et dont les contraintes sont l'ensemble des séquences satisfaites par toute occurrence dans A ".

Il existe dans les propositions de J. Barwise & J. Selgman un nombre important d'opérations possibles en utilisant les théories. Nous renvoyons à l'ouvrage pour plus de détail.

Dans le cadre des théories, on préserve l'incertitude de l'interprétation liée à la distance, à savoir :

- la possibilité d'exception,
- l'intensionnalité de l'interprétation à distance,
- la spécificité du système de transfert d'information.

Les théories constituent des inférences n'intervenant que dans la dimension intensionnelle : l'interprétation est sans lien à des phénomènes empiriques et il n'est pas possible de lier raisonnement et action. Les logiques locales permettent d'ancrer ce raisonnement dans les instances et les phénomènes du monde.

4.3.5. Présentation des logiques locales.

Les théories rendent compte de l'interprétation intensionnelle. Les logiques locales, elles, rendent compte des interprétations extensionnelles, à savoir la reconnaissance d'un état dans le monde.
Les logiques locales apportent une interprétation dirigée vers une représentation extensionnelle. Elle ne peut être assimilée à une reconstitution du token de la source mais à l'état du monde le plus plausible à la suite de l'inférence proposée par la théorie.
Les logiques locales apportent plus que simplement une représentation du token. Il s'agit de représenter l'état du monde hors de la seule propriété qui la représente dans le cadre de l'informorphisme initial. (C'est en ce sens que l'on peut représenter les contenus comme les contre- factuels).
Ce n'est pas l'individu et sa propriété qui sont caractérisés par les logiques locales, mais ce que les classifications et les théories permettent d'apprendre à propos de l'individu en tant que tel. Elles caractérisent ainsi les états complexes qu'il est possible d'inférer à partir d'un raisonnement. On caractérise ainsi les connaissances acquises à partir de l'information.

Les logiques locales se définissent par la reprise de la classification et de la théorie, mais en y ajoutant un sous ensemble de tokens, celui des tokens normaux. A la différence des théories, les logiques locales permettent d'établir des tokens d'état, satisfaisant les contraintes de la logique. Ces tokens sont l'aboutissement d'une interprétation. On présente une logique locale avec la formule suivante :

$L = <tok(L), typ(L), \models_L, \vdash_L, N_L>$

On a comme constituant de la logique, la classification, la théorie, et donc la conclusion marquée par un sous ensemble de tokens : $N_L \subseteq tok(L)$

Une logique locale est fondée si elle rejette toute instance qui ne satisferait pas les contraintes de la théorie. (Certains états peuvent ne pas satisfaire toutes les contraintes : ce sont les exceptions). Un token est normal parce qu'il satisfait toutes les contraintes du raisonnement.

La logique est complète si toute séquence qui apparaît pour tous les tokens normaux est dans la relation de conséquence $\vdash_L$.

Si l'on interprète à partir des classifications, on doit aussi pouvoir inférer que le token existe en dehors de la classification. Les logiques locales reposent sur l'explicitation d'un présupposé : il existe un patient qui n'est pas seulement décrit par un poids à un moment donné. L'objectif des logiques locales est de montrer que l'on interprète un individu qui n'est pas seulement un objet classé dans une seule propriété, et qui peut se transformer entre le moment où il avait cette propriété et le moment de l'interprétation.

Les logiques locales permettent de caractériser l'application d'un raisonnement : il ajoute à un produit intensionnel de raisonnement (marqué par les théories) l'état produit par ce

raisonnement dans le monde. L'état du monde produit par le raisonnement et son application est plus étendu que le token ayant servi à le produire parce que l'inférence en extension est construite à partir des théories. En ce sens, on peut considérer l'information comme étant partielle par rapport aux conclusions que l'on peut obtenir à partir de l'information.
Empiriquement, on raisonne à propos d'un patient précis, et non uniquement à propos des informations que l'on a à propos de lui. (C'est l'une des raisons de l'échec de programmes d' I.A comme MYCIN, et des représentations de connaissances que de ne pas intégrer les dimensions extensionnelles dans les conclusions des raisonnements).

Dans notre exemple de pharmacie, toutes les théories satisfont la séquence ; n'importe quel individu a une élimination rénale que l'on peut déduire de la mesure de cinétique, corrélée à d'autres types d'indices associés à des opérations. On interprète ainsi des phénomènes généraux comme l'élimination rénale à partir de faits particuliers, attestés par des résultats d'opération transférés. C'est le principe de tous les raisonnements de contrôle et de diagnostic.
Par les théories et les logiques locales, il est possible de caractériser les états de plusieurs individus dans le monde, et ainsi d'expliciter le fait qu'une information peut être interprétée relativement à différents individus postulés dans le domaine des individus. Par exemple, c'est ainsi qu'une expression informationnelle en pharmacie, qui est interprétée relativement à l'individu « patient », peut être également interprétée relativement aux autres individus possibles, la « bactérie » et la « molécule ».

Caractérisation du raisonnement représenté par les flux.
J. Barwise & J. Seligman ne spécifient pas exactement quel est le raisonnement en question. Nous développerons l'idée d'un raisonnement causal, aidé en cela par les propositions de J. Jayez & A. Mari (op. cit.) et de J. Collier[137].

Le raisonnement est double : d'une part, il y a l'infomorphisme qui s'inscrit à l'intérieur de l'information et qui permet de spécifier une classification par rapport à une autre. Ce raisonnement est celui dont parle J. Jayez & A. Mari. Il ne doit pas être confondu avec les théories, qui elles, proposent un raisonnement considéré comme une généralisation, et qui implique les contenus de l'information. Enfin, les classifications ne sont pas spécifiées comme étant des raisonnements. Nous reviendrons sur cette question en parlant des situations. Si l'on suit la pensée de J. Barwise & J. Seligman, seules les théories constituent des raisonnements.

4.3.6. Récapitulatif, application des flux sur exemples et contrôle.

La présentation théorique que nous venons de proposer ne permet pas de rendre compte de la façon dont on peut observer et construire les flux à l'intérieur d'un cadre empirique comme celui qui nous a permis justement de les observer. Indépendamment de l'application que l'on proposera dans le cadre du web de données, on présente ici un certain nombre d'observations et de modélisations informelles qui permettent de présenter le type de phénomènes dont les flux peuvent rendre compte. Nous nous garderons d'interpréter ces phénomènes réguliers. Nous reviendrons sur le cadre théorique dans lequel ils prennent sens dans la partie 5.3., relative aux hypothèses de la cognition située et distribuée. En effet, nous ne voulons pas d'interférence entre notre explication des phénomènes dont les flux rendent compte et la façon dont une théorie peut expliquer ces modèles.

En pharmacie hospitalière, l'information reçue est interprétée par inférences des résultats dans des modèles d'analyse, de contrôle et de prédiction. Ces inférences de généralisation sont strictement intensionnelles ; très concrètement, il s'agit de modèles de populations statistiques. Or, ces théories ne permettent pas simplement de généraliser : elles permettent d'inférer un état particulier de l'objet de référence. Cet objet est le patient : la théorie permet d'envisager le comportement du patient dans l'ensemble des états représentés par la population, la logique locale la spécificité de comportement de ce patient par rapport à la population. La première interprétation est une théorie, la seconde est une logique locale.

L'interprétation associée aux flux intègre la dualité associée aux classifications. Les canaux héritent de la propriété de l'information à produire un discours (attribut, classe ou propriété, tout dépend du cadre théorique choisi) à propos d'un objet : ce dernier est interprété par les tokens (par le produit cartésien des instances de l'information) à partir des types (à savoir des représentations qui ont été produites à partir de lui. La somme individuelle des propriétés entraîne un autre type dans le cadre d'une théorie. Ce type est la connaissance acquise à propos de l'objet, et cette connaissance s'applique sur un objet plus large que le seul individu objet qui sert de référence. En effet, l'information est toujours partielle par rapport aux individus ou objets sur lesquels elle propose du sens : c'est le principe même des classifications. Par conséquent, une interprétation à l'aide de logiques locales permet d'intégrer cette incertitude.

Caractérisation des flux sur les exemples et sur terrain.
On introduit l'analyse empirique par une représentation du cadre matériel dans lequel les flux s'inscrivent. On schématise ici les données tout d'abord spatiales avant de proposer une modélisation particulière de cette dimension.

On schématise la trajectoire de l'information depuis la chambre du patient dans l'hôpital jusqu'à la table de travail du pharmacien (où l'information est enregistrée et interprétée) afin de caractériser le cadre dans lequel les informations sont produites, circulent, sont dupliquées et interprétées. Cette schématisation a également comme objectif de fixer un cadre pour caractériser les raisonnements que l'on pourra identifier relativement aux principes de la cognition située et distribuée.

On distingue trois types de localisation dans l'activité : le corps du patient, le lieu d'interprétation de l'information et le lieu où l'information est construite en utilisant différents outils de mesure.

Les différentes localisations sont exprimées de façon verticale et la direction du flux est marquée par les flèches. Dans ces différentes localisations, on identifie des individus et des dispositifs matériels techniques:

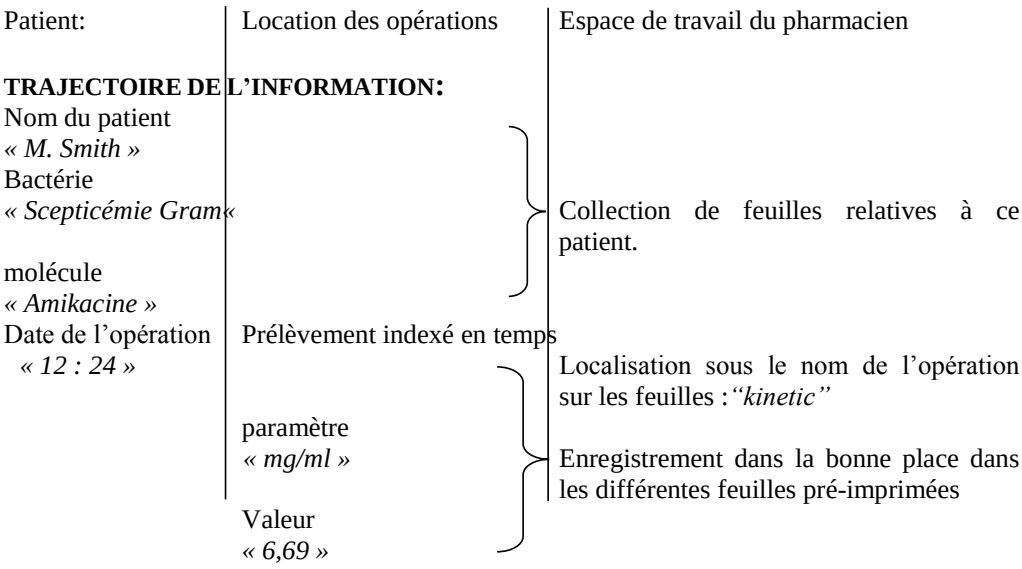


Figure 5. Schématisation des composants de l'activité et de la trajectoire de l'information.

Cette schématisation simple permet d'isoler des paramètres contextuels qui vont définir le cadre de l'information. Nous les présentons maintenant

Le paramètre individuel : la patient est caractérisé comme une pure valeur classée dans le type [x : PATIENT]. Les noms propres identifient les individus et cette désignation est hétérogène du statut d'être un patient. N'importe quel individu est imprédictible dans son comportement ; cela constitue un corolaire du statut d'être un individu et constitue une connaissance spécifique à ce monde. Elle permet d'inférer le fait que quiconque peut apprendre à propos de chacun des individus et que l'information a comme rôle de permettre la représentation de ces phénomènes impossibles à prédire. La notion d'individu est définie par P.F. Strawson.

Durée et mémoire sont les rôles et fonctions des feuilles et de la base de données. Chaque unité symbolique qui est inscrite sur la feuille est la duplication d'un « original », réalisé dans un lieu déterminé de l'activité. L'espace de la feuille est celui dans lequel l'interprétation s'insère. (Par exemple, « M. Smith" est un nom propre dupliqué depuis un lit dans l'hôpital vers la collection de feuilles dans l'espace de la pharmacie). La différence est que la dernière réalisation de l'item lexical est intégrée dans une collection et y est archivée.

Information, système d'information et opérations. L'ensemble de la trajectoire de l'information depuis les outils d'analyse et de mesure du patient vers les feuilles dans l'espace de la pharmacie constitue un système d'information. Il est régulier et donc prédictible. Il est organisé sous forme d'arborescence (non hiérarchique, mais probabiliste) : chaque bactérie et molécule est organisée suivant un choix de paramètres spécifiques et chaque paramètre choisi entraîne le choix d'une échelle de mesure particulière spécifique.

Le processus de communication. Il dépasse les limites du flux. Si la source d'information est ici le patient, l'interprétation constitue alors une nouvelle source d'information vers d'autres acteurs : les personnels soignant et les prescripteurs. Le processus de communication repose sur l'interaction entre les différents acteurs du soin.

Schématisation des opérations dans le monde.
La première schématisation était contextuelle. Elle permettait de repérer les paramètres du temps et de l'espace et de les lier à l'inscription des unités lexicales.
Il est possible de préciser cette schématisation en introduisant l'identification d'opérations. Ainsi, nous disposons d'une représentation de la trajectoire de la production de l'information et des opérations externalisées liées à ces opérations symboliques et d'une caractérisation de la production de la structure d'information.

On décrit ainsi les différentes étapes de la production de l'information. Elles sont localisées dans leur environnement propre. On introduit également quelques éléments de l'assignation de référence. Cette schématisation spécifie la précédente en caractérisant plus précisément la façon dont les symboles sont successivement produits. On adjoint dans cette schématisation une indication de la localisation de la référence de chacun des symboles transmis.

⇢ : représente la référence.
→ : représente une opération.

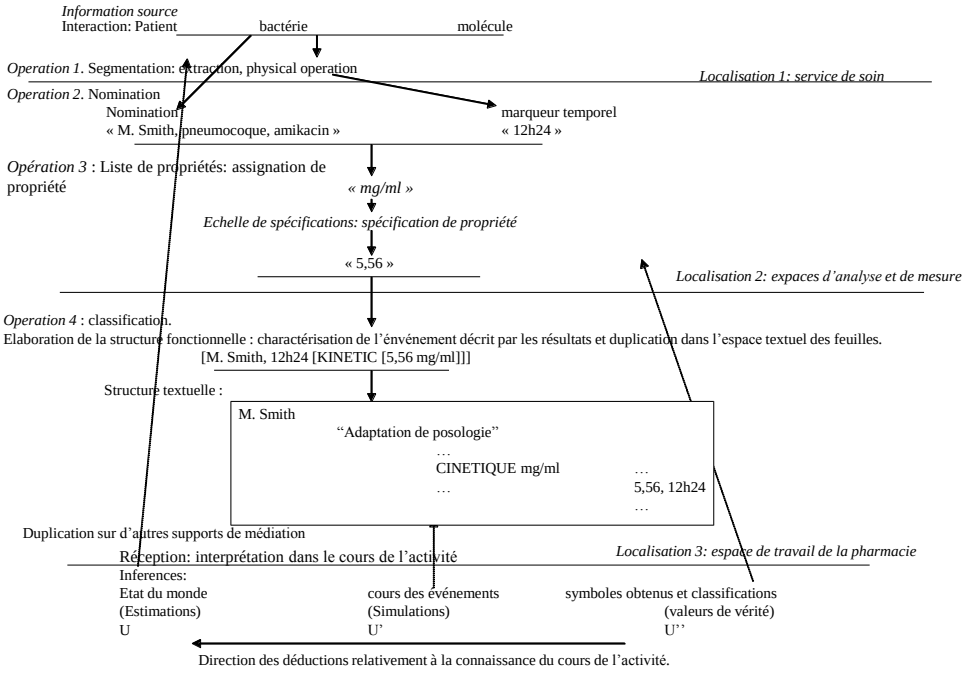


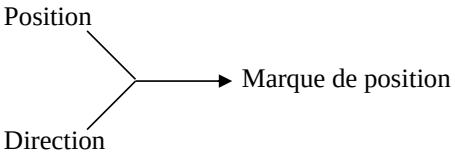
Figure 6. Schématisation du processus informationnel dans l'activité d'adaptation de posologie.

Cette schématisation permet de formuler l'hypothèse de trois mondes[202] de référence différents :
U : monde de la référence directe, limité par le patient. C'est le monde extensionnel de l'individu.
U' : monde des événements : il s'agit du monde intensionnel des événements de pharmacocinétique, autrement dit celui de l'élimination des molécules par le corps. Ce monde est intensionnel.
U'' : monde à la fois intensionnel et extensionnel des mesures et analyses ; il concerne toutes les analyses produites à propos de ce patient.
Nous reparlerons de ces mondes lorsque l'on aura intégré les paramètres de l'activité dans notre caractérisation des flux. Notons néanmoins qu'ils définissent une segmentation de l'univers suivant l'ontologie de K. Popper, reformulée dans le cadre des sciences de l'information par P. Ingwersen & K. Ja¨rvelin[138], pp. 48-9)[203].
Evidemment, nous reformulons cette ontologie, notamment parce que nous adoptons une démarche descriptive, mais qui va aboutir à une reformulation des présupposés relativement à la cognition qu'utilise K. Popper et P. Ingwersen & K. Ja¨rvelin. La véritable distinction est relative au caractère distribué de cette cognition et des conséquences que cela a sur la façon dont on peut envisager le rapport au monde et à la connaissance.

Les opérations identifiées dans le cadre de cette schématisation ne sont pas décrites avec un niveau de détail suffisant pour représenter comment une information se construit et est transférée. Nous allons donc présenter d'abord la schématisation de ces opérations, qui constituent les armatures des situations. En effet, ces opérations jouent un rôle fondamental dans la symbolisation et donc dans la construction de l'information.

On peut décrire de la façon simple les opérations en considérant, à la suite de G. Hutchins, qu'une opération se résume en l'apposition d'un paramètre à un objet. Le résultat de l'opération se caractérise toujours par l'apposition d'un symbole à côté du ou des symboles déjà obtenus.

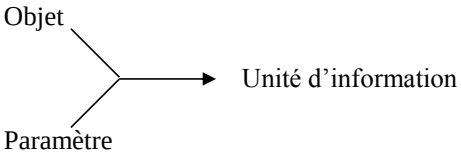
Avant de représenter de façon schématique les opérations, il convient de présenter la façon dont on adapte les propositions de G. Hutchins à notre contexte :



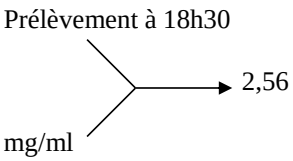
L'exemple de G. HUTCHINS est caractérisé dans la navigation : en fonction d'un lieu dans lequel je me situe (par rapport au soleil, aux étoiles), et de la direction que je suis (d'où je viens), je détermine exactement où je me situe dans ma trajectoire.

[202] Nous préférons utiliser le terme de monde plutôt que celui d'univers parce que si les espaces dont il est question ici sont des constructions humaines, ils ont une dimension empirique, donc extensionnnelle, qui ne peut être aisément représentée par la notion d'univers.
[203] L'ontologie de K. Popper sert à P. Ingwersen and K. Ja¨rvelin à formuler un modèle de la recherche d'information. Nous pourrons discuter plus loin des modifications qu'entraîne un changement de cadre cognitif sur les modèles de la recherche d'information.

On peut généraliser la schématisation de la façon suivante :

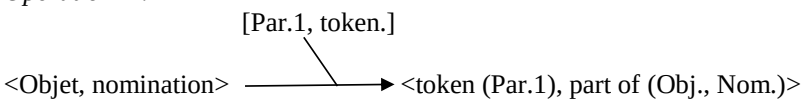


A propos d'une opération particulière, on obtient la représentation suivante :



Maintenant, en considérant l'ensemble de l'activité, et donc la succession des opérations dans le temps, on fera apparaître la schématisation suivante :

Opération 1 :



Etat représentationnel: [nomination, token (Par.1)] *« M. Smith, 12h34 »*

Cette opération a ici une forme particulière, liée au fait que l'objet de la seconde classification constitue une partie (un prélèvement généralement) du premier objet. Ce second objet est dénommé par le moment où il a été isolé du reste du corps du patient.
On postule que les entités composant les opérations existent sous forme d'une dualité entre objet du monde et entité de la pensée. En effet, une opération étant à la fois un assemblage matériel et l'externalisation d'une forme de pensée, elle ne peut être représentée que sous ce double aspect. On peut donner de cette dualité une double explication : elle peut être de nature cognitive, et nous renvoyons pour cela à la partie 5.3.2.. Elle peut également être de nature logique, et caractériser un rapport <token, TYPE>, donc une classification dans le cadre des théories de type.
Les unités composant les opérations sont considérées sous cette double dimension, qui caractérise la dualité entre entité matérielle et cognitive, entre type et token et entre concept ou propriété et réalisation de ces derniers dans des faits du monde.

Les paires <,> caractérisent des états satisfaisant les contraintes du couplage cognitif ; les [,] représentent des opérateurs et les () des états traités antérieurement. Les flèches représentent des inférences fonctionnelles. Les opérations constituent des unités cognitives localisées. Leurs résultats sont exprimés sous forme d'états représentationnels.

La première opération se limite à l'apposition d'une marque temporelle au nom propre. Matériellement, il peut s'agir d'un prélèvement.

Operation 2 :

[Par.2, Token.]

< Part of (Obj., Nom.), token.(Par.1)> ⟶ <token. (Par.2), <token. Par.1, part of (Obj.,Nom.)>>

Etat représentationnel: [Nomination, Token. (Par.1), Token. (Par.2)] *« M. Smith, 12h34, mg/ml »*

Cette opération correspond à la classification du premier objet (et non sa dénomination). On peut s'interroger sur le fait que l'on identifie pour chaque classification une opération distincte ; on identifie une opération de sélection pour chaque classification, conforme au modèle de l'opération. Ces opérations sont synthétisées dans le cadre de la structure d'information, où ces différentes opérations sont considérées comme une seule tâche.
Elle concerne le choix d'un outil concernant le phénomène à mesurer. Il décrit une propriété mesurable à partir du résultat matériel de la première opération.

Operation 3 :

[Par.3, Token.]

< token. (Par.2), part of (Obj.,Nom.)> ⟶ <Token. (Par. 3), token. (Par.2)>, <token. (Par.1), part of (Obj.,Nom.)>>

Etat représentationnel : [Nomination, Token. (Par.1), Token. (Par.2), Token. (Par. 3)] *« M. Smith, 12h34, mg/ml, 5,56 »*

Cette opération correspond à un infomorphisme ; il s'agit de l'obtention du résultat de l'opération appliquée sur le second objet, mais relativement à la contrainte fixée par la première classification.
Cette opération concerne le choix d'une valeur dans l'échelle introduite par l'unité de mesure et permet de spécifier la propriété attestée par le résultat de la précédente opération.

Operation 4 (duplication):

"M. Smith, Kinetic" & "Kinetic, mg/ml" [Paramètre sur la feuille]

[Espace d'opération] [sur la feuille]
"M. Smith, 12h34, mg/ml, 5,56" ⟶ *"M. Smith, 12h34, Kinetic, mg/ml, 5,56"*

La quatrième opération est la plus générale et subsume l'ensemble des autres. Elle caractérise le phénomène matériel que l'on cherche à observer et détermine le choix du paramètre de l'opération 2. C'est la dénomination de cette opération qui opère la classification de la structure d'information sur les feuilles.

Ces opérations constituent des cadres pour l'information, mais ne constituent pas en propre des informations au sens où elles permettent seulement de caractériser une symbolisation. L'information est considérée à partir du moment où on met en œuvre un processus de structuration de ces unités symboliques. A ce moment-là, les données s'inscrivent dans une structure d'information.
Les opérations forment la syntaxe de la structuration de l'information par les relations de succession entre les composants. Ces différentes opérations sont structurées de façon à ce que

leur résultat puisse être transmis. C'est ce processus que l'on représente à l'aide des flux. La transcription de la représentation depuis les opérations vers les flux peut se représenter de la façon suivante :

La construction de l'objet comme token: <Object, nomination> : a.A
Construction de la fonction au niveau des tokens : <Part of (obj., nom.), token. (Par. 1)>: a\b.B
Construction du niveau des types : (Par.2, token. (Par. 2)), (Par.3, token. (Par. 3)): $\alpha \in \Gamma$, $\beta \in \Delta$
Traduction des opérations :
Relations de classification entre symboles : $<\vDash, a, \alpha>$.
Relations fonctionnelles entre classifications : paire contra-variante de fonctions :
$< f^{\wedge}, f^{\vee}>$.

Les opérations peuvent maintenant être ordonnées de façon à représenter un infomorphisme:

- Un type $\alpha \in \Gamma$ peut être traduit dans un type plus précis $\beta \in \Delta$ sous la condition que α classifie *a*, donc β classe *b*. (Sit: situation)).

$$\begin{array}{ccccc} \alpha\in\Gamma & — & f^{\wedge} & \rightarrow & \beta\in\Delta \\ s_1 \vDash \ | & & & & s_2 \vDash \ | \\ a.A & \leftarrow & f^{\vee} & — & b.B \end{array}$$

Cet informorphisme et le transfert de l'information vers le cadre de réception permettent la représentation du schéma général de flux suivant (similaire à celui que nous avons présenté plus haut, et pour lequel inst. (ou instance) est remplacé par token, sachant que les token constituent des instances) :

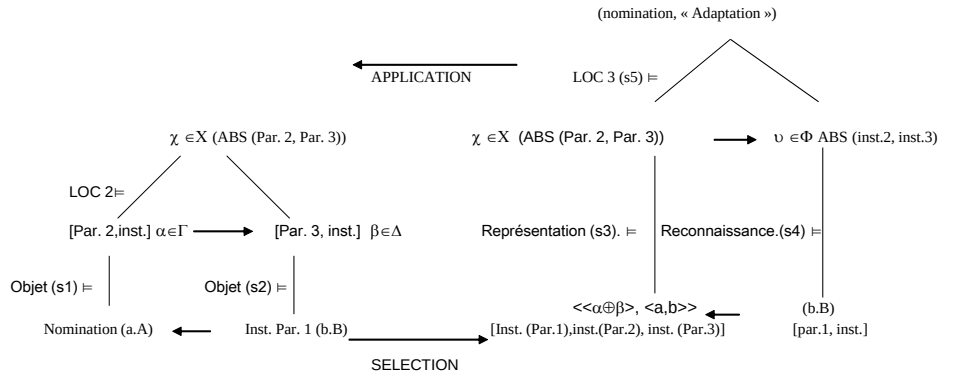


Figure 7. Représentation des flux intégrant les opérations dans le cadre de l'adaptation de posologie.

Nous pouvons préciser maintenant les situations d'interprétation, contenues dans l'espace de réception de l'information.
Situation d'interprétation en intension : l'expression "*M. Smith, 12h34, Kinetic, mg/ml, 5,56"* est classé dans la représentation intensionnelle du patient en s_3 :
$s_3 \vDash <<<\alpha\oplus\beta>, <a,b>> , < \chi,<\Gamma,\Delta>>>$.

Situation d'interprétation en extension : le concept $\upsilon \in \Phi$ (où Φ: spécification des concepts de X) est une interprétation au niveau des tokens (s_4):
$s_4 \vDash$ <<b>, < a,<$\alpha \oplus \beta$>, υ>>.
Un type χ of X est fonctionnellement spécifié dans un type υ de Φ .

Contrôle : flux et transition : déroulement de l'activité.
Nous avions présenté plus haut les logiques de transition comme des outils permettant de valider les inférences produites par les flux, en attestant tout simplement d'une transformation d'état opérée dans le contexte.

Or, pour cela, il faudrait proposer l'analyse de la totalité des informations contenues sur les feuilles et relatives au même objet. On disposerait ainsi de données suffisantes pour caractériser les changements d'états. Il est à noter que cette question est abordée à l'intérieur de la théorie des flux par la modélisation des états et des changements d'états. Très généralement, cette modélisation repose sur les mêmes principes classificatoires que les flux, mais substitue une paire covariante de fonctions à la paire contra variante. Ainsi, on rend compte de la succession des états, mais pas de l'information et de sa circulation. En définitive, les changements d'états représentent la conséquence des flux sur les représentations du monde.

Nous expliquons maintenant les systèmes, qui constituent le concept permettant de verrouiller la structure, de la limiter et de la circonscrire.

Rôle des schématisations.
Les schématisations précédentes ont comme rôle principal de construire un socle d'analyse permettant de montrer l'ensemble des phénomènes que les flux sont capables de représenter, mais également les limites du modèle. Ce cadre d'observation servira ensuite d'outil pour mettre en œuvre le modèle dans toute situation conforme à celle-ci, à savoir où l'on retrouve une production d'information fondée sur des opérations, un transfert d'information et l'intégration de cette information dans un cadre d'activité.
Pour cela, il suffit de remplacer par des variables les unités symboliques attestées dans le cadre de la pharmacie.

4.4. Pertinence des flux.

La précédente description des conditions d'observation et de modélisation des flux sert à montrer l'ampleur et les limites de l'application des flux à des situations empiriques. Cette description vise en partie à étendre les situations d'applications des flux. Dans le cadre du web sémantique, les flux s'intègrent dans les problématiques d'alignement de données structurées et d'ontologie, et leurs usages pourraient être nettement plus étendus.
En effet, la théorie des flux se présente comme une théorie mathématique, susceptible de rendre compte de systèmes fonctionnant à partir de données hétérogènes, expliquant les possibles traductions entre ces données. Ce programme très général dépasse largement les contraintes d'un système, d'un langage ou d'un outil de structuration. Le premier travail consistera à caractériser l'usage des flux dans le contexte du web sémantique. On affinera progressivement les choix d'outil ou de modèle dans le web de données dans le cours de la présentation.

Le programme très général de flux demande quelques précisions, entre autre parce qu'il mêle des questions de connaissances et de signification. Autrement dit, si le raisonnement constitutif des flux a une portée limitée, les cadres d'application sont beaucoup plus larges : ils concernent tous les domaines dans lesquels il est nécessaire de mettre en relation des structures de données hétérogènes de façon à produire une représentation d'état adéquate à une activité.
Par ailleurs, les applications des flux ne concernent jusqu'à présent que des parties relativement restreintes de la théorie. En effet et notamment, les questions d'alignement d'ontologies ne prennent pas du tout en compte la dimension interprétative. Comment peut-on alors caractériser la dimension sémantique des flux, sachant que l'on n'a pas ici seulement un cadre linguistique, mais bien un transfert d'information ? On formulera ici quelques hypothèses relativement à la pertinence du modèle, avant d'entrer, dans la partie 5 (consacrée à l'analyse des structures d'information et à la sémantique), dans les arguments pour une formulation étendue des flux.
Chacune de ces trois représentations peut être déduite de l'autre :

- la première schématisation est la plus analogique qui soit relativement à la situation que l'on cherche à décrire ; elle ne caractérise les flux que comme structure d'inférence entre deux classifications.
- la schématisation de la succession des états est dotée d'une sémantique dans le deuxième schéma ; elle inclut la représentation du transfert de l'information depuis une localisation vers une autre.
- la troisième représentation caractérise les relations de dépendance qu'il est possible d'identifier, dans une expression, et qui seraient les traces des flux.

Enfin, on peut expliquer ici les difficultés qu'il y a à définir l'information : une seule dimension (qu'elle soit mathématique ou linguistique) ne peut rendre compte de l'information. Les deux dernières représentations intègrent la dimension de la communication, et donc introduit d'autres paramètres que ceux qui concernent seulement l'élaboration de l'information. Ces nouveaux paramètres intègrent les supports et les conditions de réception.

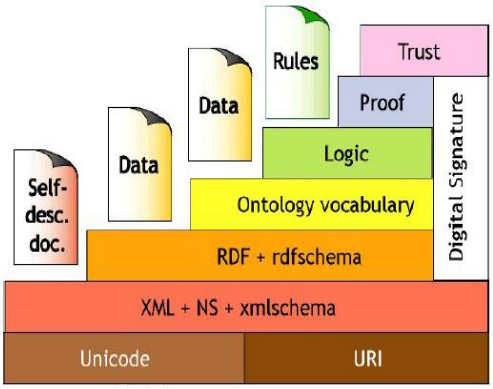
4.4.1. Sur quelques outils proches : problématiques du web sémantique.

L'idée de proposer un modèle du contexte pour ensuite caractériser les contenus ou les connaissances transmises ne constitue pas une idée nouvelle. Il existe en IA[139] comme en linguistique de nombreux travaux sur le sujet, liés autant à l'explication causale du comportement de certains agents qu'à des paramètres de spécification de contenus ou de signification.
Or, la question du contexte est posée différemment dans le cadre des flux : comment un canal d'information transmettant régulièrement des messages permet-il d'associer et d'inférer des états ? En d'autres termes, peut-on construire des relations signifiantes à l'aide de contraintes contextuelle associées à des structures, considérant les connexions entre structures de données signifiantes que ces canaux rendent possibles ?
Ces questions amènent à traiter la question des flux non comme phénomène périphérique, caractérisant des contraintes sur l'enchaînement des propositions constituant une structure, mais bien comme un modèle représentant des processus amenant à mettre en relation des structures de données hétérogènes dans le cadre de la structuration progressive du web. Si en premier lieu on peut penser à des structures comme les ontologies, les thésaurus et les terminologies, les propositions que l'on peut faire à propos de la structure d'information orientent l'analyse vers l'analyse des expressions linguistiques proprement dites.

On aimerait présenter maintenant le domaine le plus immédiat d'application des flux, à savoir le web sémantique, par rapport auquel on aimerait situer la problématique des flux d'information. En effet, le web constitue un espace en structuration dans lequel les propriétés des flux peuvent apporter une aide à certains niveaux d'abstraction. Ainsi, on limitera aussi la portée de la théorie.
On présentera d'abord les différents niveaux d'abstraction du web, afin de comprendre où peut se situer l'information.

Les différents niveaux d'abstraction du web sémantique.
Historiquement, la structuration du web suit une trajectoire qui va de règles syntaxiques simples de structuration des données (comme le HTML), vers des outils prenant de plus en plus en charge la sémantique des données : XML, puis RDF enfin OWL. Ce dernier niveau établit un lien direct entre le web et les outils de représentation de connaissances, nettement antérieurs, comme les réseaux sémantiques par exemple.

La schématisation qui suit représente une première structuration du web et de l'empilement des différents niveaux d'organisation.



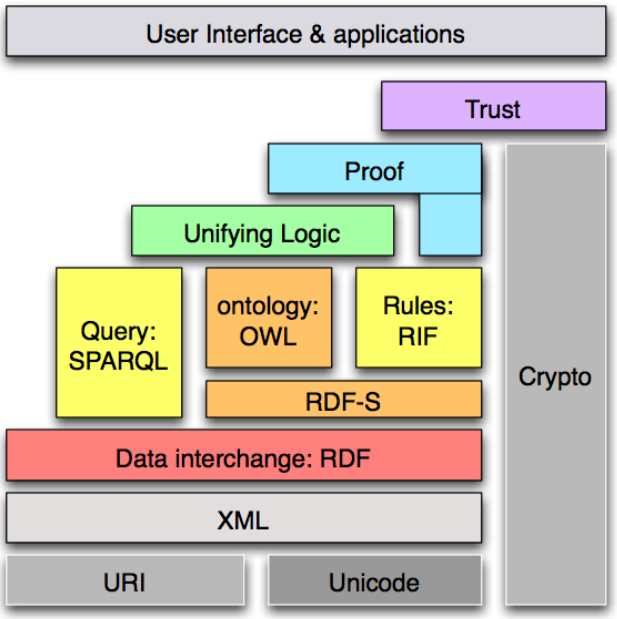
[204]

Néanmoins, ce premier modèle pose problème.

- La couche « ontologique » possède une alternative : soit OWL, soit un langage à base de règles (comme DATALOG par exemple[140])
- Il existe des logiques descriptives qui constituent les relations entre les ontologies et la logique.
- L'architecture du web de données est encore en discussion.

On pourra préférer ce modèle, plus détaillé et surtout qui rend compte de la spécificité d'usage de chacun des différents outils du web et de leurs relations.

[204] http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html



205

XML demeure la syntaxe de base de toute structuration du web (encore que l'adoption de TURTLE comme syntaxe simplifiée pour l'écriture des triplets RDF peut être vue comme un détachement par rapport aux règles d'écriture fixées par XML[206]. La couche RDF caractérise le modèle de données pour caractériser les faits, et RDFSchema constitue un langage d'ontologies simple. SPARQL constitue le langage d'interrogation adapté au web de données. C'est donc cette couche qui apparaît la plus pertinente pour accueillir les questions d'information.

La couche ontologie est peuplée par un langage plus expressif que les schémas RDF. Il s'agit du standard courant du web OWL. Néanmoins, un langage de règles, utilisant une théorie de types, permet de se passer d'OWL.

La couche logique augmente les capacités des langages d'ontologie. On entendra alors non seulement les logiques descriptives, mais l'ensemble des logiques permettant de contrôler les raisonnements mis en œuvre dans les niveaux moins abstraits. Les logiques jouent alors le même rôle que les logiques de transition par rapport aux flux.

La couche preuve sert à la génération de preuves, à l'échange et à la validation. La couche « trust » (ou « gouvernance ») contient les signatures digitales, ratifications, etc.

Dans ce cadre général, la question d'un niveau spécifique de l'information constitue un des enjeux fondamentaux du web de données : les relations entre entités sont gouvernées et justifiées par une sémantique, quel que soit par ailleurs le niveau d'abstraction de cette sémantique : terminologie, sémantique lexicale, ontologie. Or, l'information ne constitue pas

[205] http://www.w3.org/2009/Talks/0120-campus-party-tbl/#(1)
[206] http://www.w3.org/TeamSubmission/turtle/

l'objectif spécifique d'une recommandation, si ce n'est une référence très générale concernant RDF.
Dans un tel contexte de formats de représentation universels et neutres par rapports à des théories du langage, de la connaissance et de l'information, les projets sont très nombreux ; ils vont de la représentation ontologique de domaines à la construction de structure permettant de mettre en relation et de mettre à profit les relations entre des données structures. Par exemple l'Open Knowledge Project propose un certain nombre de modèles d'interaction qui constituent des contextes de transmission de connaissances[141]. Bien évidemment, nous ne pouvant en faire une liste exhaustive. Néanmoins, on peut positionner la question des flux à l'intérieur de thématiques propres au web de données.

Il existe un nombre important de travaux relatifs à la caractérisation d'inférences entre des données structurées hétérogènes : ce sera le cas plus particulièrement pour l'intégration sémantique[142], où il sera question de la mise en relation d'ontologies et de la distribution des sources d'information[143]. Toute relation établie entre structures de connaissances est donc située par rapport à un modèle de la traduction. Celui-ci peut être géré dans un outil de production d'ontologies comme DOLCE[144].
Néanmoins, ces projets n'intègrent pas la modélisation du passage d'un niveau d'abstraction vers un autre, à savoir les classifications, ni la façon dont on peut lier des structures de données, considérant leur hétérogénéité, au travers notamment des canaux. La caractérisation de la signification et la distinction entre langage naturel et langage normé, la distinction entre connaissances et termes constituent un autre enjeu de l'élaboration du web de données, dans lequel les flux ont leur pertinence.
Enfin, il n'existe pas de niveau spécifique pour caractériser les unités linguistiques dans le web de données. Cette absence constitue un problème important, pour lequel les flux peuvent proposer quelques solutions.

Nous situerons l'enjeu des flux du côté de RDF, parce que les flux caractérisent des relations entre instances et types, donc se situent à la charnière des entités de faible niveau d'abstraction et des structures de haut niveau comme les ontologies.

Caractérisation de plusieurs niveaux d'abstraction pour les flux : depuis les instances vers les mises en relation d'ontologies.
Les standards du web constituent des outils de représentation ayant des propriétés syntaxiques internes au web. Pour nous, le problème est légèrement plus complexe puisque si le contexte influe radicalement sur la sémantique de l'expression transmise, il est aussi associé à la référence. En effet, dans nos exemples, les opérations (comme les mesures de durée ou les opérations de mesure) tout comme les entités matérielles du monde (comme le corps du patient notamment) jouent un rôle fondamental dans la définition du flux. Ainsi, si l'on utilise un standard, c'est relativement à une problématique plus large, dans laquelle les langages du web constituent des outils de formulation.

Les flux semblent relativement proches des ontologies (telles que définies par B. Smith). En effet, un lien s'opère entre instance et type, des relations spécifiées sont établies entre les concepts et il est question de référence au monde réel. La différence essentielle tient d'une part dans l'absence de construction d'un domaine par les flux (puisqu'ils dont le lien entre des structures appartenant à des domaines pouvant être distincts), et d'autre part dans l'absence de dénomination des liens spécifiés entre les entités (les relations sont marquées dans les flux par des inférences et non par des relations dénommées par un concept spécifique). Ce sont les

raisons pour lesquelles les flux caractérisent des relations entre des structures existantes et ne permettent pas la structuration d'un domaine, donc l'élaboration d'une ontologie.
Cette distinction permet de situer les enjeux des flux dans le cadre du web de données ; il s'agit d'un outil permettant de lier des structures représentant elles des domaines de connaissances. En eux-mêmes, les flux apparaîtront comme un outil générique, à savoir une méthodologie susceptible d'être utilisée pour des domaines et des circonstances relativement différentes.

En ce sens également, les flux se distinguent nettement des outils de caractérisation des relations lexicales[145], notamment MARTIF et les outils de la TEI. Ils opèrent seulement sur des structures linguistiques et ne font pas le lien vers des niveaux d'abstraction plus élevés. Ce sera d'ailleurs un enjeu fondamental des flux que de rendre compte de la conversion entre ces différents niveaux.

Enfin, mis à part les propositions fondées sur les théories de l'interaction, il n'existe pas de théorie générale des flux autre que celle de B & S. Les modèles de cognition située et distribuée (P. Agre[146] notamment) n'intègrent pas la dimension technique des réseaux. Nous reviendrons sur cette dimension technique dans la dernière partie de ce travail, celle qui concerne la mise en œuvre d'applications.

4.4.2. Flux et web de données

Globalement, les flux seront considérés comme un modèle permettant de représenter des phénomènes qui utiliseront les langages de représentation du web. C'est en ce sens que les flux, en tant qu'outil mathématique de représentation d'un raisonnement, servent d'outil logique pour la caractérisation de relations entre données structurées. Les langages de représentation constituent des outils et les flux, en tant que modèle, assure la conformité des représentations à la portée de la logique.
Cette construction permet de réaliser l'articulation entre connaissances, métadonnées et structure d'information. Dans ce cadre, la structure d'information constitue une traduction des flux à un plus bas niveau d'abstraction que les métadonnées. C'est aussi une expansion du modèle des flux dans le cadre du traitement d'expressions linguistiques.

Après avoir présenté en quoi les flux pouvaient servir dans la cohérence générale du travail, on pourra ensuite les situer dans les modèles de description documentaire afin d'expliquer en quoi la problématique des métadonnées est enrichie par celle des flux : en distinguant clairement les tokens, les valeurs et les types, on intègre une dimension sémantique structurante des entités composant les métadonnées. Nous présenterons les dimensions techniques, mais également organisationnelles des langages du web dans la dernière partie de ce travail, en vue de l'ancrer dans une perspective applicative. Nous nous limiterons ici aux dimensions scientifiques de ces langages et outils.

On pourra néanmoins ici très rapidement identifier trois types d'entités qui devront être distinguées :

- Les langages du web, qui constituent des outils d'écritures, présentés sous forme de recommandation, mais qui offrent des possibilités étendues d'expression.
- Les métadonnées, qui utilisent les langages précédents pour leur écriture, et qui constituent des vocabulaires communautaires (au sens de communautés professionnelles) permettant la description des documents,

- Enfin, les outils, utilisant également les langages et pouvant être associés à certaines métadonnées, qui proposent des fonctionnalités au travers de schémas ou de DTD déposés.

Notre travail, outre les propositions que l'on formule, pose comme question celle de la nature même des langages du web et de leurs possibilités expressives. S'agit-il de langages artificiels clos, ou au contraire, parce qu'ils possèdent la propriété de prédication, ont-ils les mêmes capacités d'infinitude que les langues naturelles ?

Nous proposons de caractériser les flux en montrant ce qu'ils ont jusqu'à présent apporté, à savoir les applications liées à l'alignement d'ontologies. Par ailleurs, les flux contribuent largement au travail philosophique relatif à la définition de l'information. On se limite donc ici aux aspects les plus proches des objets concrets des sciences de l'information.

Utilisation des flux pour la mise en relation d'ontologies (considérées comme des structures de données).

Les flux ont donné lieu à deux travaux majeurs concernant l'application des flux au web de données. Cela peut sembler relativement peu pour une telle théorie. Néanmoins, outre son accès difficile, il existe des modèles répandus comme la FCA (Formal Concept Analysis) et les logiques de description, qui proposent de caractériser des raisonnements sur deux niveaux de description[147] et à partir de trois sortes d'objets. Ces objets sont les individus, les concepts et les propriétés associées. Les deux niveaux de description sont les domaines de connaissance et les connaissances factuelles.

Ces deux principales références théoriques pour modéliser les structures et les relations entre elles (y compris la distribution) ne caractérisent pas globalement la question de l'information et la dynamique qui peut y être associée.

L'application des flux dans le cadre du web de données s'est essentiellement manifestée autour de l'alignement d'ontologies. Les ontologies de domaine étant des structures de connaissances homogènes, leur alignement, à savoir la possibilité d'inférences de l'une dans un autre, est apparu comme un domaine d'application particulièrement pertinent pour les flux.

Ainsi, les flux se caractérisent comme un outil descriptif de haut niveau d'abstraction, permettant de relier des structures de données comme par exemple les ontologies. R. Kent[148] d'une part et Yannis Kalfoglou et Marco Schorlemmer d'autre part, se situent dans le cadre de l'alignement d'ontologies. Ce qui retient d'abord leur intérêt, c'est la capacité des flux à appréhender et mettre en valeur la distribution.

Yannis Kalfoglou et Marco Schorlemmer [149] caractérisent l'intérêt de la théorie des flux de la façon suivante :

"[Information Flow] is based on the understanding that information flow results from regularities in a distributed system, and that it is by virtue of regularities among the connections that information of some components of a system carries information of other components. The more regularities the system has, the more information flows; the more random the system is constituted the less information will be able to flow among its components."

Pour Yannis Kalfoglou et Marco Schorlemmer, un contexte limité et circonscrit est un domaine. Il est alors possible d'y connecter différentes structurations de données et de connaissances comme des ontologies. En effet, le domaine reste le même, même si la langue

ou la structure du vocabulaire (et ce qu'elle est censée représenter) diffèrent. On associe alors le contexte au domaine.
La particularité du travail de R. Kent réside dans l'utilisation de la théorie des types pour représenter un flux d'information. L'idée consiste à caractériser des informorphismes entre des ontologies hétérogènes, sachant que, puisque les ontologies disposent de plusieurs niveaux de représentation, il devient alors possible d'utiliser le modèle.

Ses premiers travaux ont tout d'abord visé à associer aux flux un vocabulaire formel pertinent par rapport aux ontologies. Ils montrent l'articulation des flux à la théorie des types et rendent alors possible l'application des flux à des ontologies alignées. Les travaux menés par M. Schorlemmer et Y.Kalfoglou vont fournir de premières expériences d'application. Celles-ci seront développées ensuite notamment par M. Schorlemmer, qui précisera l'intérêt des flux pour caractériser des raisonnements utilisant des données distribuées dans le cadre d'une situation occurrente, contenant des phénomènes d'interaction.

La perspective proposée par les flux à propos de l'intégration sémantique (à savoir la façon dont on va pouvoir caractériser des relations sémantiques entre des structures hétérogènes mais syntaxiquement interopérables) se précise peu à peu : elle consiste en la prise en compte progressive des propriétés liées au contexte, à l'environnement.

Les propositions d'utilisation des flux dans le cadre de l'intégration sémantique mettent en œuvre des outils logiques complémentaires des flux. M. Schorlemmer et Y.Kalfoglou établissent[150] l'ensemble du cadre logique dans lequel l'intégration sémantique (entre des ontologies hétérogènes) est opérationnelle.

Les travaux de R. Dapoigny[151] montrent qu'à partir d'une logique de but, et plus globalement des logiques dynamiques, il est pertinent d'utiliser les flux pour caractériser des changements. Ces applications restent, comme A. Zimmermann & alii.[152] l'ont montré, encore relativement théoriques et n'ont pas jusqu'à présent été appliquées à grande échelle.

Utilisations des flux en philosophie de l'information.
Nous venons d'évoquer les applications les plus immédiates des flux dans le cadre du web de données. Néanmoins, pour bien comprendre les enjeux des flux, il est important de les situer par rapport à d'autres domaines d'application, relativement au raisonnement impliquant des formes visuelles, des diagrammes et autres représentations analogiques.

L'idée fondamentale est que les raisonnements à fondements visuels reposent sur la juxtaposition d'objets hétérogènes, de dimensions et de systèmes sémiotiques différents. La proposition de J. Barwise est par définition prédicative, et vise à représenter la trajectoire de la production d'une information. Cette représentation se distingue de la méréologie (qui est une théorie topologique considérant que le raisonnement se représente par des marqueurs spatiaux). Elle se distingue également des théories propositionnelles présentées notamment par J. Van Benthem (pp.71-74)[153] qui elles caractérisent des raisonnements dans l'espace comme des identifications de similitudes et de différences entre objets.

La théorie de J. Barwise & G. Allwein à propos des diagrammes intègre un certain nombre d'éléments déjà présents chez H. Kamp et qui ont trait à l'articulation entre les marqueurs pragmatiques et sémantiques, à savoir les instructions (voir plus loin P. Blakburn et alii).

L'hypothèse centrale est que les formes visuelles peuvent être autant que les expressions symboliques des éléments de preuves mathématiques : les diagrammes peuvent servir d'outils de calcul, et donc il est possible de les inscrire dans les mécanismes de déduction.

En considérant les déductions comme l'extraction et l'explicitation d'informations implicites dans les informations données, on valide généralement les raisonnements par des déductions phrase à phrase. En particulier, on considère les preuves déductives comme les structures extraites de ces phrases en utilisant certaines règles formelles.

L'idée centrale consiste à considérer que les raisonnements peuvent utiliser des données non linguistiques. Les représentations linguistiques ou visuelles d'une même information n'ont pas les mêmes propriétés (l'indication d'une direction en utilisant une carte). Certaines représentations sont plus efficaces que d'autres dans certains cas. Les raisonnements concilient les différentes formes de représentation de l'information. Ce qui est intéressant dans ce cadre, c'est le fait que l'on puisse caractériser le raisonnement en utilisant à la fois des marques symboliques et visuelles.

En fait, l'hétérogénéité repose sur le principe que des informations visuelles et symboliques constituent la majorité des raisonnements. (J. Barwise & G. Allwein, p. 23 et sq.). La question qui se pose alors est double : pourquoi les entités visuelles sont-elles à ce point pertinentes, et comment décrire les entités visuelles dans le cadre d'un raisonnement ?
(J. Barwise & G. Allwein prennent comme point de départ une démarche similaire à celle de Frege, mais en distinguant bien le fait que l'on ne cherche plus à représenter le raisonnement dans les expressions symboliques, mais dans l'ensemble des données observables).

4.4.3. Dimensions méthodologiques des flux. Niveau (ou couche) logique pour l'écriture des relations entre structures de données.

Nous avons présenté plus haut des propositions relatives aux niveaux d'abstraction et aux logiques de transition. Ces outils servent à contrôler l'utilisation de certains résultats scientifiques depuis un domaine vers un autre.
On peut considérer alors que les flux ont une utilité similaire, mais dans le cadre de la mise en relation d'analyses et de descriptions documentaires hétérogènes : en effet, lexiques, thésaurus et terminologies se diversifient, entre autre du fait de la prise en compte des besoins propres des utilisateurs comme de la multiplication et de la meilleure accessibilité des sources. Par ailleurs, toute généralisation de ces outils entraîne, par la standardisation, de nouvelles règles et possibilités expressives. (Nous l'avons vu avec les travaux de Y. Kalfloglou et R. Kent). Nous aimerions maintenant montrer en quoi les applications des flux ne se limitent pas à l'alignement d'ontologie, mais plus généralement, à la caractérisation de raisonnement à partir de données représentées à l'intérieur des langages du web.

Flux et langages de représentation.
Les flux constituent des outils logiques permettant de valider des relations établies dans le cadre des langages fondés sur les recommandations du W3C, et notamment RDF[154], OWL et surtout SKOS[155], Ce dernier, fondé sur les acquis des précédents outils de structuration, permet de relier des thésaurus entre eux. Or, ces outils ne reposent pas sur une théorie des phénomènes qu'ils représentent, et donc ne peuvent contrôler la validité de ce qu'il est possible de représenter. Ils sont élaborés de façon à apparaitre comme neutres théoriquement.

En effet, de par leur capacité à expliciter la signification de termes (appartenant à des structures distinctes) par leurs relations à d'autres unités d'autres structures, les flux spécifient des usages ; ils accroissent les possibilités descriptives de par les liens qui sont établis entre unités ayant des similarités tout en étant d'autres domaines. Ainsi les flux peuvent contrôler les connexions entre ces différents outils.

A partir de XML, mais surtout RDF et SKOS, on dispose d'outils qui par des assemblages prédicatifs binaires puis ternaires, permettent d'exprimer des relations. On peut prendre comme exemple une relation [auteur-publication]. Cette relation peut être contrôle par une ontologie. Un niveau d'abstraction supérieur régule les prédications. Par contre, on ne sait pas quelle information circule entre ces deux entités, à savoir comment l'une enrichit l'autre. Qu'est-ce que l'[auteur] apprend sur la [publication] et inversement. A l'aide des deux niveaux de structuration des flux, on peut spécifier ces apports. C'est ce à quoi on s'emploie dans notre projet.

Logique et ontologies.

Les flux s'inscrivent alors dans le débat sur le rôle des logiques de description, et plus généralement celui de la logique dans le contrôle des opérations effectuées sur le web. O. Kutz & alii.[156] montrent en quoi les ontologies permettent de poser de façon renouvelées les questions de logique universelle, puisque les ontologies sont nécessairement hétérogènes et doivent pouvoir être conciliées de façon sémantiquement fondée. O. Kutz propose d'utiliser les logiques institutionnelles de J. Goguen[157] pour rendre compte de la pluralité des logiques utilisées dans le cadre de la conception des ontologies. Les logiques institutionnelles sont des outils qui permettent de raisonner à partir de catégories, mais ne disposent pas des mêmes possibilités expressives que les flux : ils ne reposent pas sur une théorie de la signification ni n'intègrent des propriétés essentielles de l'information comme la duplication. Cette question apparait aujourd'hui essentielle, illustrée par exemple par les travaux de A. Gangemi, qui visent à reformuler les outils du web de données sur la base de principes sémiotiques explicites.

En définitive, les flux ne travaillent pas au même niveau que les ontologies, qui constituent des représentations conceptuelles d'un niveau d'abstraction plus élevé que les raisonnements opérant sur l'information. Néanmoins, ces deux types d'outils ont vocation à s'enrichir mutuellement.

En effet, flux et ontologies utilisent de mêmes langages d'expression, qui sont les standards du W3C. Ces langages ne sont pas des ontologies, juste des outils de représentation.

Si l'on suit toujours O. Kutz, les ontologies se définissent d'abord comme la représentation conceptuelle des relations entre les concepts d'un champ spécifique, telle qu'elle forme un consensus entre les acteurs de ce champ. Cette définition des ontologies (comme outils) se distingue de celle de l'Ontologie, comme projet philosophique. Sachant cela, on peut distinguer des ontologies formelles (qui seraient des méta-ontologies) et des ontologies de domaine, qui peuvent être également appelées ontologies terminologiques. On peut enfin ajouter les ontologies de tâches.

Les flux, en ce qu'ils s'appuient sur des données structurées, opèrent sur des ontologies. C'est le sens des propositions d'utilisation que nous avons présentées. Le problème se complexifie à partir du moment où les concepteurs d'ontologies se sont posés des problèmes de signification au sens linguistique. De cette question est née également la possibilité de structurer les ontologies en différents niveaux d'abstraction.

Sachant que par ailleurs il n'existe pas d'ontologie universelle mais des vocabulaires de haut niveau d'abstraction utiles à la structuration des ontologies de domaine (BFO, DOLCE), la question est relative aux relations entre ces différents niveaux de représentation.
Cela dit, les concepteurs d'ontologie utilisant plusieurs niveaux d'abstraction ne considèrent pas cette question comme un problème logique. Ainsi, les problèmes logiques sont posés entre des structures de données, et non entre des structures linguistiques et des structurations de connaissances. Les questions de connexion, raffinement et d'intégration conceptuelle sont certes des questions logiques posées aux mises en relation d'ontologies, mais ne sont guère associées aux questions de signification et de relation entre un contenu linguistique et un concept. Or, comme nous l'avons déjà évoqué, les flux ont la capacité de traiter les relations entre des données de niveau différent en raison du recours à une typification.

4.4.4. Flux et structuration de l'information. Enjeux par rapport aux métadonnées.

La pertinence des flux concerne également les dimensions plus traditionnelles des bibliothèques, notamment les fondements des classifications et la complémentarité des différents niveaux de structuration de l'information. En effet, les classifications telles que considérées par les flux ne reposent pas sur une « épistémologie sociale » mais sur un raisonnement explicite dans une certaine situation. Comment peut-on alors corréler différentes classifications et catégorisations de documents et surtout de quelle façon relier les catégorisations liées aux usages à celles qui fondent l'organisation des bibliothèques ?

Les flux reposent sur le traitement des symboles à trois niveaux d'abstraction différents : les tokens, les classes et les canaux. Ils concernent ainsi l'organisation même des bibliothèques : K. Herold[158] considérant les liens entre information et modèles de bibliothèque, place au centre l'idée selon laquelle les bibliothèques constituent le meilleur exemple du continuum entre données-information-connaissance. Encore faut-il que l'on puisse caractériser les relations entre ces différents niveaux de structuration, sachant que les documents sont de plus en plus numérisés, ce qui permet de penser différemment le niveau des tokens : on peut aujourd'hui, par le biais d'annotations, structurer partiellement les contenus.

Cette pertinence des flux s'illustre également dans la recherche d'information, notamment dans le lien entre les différents niveaux de formulation des questions et leurs traductions dans les langages de recherche. Les problèmes de l'hétérogénéité entre les outils de description et de leur traduction dans un langage propre de recherche peuvent également être concernés par un modèle de flux.

On se servira des flux pour construire des modèles permettant de relier et de rendre compatibles différents outils de description de documents (notices, métadonnées, etc.) ou différents points de vue sur les publications. Les flux permettent à la fois d'étager les outils en fonction de leur niveau d'abstraction (métadonnées, représentations de connaissances, ontologies) et de contrôler les processus d'échanges entre ces différents niveaux.

C'est dans le cadre des métadonnées que l'enjeu des flux est le plus important. En effet, les flux privilégient une faible hiérarchisation mais plusieurs niveaux d'abstractions, illustrée par le rapport token-type, face aux taxonomies et classifications, qui elles développent sur un seul niveau d'abstraction des hiérarchies complexes.
Depuis quelques années, les travaux ont porté sur les compatibilités et complémentarités entre les différents jeux de métadonnées et des langages normés (par le biais de profils Dublin Core

par exemple). Cela permet de disposer des descriptions les plus riches possibles. Or, il ne s'agit là que d'une exploitation minimale des possibilités offertes par les métadonnées, surtout si on combine les métadonnées avec d'autres outils d'exploration et de description des textes. Ce sera le rôle de notre projet que de montrer que des descriptions d'un certain document portent de l'information à propos d'un autre document qui lui est empiriquement lié.

Cet inventaire rapide est limité à la mise en relation de structures. Les flux ne peuvent produire eux-mêmes des structures de données, de connaissances, de termes ou de notions sauf en ce qui concerne la formation et l'extraction de structures d'information.

Nous venons d'évoquer la pertinence des flux dans la perspective des outils et processus étudiés par les Sciences de l'Information. Néanmoins, nous n'avons pas encore traité la pertinence du modèle à l'intérieur des problématiques plus générales que sont le langage et l'activité humaine.

Nous avons rapidement présenté les usages des flux. Ils participent à la structuration du niveau logique du web de données. L'objectif consiste à utiliser des données structurées (depuis les lexiques électroniques jusqu'aux ontologies) comme des matériaux sur lesquels on peut caractériser des raisonnements et ainsi accélérer la mise en valeur des outils et des représentations existants. La structuration même que propose R.E. Kent (2011, op. cit. p.1) des différents niveaux des flux répond à cette ambition : « In the IFF, there is a precise boundary between the metalevel and the object level. The modular architecture of the IFF consists of metalevels, namespaces and meta-ontologies. There are four metalevels, Lower, Upper, Top and Ur, corresponding to the set-theoretic distinctions between the "small", the "large", the "very large" and the "generic", respectively.
Each metalevel services the levels below by providing a metalanguage used to declare and axiomatize those levels. Corresponding to the four metalevels are the four nested metalanguages, meta-lower ⊒ meta-upper ⊒ meta-top ⊒ meta-ur, where each metalanguage axiomatization includes specialization of the one immediately above. Within each metalevel, the terminology is partitioned into namespaces, and various namespaces are collected together into meaningful composites called meta-ontologies.
De cette façon, on peut aisément caractériser des relations entre données structurées à différents niveaux du web de données. Les travaux actuels à propos de la théorie des flux, mais également des catégories et des types, vont dans le sens d'une mise en valeur de l'hétérogénéité des données structurées du web de façon à produire des informations et des connaissances nouvelles.

PARTIE 5. Flux, questions de sémantique et de représentation de l'information.

Qu'y a-t-il de commun entre les usages du langage dans la vie quotidienne et la mise en relation de structures de données symboliques (ou autres usages spécialisés, comme en pharmacie hospitalière) ? Lorsque l'on regarde les exemples de résultats d'analyses médicales ou d'états des marchés financiers, ce n'est pas la syntaxe qui apparaît commune avec le langage naturel, mais bien la sémantique. Nous allons donc développer cette dimension des flux, parce qu'elle est centrale pour le web sémantique et pour la relation entre langage et information.

Une information a un sens en considérant les opérations de construction d'information : " M. Smith, kinetic, concentrations, 5,56 µg/l, T0, 12h32 " a un sens si l'on reconnaît le fait qu'une entité marquée dans le temps ("T0, 12h32 "), contient une propriété ("µg/l"), qui est mesurée par une opération dont le résultat est le numérique "5,56", choisi sur une échelle. Ce fait représenté est une partie du processus de cinétique de l'individu (" M. Smith").

On voudrait donc expliciter la portée des flux pour expliquer ce type de réalisation linguistique appelé « structure d'information ». La structure d'information représente un usage régulier des structures linguistiques de façon à satisfaire au mieux la fonction d'information. La structure d'information constitue un outil, et peut donc être raffiné, adapté et spécifié pour réaliser la fonction impartie.
On ne postule pas alors une théorie générale du langage et l'on ne considère pas la structure d'information autrement que comme un outil élaboré dans l'usage du langage et permettant de faire circuler des contenus sans générer d'ambigüités. Une telle position est considérée comme minimaliste. On reprendra la proposition deP. Dekker, qui définit le minimalisme comme la sélection d'un petit nombre d'aspects de l'interprétation du langage naturel (considérés comme particulièrement intéressants). Il utilise la notion minimale de structure d'information (entendue au sens de la sémantique dynamique), pour centrer son travail sur un domaine restreint de phénomènes. La généralisation n'est alors possible que si l'on conserve la logique sous-tendue.

Notre démarche assimile les flux à des contraintes sémantiques sur les structures symboliques ; les questions de syntaxe sont secondaires puisqu'elles dépendent d'autres types de structure, comme l'organisation de l'espace d'affichage. On postule qu'il n'y a pas de déterminisme du langage sur la fonction informationnelle, et donc que les réalisations linguistiques portent des indices des flux d'information lesquels sont directement expliqués par la fonction informationnelle.

Nous commencerons donc par caractériser la dimension sémantique des flux en spécifiant de quelle façon les flux interviennent sur la signification. En effet, comme on l'a vu, les flux se caractérisent par des régularités fixées par le système d'information : les contraintes ne s'appliquent par au niveau du choix des unités, mais à celui des catégories. (Le choix des unités est dû, dans nos exemples, à des calculs ou des mesures externes à la fois aux opérations linguistiques et aux flux d'information).
On pourra alors considérer une sémantique relative aux flux s'appliquant à des catégories linguistiques.

La dimension sémantique des flux amène à une caractérisation de la structure d'information : les liens que l'on peut établir entre les différentes unités lexicales constituantes des flux seront caractérisés comme des relations sémantiques ; ces relations sont contextuelles et sous-déterminées, ce qui signifie qu'elles décrivent une structure particulière inscrite dans le discours. Suivant en cela M.A.K. Halliday, on ne prétend pas représenter la totalité de la signification d'une expression, mais seulement une de ses dimensions.

Si elle est d'abord sémantique, cette caractérisation pourra contenir une dimension syntaxique dès lors que les principes sémantiques auront été établis. Le procédé peut sembler inhabituel puisqu'il est fondé sur une inversion de prémisses par rapport à la démarche linguistique traditionnelle (par laquelle on part toujours d'une description syntaxique de l'unité que l'on étudie, elle-même spécifiée par l'identification des catégories qui composent cette structure). Ici, on part d'une construction contextuelle et fonctionnelle pour envisager la signification.

Nous reprenons ici nos scènes en exemple, incluant à propos de la pharmacie hospitalière la schématisation du cadre des flux ; comment peuvent-elles être modélisées dans le cadre de la détermination d'une sémantique ?
On s'intéresse ici à la façon dont les agents interprètent, et non à la sémantique intrinsèque des unités du langage. (Cette question de l'interprétation est en réalité plus complexe et sera détaillée dans la suite de cet exposé). Le cadre est donc le suivant : l'interprétation d'un résultat d'analyse biologique ou d'une mesure physiologique contient certes l'identification de l'état du monde, mais également la référence à une opération qui s'est déroulée, et qui a permis l'obtention de ce résultat et sa transmission par le biais d'une symbolisation. L'interprétation n'est pas valide seulement dans l'univers qui contient les objets (comme le patient et ses prélèvements), mais également dans celui qui contient les opérations de mesure, les analyses, etc.
Il y a donc, dans cette sémantique, un aspect qui n'est pas pris en compte par les approches linguistiques : celui des mondes dans lesquels l'information est obtenue.

Cette partie a comme vocation la spécification de la structure d'information, à savoir la façon dont on peut inférer, à partir des flux, l'association d'entités symboliques différenciées, et plus précisément hétérogènes. On s'interrogera ensuite sur les raisons pour lesquelles l'information peut être considérée comme initiatrice d'une modélisation plus générale des dynamiques d'activité.
Nous commencerons donc par caractériser la question de la référence, puis nous en déduirons la structure d'information. Enfin, nous essaierons d'élargir le problème aux questions d'activité.

5.1. Référence et inférence.

Le premier problème est celui du niveau de description ; les flux contiennent une dimension sémantique, ne serait-ce que par les relations classificatoires et les inférences. On aimerait montrer que ces relations concernent effectivement des catégories linguistiques et non de seules dimensions conceptuelles. Si les flux sont appliqués à un niveau d'abstraction qui dépasse celui des entités linguistiques usuelles (disons les unités lexicales), il n'en reste pas moins que ce sont bien des catégories linguistiques que l'on observe. On valide la distinction entre une représentation conceptuelle et une représentation sémantique par l'intégration de la dimension référentielle. Toute catégorie et structure décrite doit avoir des propriétés intensionnelles et extensionnelles.

Comment alors peut-on appliquer une problématique sémantique aux expressions que l'on a identifiées dans nos exemples ?

On peut répondre de trois façons à cette question :

- tout d'abord en considérant que la sémantique référentielle ne constitue pas une seule question linguistique, mais qu'elle concerne l'ensemble des entités symboliques, (notamment les expressions diagrammatiques et planaires[159]) ;
- ensuite, que cette sémantique se situe au niveau d'abstraction des flux ; elle est donc distincte de celle relative aux unités, voire aux catégories du langage. On devra donc grouper les entités linguistiques dans des types qui dépassent en généralité les catégories lexicales et grammaticales ;
- enfin, la question de la référence est associée aux flux du fait de leur rôle contraignant par rapport à la fois à la représentation du monde et aux moyens de cette représentation. La définition des univers de référence sera donc déterminée par les flux.

5.1.1. En quoi et pourquoi les flux constituent-ils un ensemble de contraintes utiles pour l'interprétation des expressions informationnelles ?

J. Barwise (op.cit.) a toujours refusé le postulat des mondes possibles et des logiques modales pour caractériser les l'interprétation des expressions. L'argument, que l'on retrouve chez F. Récanati[160], consiste à dire que les mondes possibles sont de pures hypothèses, et donc ne peuvent être considérés comme les fondements d'une caractérisation de l'interprétation. Cette position rejoint celle de B. Smith pour lequel une ontologie doit représenter quelque chose du monde, tout simplement parce que l'information sert à agir sur ce monde. La pharmacie hospitalière en donne un exemple concret : de la qualité de l'information symbolique dépend la fiabilité et la qualité du soin.

L'hypothèse que l'on veut vérifier maintenant consiste à dire que les classifications décrivent des situations dans le monde. Ces situations constituent ainsi des cadres permettant d'interpréter l'information.
Ainsi, par exemple, il existerait une situation dans laquelle M. Smith est corrélé à « kg ». C'est très simple, c'est lorsqu'il est sur une balance. Cette situation est incluse dans une autre, plus vaste, dans laquelle apparaît, à un certain moment, l'opération de mesure du poids, avec l'indication du résultat et la notation du moment où cette mesure a été effectuée.

Néanmoins, les situations que l'on vient de décrire ne concernent que l'élaboration de l'information. Le patient, et l'ensemble des informations se référant à son état, autrement dit, la somme des différents états se rapportant à lui, constitue un autre ensemble de situations représentant à la fois la pathologie et la thérapeutique.
Enfin, à ces situations s'ajoutent celles qui sont relatives à la réception, à savoir le pharmacien interprétant dans son espace de travail à l'aide de ses outils.

La théorie des situations propose de structurer cet univers en trois : l'univers de [207]référence, l'univers de communication (comprenant la situation d'interprétation), et la situation

[207] Nous nous référons ici à la théorie proposée par J. Barwise & J. Perry en 1993 et qui connait depuis de nombreuses évolutions. Rappelons qu'elle constitue le cadre de référence pour positionner le rôle des flux dans la théorie de J. Barwise & J. Seligman.

« ressource », dans laquelle sont caractérisées des connaissances communes relatives à l'interprétation. (Nous traduisons en utilisant la terminologie de J. Barwise & J. Perry).
Notre objectif consistera à spécifier cette structuration relativement à l'information, en caractérisant la situation « ressource » comme le lieu où sont représentées les opérations produisant l'information.
Ainsi, les domaines, dans le cadre des situations ressources, constituent des constructions contextualisées limitées, relatives à un monde de référence et un monde d'outils (produisant des mesures mais également des symboles), et non un ensemble de connaissances génériques.

Enfin les flux caractérisent des phénomènes d'inférences entre des structures hétérogènes de données. Ils ne caractérisent pas explicitement la signification, mais certaines contraintes sur cette signification. Historiquement, la théorie des flux vise à répondre à certaines questions que se pose la théorie des situations, concernant notamment l'enchaînement de différentes situations. Ainsi, comment peut-on passer d'une situation à une autre ? Quelles sont les limites d'une situation ?
La complémentarité des situations et des flux constitue un point important parce qu'elle permet de lier les questions de raisonnement et celles de signification. Nous verrons enfin en quoi nous avons besoin de la théorie des situations, pour caractériser les entités que l'on intègre à l'intérieur des infomorphismes. En effet, les flux ne caractérisent pas les unités qui sont classées : il s'agit d'un raisonnement, qui s'applique à toute entité pouvant être classée.

5.1.2. Notion de situation et « d'infon » : cadre pour la représentation de la signification de la structure.

Afin de décrire comment chaque expression est interprétée dans chacun des mondes, nous reprenons les principes de la sémantique des situations (application à la sémantique du langage naturel de la théorie des situations de Barwise et Perry) tels qu'ils s'appliquent aux composants des flux.
La théorie des situations fait partie de l'ensemble des travaux qui visent à représenter la signification au travers de structures locales. Elles représentent des parties du monde dans lesquelles des informations sont interprétées. Ces structures sont caractérisées à la fois au niveau des types et à celui des instances, ce qui permet de généraliser les modélisations. (Ainsi, on s'approche des frames définis par L. Barsalou[208]). On utilise alors un ensemble d'outils de paramétrage afin de caractériser ces catégorisations.
L'hypothèse des situations serait de dire que le monde est limité par des situations dans lesquelles sont produites et interprétées les informations. Ces situations auraient à la fois une dimension matérielle et une dimension mentale.

En tant que théorie sémantique, la théorie des situations se fonde sur un ensemble de postulats relativement à la signification empruntés à F. Dretske : la signification se caractérise par la classification d'un token dans un type. Le type est défini par une certaine situation : c'est la relation classificatrice d'un token et d'une situation type qui donne la signification d'un phénomène particulier, linguistique ou pas[209].

---

[208] Nous reviendrons sur cette parenté entre la théorie des situations et certains courants de la psychologie cognitive. En effet, au-delà des notions de frame et de contraintes, c'est la caractérisation de la structuration des connaissances et corrélativement de leur représentation, qui est posée à la fois en psychologie et en sémantique.

[209] Cette définition très large de la sémantique permet de caractériser des phénomènes de symbolisation comme l'obtention des valeurs des propriétés des patients par exemple. Une des raisons de notre intérêt pour la théorie des situations réside dans le fait qu'elle rend compte d'opérations de symbolisation ne faisant pas intervenir que des unités linguistiques. C'est le cas par exemple des opérations de mesure.

Pour J. Barwise & J. Perry, cette opération est paramétrée, ce qui signifie qu'il n'y a de classification que relativement à certains paramètres contextuels, comme des connaissances déjà traitées par exemple.

Définition des situations.
Les situations sont ces unités à la fois du monde concret et mental auxquelles réfèrent les structures d'information. Les structures d'information sont des propositions permettant la transmission d'une information. Elles seront caractérisées dans la théorie des situations comme « infons ».
Les situations dépassent la seule dimension des unités du langage pour introduire des objets contextuels que sont les facultés cognitives, considérées comme la faculté d'abstraction et celle d'ancrage. C'est en ce sens que la théorie des situations est une théorie contextuelle : elle introduit des mémoires de configurations des objets du monde construits que sont les situations. Elle s'inspire en partie des « Affordances » de J. Gibson[161]. Ces affordances constituent des unités de traitement des signaux perçus de façon à les reconnaître et les classer dans des situations ou des schèmes déjà construits.
Chez Barwise & Perry, la signification ne se fait pas dans un univers interprétatif intensionnel (ayant sa propre structuration), mais dans une relation entre :

- des événements expressifs, comme une phrase énoncée (ce qui est exprimé) et
- d'autres aspects de la réalité objective, dans laquelle une communauté linguistique détermine l'affectation de choses avec les mots. Il s'agit là de la relation entre des unités du langage et des constructions du monde. (Le langage réfère non pas directement au monde, mais à la façon dont culturellement et cognitivement le monde est construit).

Dans ce cadre, le rôle d'une sémantique est de marquer le lien entre d'une part des mémoires de situations (qui sont des types de scènes) et des situations réelles (ou occurrentes). Ce lien implique le langage dans sa dimension sémantique.
La propriété sémantique associée au langage se marque par la relation entre la capacité de référence et de signification cognitive (ou dans une terminologie sémantique, le sens).

On prendra l'exemple de la carte postale envoyée par Toto :
« Je suis heureux » Toto, Fort-de-France, 28/12/2001.

- On a une relation R, marquée par « heureux », et qui constitue un prédicat à propos d'un individu, ayant un rôle SUJET, marqué par « je ».
- Cette relation et ce rôle sont spécifiés par deux autres rôles, celui marqué par le LIEU et celui marqué par le TEMPS.
- Ces rôles sont individualisés par des arguments : {je, Toto, Fort-de-France, 28/12/2001}. Ces arguments servent à individualiser des rôles, donc à identifier une situation précise, unique.

La relation se marque par le fait qu'un état mental, exprimé à un certain moment, peut être interprété comme cet état mental (à ce moment, en ce lieu et par cet acteur) en un tout autre lieu et par un tout autre acteur. La propriété « être heureux » requiert un individu et un temps et lieu d'assertion, dans lequel se vérifie cette propriété assertée.

La situation ne caractérise pas un contenu : l'état mental n'est pas représenté par la situation. L'état mental n'est pas décrit par la situation. Il est simplement associé à une situation pour être vérifié. (La question du contenu des états mentaux est d'une toute autre nature). La situation est simplement le cadre dans lequel cet état mental se vérifie. (Autrement dit, la

situation peut permettre de transmettre de toutes autres informations, n'ayant que peu de relation avec l'état mental ; simplement, c'est cet état-là qui est exprimé).
L'univers interprétatif est caractérisé par l'ensemble des relations entre les constituants, faisant en sorte que l'état mental se vérifie bien dans le monde. Cette validation se fait grâce à cette situation précise. Il y a donc distribution dans l'univers interprétatif, et donc une caractérisation distribuée de la proposition.

Mais en même temps, on identifie deux situations :
3. celle où Toto est heureux en temps et lieu (cette information peut s'étendre au-delà de la situation indiquée)
4. celle où Toto dit : « je suis heureux ». Cette information est limitée par le temps et lieu de l'énonciation.

On a donc une situation d'énonciation et une situation décrite. La situation est donc un cadre, ou un contexte, dans lequel un certain nombre de phénomènes se produisent.
Ce que l'on peut remarquer dès à présent, c'est que la situation est toujours partielle par rapport aux individus et par rapport aux temps et lieux.
On distingue les ETATS, où une information perdure dans une certaine situation et les EVENEMENTS, où dans une même situation, on a plusieurs informations voire des révisions d'informations.
Une information médicale quelconque s'interprète par trois situations :
- un fait dans le monde à un temps T
- une opération qui permet de représenter par les valeurs numériques ce fait.
- Un événement, à savoir le fait de posséder en continu un poids évolutif et que le fait ne constitue qu'un état des affaires du poids. (Les feuilles pré-imprimées dans lesquelles sont notées successivement toutes les informations représentent des événements).

Caractérisations fondamentales de la signification dans la théorie des situations.
La théorie des situations identifie trois situations- types.
La situation d'énonciation correspond aux faits de l'énonciation contenus dans l'expression. La situation d'énonciation correspond à une structure d'information, telle que l'on peut la représenter par des règles catégorielles. (On entend par là les catégories syntaxiques).

La situation décrite correspond à la vérité ou la fausseté de l'expression interprétée. (J. Barwise et J. Perry, p.6). Ce sont donc les faits dans le monde. Les situations décrites correspondent aux situations dans lesquelles les constituants de l'expression se vérifient.

La situation ressource correspond aux descriptions définies, à savoir à des situations constituant des connaissances nécessaires à la compréhension. Ainsi, dans "le chien mord le passant", on a une situation ressource lorsque l'on identifie le fait que le chien existe et entre autre peut mordre. La situation ressource est caractérisée comme persistante, à savoir caractérise le fait qu'elle contient des propriétés et des attributs qui dépassent largement ceux qui sont présentés dans l'expression.

La signification d'une phrase déclarative se caractérise par la relation entre l'expression et les situations décrites en utilisant les situations ressource. L'interprétation d'une proposition faite par une phrase dans une occasion spécifique identifie une situation décrite.
La signification d'une phrase nominale : relation entre une expression et un individu.
La signification d'un verbe : relation entre une expression et une propriété.

La phrase transmet une information (concernant le monde extérieur ou à propos d'états

d'esprit) où la signification est une relation entre 1 expression et la situation décrite par l'expression. L'interprétation d'une expression à un moment spécifique est la situation décrite.
Ainsi, la signification de la phrase «je suis assis » se caractérise ainsi :
a. On a une expression U où <a> parle dans <L> et
b. on a une situation E où <a> est assis dans <L>.
Ainsi, on obtient la relation entre la situation d'énonciation et la situation décrite.

La situation ressource caractérise la succession des opérations permettant de lier une propriété à un objet. Les propriétés sont associées aux descriptions définies (J. Barwise et J. Perry, p. 146) et sont utilisées pour identifier un objet dans une certaine situation.
Les propriétés (« être rouge, comestible, menaçant »), sont affectées à des objets, lesquels sont en fait des individus pouvant être impliqués dans des situations différentes. Les propriétés, affectées à un individu, servent en fait à isoler une situation dans laquelle se trouve l'individu. (Et où l'individu a cette propriété).
Les propriétés constituent des primitives (des outils d'analyses indécomposables), correspondant à des uniformités entre des situations réelles.
Les propriétés se distinguent des relations parce que leur structure est distincte : <être une chaise, t> vs < R, assis, x, sur y, t>.
Ainsi, l'événement *e* où un individu particulier *a* est assis sur une chaise particulière *b* à *L*, s'écrit de la façon suivante :
En *e* : à *L* : chaise, *b* ; oui.
Assis, *a*, *b* ; oui.

La propriété d'être une chaise est donnée par l'abstraction de l'individu. C'est alors un type d'objet.

Simplement, <être assis> peut être considéré comme une propriété à partir du moment où l'on aura à la fois abstrait l'objet *a* et la localisation *L*. Il s'agit dès lors de l'événement- type E. Cet événement sera alors commun à l'ensemble des situations dans lesquelles les propriétés sont vérifiées, quelle que soit l'hétérogénéité de ces situations.
Les propriétés ne sont pas en nombre limité ; elles peuvent être structurées autour **d'uniformités (partielles) entre des situations distinctes** ; elles servent donc à classer ces différentes situations d'une certaine façon et ainsi établir des relations entre des phénomènes par ailleurs très éloignés.
C'est en ce sens que l'on peut analyser des discours, à savoir le fait que l'interprétation d'une situation nécessite une situation antérieure. (Cette question est distincte de la persistance). (J. Barwise et J. Perry, p.21)

La théorie des situations caractérise les propriétés comme situation ressource. Par l'architecture des paramètres et des situations, une propriété ne constitue pas un ensemble de catégories dans lesquelles on inclut les occurrences, mais une proposition qui se vérifie ou pas dans une situation décrite.

De façon plus sémantique, la propriété constitue une caractéristique prédicative. Le prédicat sert dès lors :

- à construire une structure informationnelle (en affectant des rôles et des arguments satisfaisant ces rôles)
- et à relier la situation décrite à une situation ressource, à savoir une propriété identifiant l'individu dans la situation.

« Marie court » : on a une situation et une propriété :
- la situation est celle de Marie, qui est une personne, à [t], et
- la propriété de <courir>, à [t], et identifiant Marie dans cette situation.

L'interprétation se produit donc dans la mise en relation d'une situation décrite et d'une situation ressource.

Le concept de situation explique comment une même forme linguistique peut exprimer des choses différentes dans des contextes différents, mais dans un contexte unique, n'exprime qu'une seule chose. « Le train de Paris a un retard de 10 mns » a une et une seule signification dans un seul contexte, même si par ailleurs l'information peut circuler entre des personnes situées dans des lieux très différents, et alors avoir des significations différentes[162]. Tout simplement parce que la ressource utilisée ne serait pas la même.

Une signification porte sur une relation entre deux espaces différents, celui d'énonciation et celui décrit. Ainsi, le langage a une **signification externe**, à savoir qu'il sert à identifier des objets, mais également une **signification mentale**, caractérisant la relation entre l'expression et les situations identifiées.

Pour expliciter la signification, J. Barwise & J. Perry caractérisent deux facultés cognitives (qu'ils n'expliquent pas, et renvoient à la psychologie, et par ailleurs à F. Dretske) : la discrimination, ou capacité à segmenter, et l'abstraction, ou capacité à classer dans des types toute occurrence. (La discrimination consiste à classer dans un type afin de transmettre l'objet (c'est ce que fait le flux), et l'abstraction, qui consiste à inscrire l'objet transmis dans un type de situation, marqué par les objets environnants.

Une telle caractérisation permet une prise en compte plus affirmée du **contexte** dans l'interprétation des expressions linguistiques.
La seconde caractéristique fondamentale de l'interprétation qui apparaît dans cette définition de l'interprétation, c'est son caractère **distribué**.

Valeurs de vérité dans la théorie des situations.
Une autre innovation de la théorie des situations est la caractérisation des valeurs de vérité de façon distincte par rapport à la théorie classique des propositions. En effet, les valeurs sont définies par 0/1, au sens où elles se valident ou ne correspondent à rien, et non par vrai ou faux.

Que sont exactement les situations ?
Cette définition des situations peut sembler relativement technique et insuffisante par rapport aux ambitions, relatives à la cognition notamment. La référence aux frames et « affordances » n'est pas totalement satisfaisante.
Les situations et les situations -types sont considérées dans un premier temps par J. Barwise & J. Perry comme des ontologies. (B. Partee, op.cit.).
Ensuite, A. Kratzer considère qu'il s'agit de situations "louées" ou parties d'univers, fonctionnant à la fois comme des individus (de façon analogue à des événements et jouant un rôle direct dans l'interprétation des nominaux événementiels) et comme "substituts du monde". Dans ce dernier cas, les propositions sont interprétées comme ensembles de

situations possibles et les expressions sont évaluées à certaines situations plus qu'à certains univers -temps.

Précisions sur les concepts utilisés en théorie des situations : relations.
La théorie des situations est fondée sur des relations opérant entre des objets paramétriques. On illustre maintenant ce qui est entendu par ces différents concepts et on illustre sur l'exemple de l'adaptation.

Comme nous l'avons vu, il existe plusieurs formulations de la théorie des situations. Celle proposée par J. Seligman & L. Moss[163] insiste sur une caractérisation de la structure d'information à partir d'une relation entre un prédicat et un certain nombre de types qui définissent des rôles et acceptent des arguments.
Les relations sont insensibles au rapport instance/type parce qu'elles ne sont pas définies par un degré d'abstraction (à la différence des rôles et des arguments). Ainsi, les relations caractérisent des événements intervenant entre des entités structurées distinctes. Celles-ci peuvent être des types d'objets, donc être caractérisées par des classifications. Le lien qui pourrait être établi entre les relations au sein des structures d'informations (et notamment les flux), et les relations intrinsèques au sein des ontologies sera questionné afin de montrer la complémentarité des flux et des relations marquées dans le cadre des ontologies.

Les rôles caractérisent des types définis par la relation et les arguments des instances conformes au rôle. Les rôles étant définis par les relations, celles-ci caractérisent la structure de l'infon.

Contraintes.
Le concept de contrainte est essentiel dans la théorie des situations.
On peut rapprocher la définition des contraintes que propose la théorie des situations avec celle qui a cours en psychologie cognitive (voir notamment les propositions de L. Barsalou , op. cit.). Les contraintes constituent des relations nécessaires entre attributs qui s'étendent aux valeurs. Les contraintes sont ainsi distinctes des invariants structurels, qui constituent des relations nécessaires entre propriétés pour constituer un cadre conceptuel.
Les contraintes définissent les relations qui s'établissent entre deux situations, sachant que la première implique la seconde. Cette implications est caractérisée par l'information que l'on transmet de l'une vers d'autre. Cette information se définit par un flux.

Les situations sont contraintes, à savoir qu'elles sont limitées par le fait qu'elles requièrent un flux d'information produit par une situation précédente pour caractériser une succession d'état. Cette information est véhiculée par les « infons ». Le cadre de la théorie des situations permet de spécifier comment, en prenant en compte les flux (donc la transmission de l'information), on peut interpréter des informations au sein d'une situation et relatives à une situation antérieure et conditionnelle.
Si les situations constituent des cadres, les infons constituent les structures permettant de représenter l'information pertinente dans ces cadres. Les « infons » sont des structures d'information. Elles constituent des expressions alors que les situations sont des structures interprétatives, autrement dit sont du domaine des connaissances et du raisonnement. Ce sont des propositions fondées sur une segmentation du monde et qui permettent la transmission des phénomènes occurrents. (Comme nous l'avons vu, cette caractérisation sera reprise par L. Floridi avec l'objectif d'exploiter les dimensions épistémologiques d'une telle assertion).

Les contraintes constituent les règles expliquant comment une certaine situation permet une autre ou l'empêche. Les contraintes rendent compte de l'enchaînement des situations dans une structure d'information.

Enfin, les situations sont paramétrées, ce qui signifie que les relations entre les types et les tokens sont contraintes par des paramètres, à savoir des traits permettant de fonder la relation de classification entre les deux entités. Ces traits sont contextuels et sont produits par les contraintes. Ainsi, une échelle de mesure est paramétrée par rapport à une unité.

Eléments de modélisation concernant la théorie des situations.

La théorie des situations s'applique d'abord à rendre compte de l'hypothèse formulée précédemment d'une triple dimension de l'information : les structures véhiculées par les flux sont les représentations d'une opération de symbolisation, inscrite dans l'espace de travail ET de situations dans l'espace de référence. A ces deux univers, s'ajoute leur dimension mentale (ou intensionnelle).

Pour montrer cela, nous devons expliciter le lien entre les dimensions sémantiques et dynamiques. Dans la théorie des situations, il est marqué par l'introduction des modèles de H. Kamp, qui sont relatifs à la dynamique discursive. Ce travail a été réalisé par R. Cooper[164]. Il est intéressant et potentiellement très productif de noter que R. Cooper lie sa modélisation des expressions informationnelles aux grammaires de C. Fillmore[165]. Cette mise en relation pourra être retenue pour une exploitation ultérieure.

La théorie des situations constitue à la fois une théorie logique mathématique dans l'acception de K. Devlin (op. cit.) (il propose une application en ethnométhodologie dans un de ses ouvrages) et une théorie sémantique. Cette dernière, notamment dans les propositions de R. Cooper, diverge quelque peu avec les propositions précédentes entre autres parce qu'il est nécessaire d'appliquer ce cadre théorique à des problématiques linguistiques. Une de ses principales réussites aura consisté à introduire la dynamique du discours à l'intérieur de la sémantique distribuée que constitue la théorie des situations.

Quel que soit par ailleurs le caractère séduisant de l'approche de R. Cooper, elle s'inscrit dans une perspective de communication, l'objet étant non l'information mais comment elle est organisée dans les discours et interprétée dans ce contexte.

Pour éviter de mélanger flux et situations, il faut considérer que les flux constituent des contraintes sur l'interprétation, elle-même représentée par les situations. La dynamique introduite par les flux ne concerne que les successions d'état et les événements.

Nous pouvons représenter l'interprétation des expressions portées par les flux en distinguant trois étapes :

- La construction des unités de référence directe (ou tokens)
- Les classifications et inférences liées à la structure d'information.
- Les canaux.

Nous présentons maintenant une modélisation simplifiée de notre exemple de circulation d'information en pharmacie hospitalière. Il permet de caractériser les différentes situations identifiables et par là-même à propos desquelles une expression informationnelle pourra être interprétée.

La construction des unités de référence directe représente l'attribution d'un symbole à un objet. Cette attribution est relative à une situation précise. S_1 et S_2 représentent ces situations.

$S_1 \vDash$ (x: "M. Smith")
$S_2 \vDash$ (y: "12:24")

Les situations représentées ici caractérisent la construction des tokens, à savoir les dénominations.

Quant aux opérations, on peut les associer à une situation qui affecte à un token précédemment défini une propriété permettant d'associer un type à ce token ; c'est donc une classification.
La construction des univers intensionnels est intégrée dans les opérations, ce qui est conforme à ce que l'on sait de la cognition située et distribuée.
$S_{OP} \vDash (T_1, T_2)$

Définition de T_1 et T_2. Ce sont des constructions intensionnelles qui reprennent la structure définie auparavant : $\sum_A : \alpha \in$ typ(A).

La situation présentée ici schématise l'opération caractérisée précédemment.

Les classifications peuvent s'exprimer ainsi :
$S_{ClassA} \vDash$ Rel (S_1, T_1)
$S_{ClassB} \vDash$ Rel (S_2, T_2)
Où $S_{OP} \subset <S_{ClassA}, S_{ClassB}>$

Ces deux situations traduisent les classifications présentées dans les flux. Dans ce cadre, la situation qui caractérise les deux classifications est l'opération. Chacune des deux classifications représente une situation, la première avant l'opération (l'attribution d'une propriété), la seconde après (la spécification de la propriété).

L'information, au travers de l'infomorphisme, se caractérise par une situation pouvant être considérée à la fois en intension et en extension.
$S_{OP} \vDash <(x \triangleright y), (T_1 \oplus T_2)>$

Les canaux caractérisent une situation mettant en relation les deux classifications précédemment présentées :
$S_{ClassAB} \vDash$ Rel (S_{ClassA}, S_{ClassB})

On peut maintenant synthétiser de la façon suivante l'ensemble des situations observables :

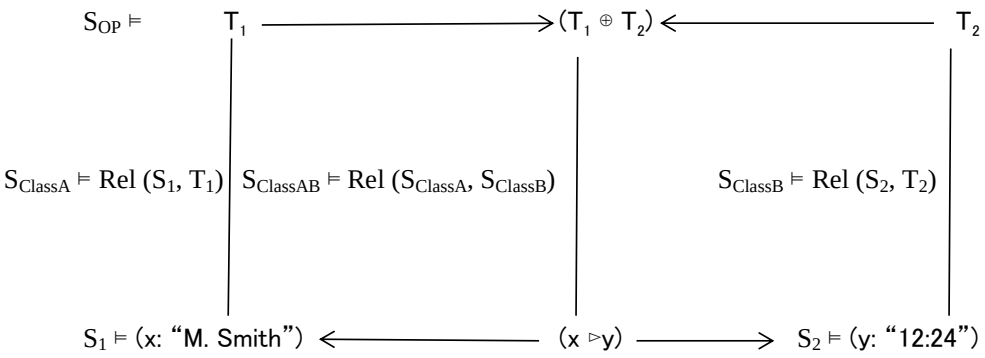


Figure 8. Intégration des situations dans les flux. Exemple sur le cas de l'annonce en gare.

Information, révision de connaissances et croyances.

Nous reprenons ici un certain nombre de thèmes récurrents des théories de l'information, présupposant un rôle de l'information dans la transformation des connaissances et des croyances. Ces théories empruntent des éléments aux théories de la temporalité et des changements d'état que nous avons précédemment considérées. La théorie des situations, notamment dans la formulation intégrant les DRT, ouvre des voies vers la caractérisation de phénomènes dynamiques par la théorie des situations. La théorie des situations, notamment les propositions de D. Israel & J. Perry[166] et par ailleurs de R. Cooper[167], ont cherché à modéliser, notamment au travers du paramètre temporel, les dynamiques discursives et les changements d'états.

L'idée de ces travaux consiste à rendre compte de cadres (comme le sont les situations), au travers de modifications marquées par l'apport de nouvelles informations. Ces propositions ne font

Pistes de travail.

On peut considérer qu'une approche formelle de la signification n'aurait que peu d'intérêt dans un contexte comme celui des problématiques de la description documentaire. Or, les unités langagières qui nous préoccupent sont inscrites dans des outils et des processus dans lesquels la signification est systématiquement contrôlée et limitée par les fonctionnalités des outils. Elles ne peuvent être appréhendées indépendamment de ce contexte opératoire.

Or, un modèle contextuel comme la théorie des situations permet d'intégrer les paramètres externes comme les fonctionnalités des outils documentaires et en même temps, l'appareillage est suffisamment abstrait pour caractériser les uniformités de production de la signification entre ces différents contextes.

En effet, depuis longtemps, les concepts et les mots ont donné lieu à des travaux explorant leurs distinctions. (Le langage constituant à la fois une théorie, un outil de représentation, une faculté cognitive et pouvant lui-même donner lieu à des théories très variées à propos de son unité, de sa plasticité, etc. le débat est beaucoup trop volumineux pour pouvoir être relaté dans ce travail). Par contre, entre termes d'une part, et valeurs, attributs de métadonnées d'autre part, peu de travaux ont été proposés qui prendraient en compte l'ensemble des dimensions impliquées dans la signification. Enfin, entre les éléments de métadonnées et les instances d'ontologies, on dispose également très peu d'études. Les critiques formulées à l'égard de

l'approche de lexique proposée dans le cadre des ontologies par B. Smith indiquent que la caractérisation sémantique des unités doit être raffinée.

Nous avons parlé de l'intérêt de la théorie des situations comme théorie de la signification, intérêt dû au fait qu'elle n'était pas fondée sur des principes seulement linguistiques mais plus généralement liés à la signification naturelle.
Nous avons présenté la façon par laquelle les théories de la cognition située et distribuée rendaient compte des opérations de symbolisation. Nous allons maintenant expliciter cette mise en relation qui permet de fonder la théorie des situations dans un cadre cognitif. Nous prendrons comme point de départ la notion d'opération et le lien que l'on peut alors établir entre la caractérisation des opérations par E. Hutchins et la définition des situations.

On propose de caractériser les dispositifs documentaires (les thésaurus et les différents langages documentaires) comme des outils au sens d' E. Hutchins. Cette proposition nous permet d'envisager une caractérisation de la signification des entités symboliques dans le cadre des outils documentaires qui prenne en compte justement ce que sont ces outils. Une perspective seulement linguistique ne permet pas de représenter le contexte opératoire de ces entités.

Cette intégration permet d'expliciter la relation que l'on peut intuitivement établir entre d'une part la symbolisation et d'autre part les opérations cognitives.

La modélisation des opérations cognitives minimales proposée par E. Hutchins ne repose pas sur une représentation de la notion de symbolisation. C'est d'ailleurs une faiblesse reconnue par l'auteur à propos de son propre travail. Justement, on propose d'intégrer, par le biais de la théorie des situations, cette dimension symbolique. Nous utilisons pour cela les propositions des théories de types, qui instaurent des suites d'opérations binaires. Nous utiliserons pour cela les grammaires catégorielles.

Pertinence de la théorie des situations pour l'analyse des expressions circulant dans le cadre d'un système d'information.
Cette relation entre modélisation sémantique et cognition située et distribuée est importante parce qu'elle permet de répondre à une question que l'on s'est posée lors de notre analyse de la circulation de l'information lors de la tâche d'adaptation de posologies. En effet, comment se fait-il que des suites d'unités symboliques aient une signification suffisamment précise pour permettre une action risquée dans le monde alors même que leur expression est très concise et n'obéit pas aux règles communément admises de la bonne formation syntaxique ? Comment la signification est-elle possible dans un tel contexte et avec de telles contraintes ?
L'apport de la théorie des situations réside déjà dans le fait qu'il s'agit d'une théorie générale de la signification (fondée comme on l'a rappelé sur la théorie du signe de F. Dretske), et non une théorie de la sémantique du langage naturel. Ensuite, elle s'appuie sur la notion de distribution, à savoir le fait que la signification est fondée sur des relations entre des entités (matérielles, cognitives) hétérogènes. Enfin, la notion de contrainte explique de quelle façon une entité interprétée influe sur le traitement sémantique d'une autre. Cette dernière notion permet de résoudre la question de l'absence d'une syntaxe identifiable dans les expressions rencontrées.

Rappelons que l'information, dans notre cadre théorique, est produite. Elle est donc le résultat d'une certaine opération, et représenter cette opération permet à la fois de comprendre pourquoi nous avons effectivement une information, mais également la régularité de cette opération, et enfin de quelle façon cette opération est représentée de façon à former une structure d'information qui permet la circulation de son contenu au travers d' flux d'information.
Cette caractérisation des opérations peut être étendue et validée dans d'autres contextes. En premier lieu, le renseignement de métadonnées, l'indexation, peuvent être considérées comme des opérations de symbolisation paramétrées.

Enfin, si l'on observe la structure relationnelle qui est définie ici, on ne peut manquer d'opérer un rapprochement avec les formulations de RDF.

5.1.3. Questions de sous-spécification des propriétés sémantiques des unités symboliques.

Comme on l'a vu, une des hypothèses fortes de la linguistique contemporaine, notamment formelle, est la sous-détermination (ou sous-spécification) de la signification. Partant de l'idée que les règles de formation des expressions et d'interprétation ont un fondement lexical, on impartit à ces unités la propriété d'être sous-déterminées et de trouver un sens précis dans leur contexte immédiat de réalisation. Il s'ensuit que l'on caractérise alors des structures syntagmatiques spécifiant la signification.
Plus concrètement, l'objectif est de montrer ici que les situations précédemment évoquées et leur distribution dans l'espace correspondent effectivement à différentes acceptions de la signification des entités linguistiques. Ces différentes acceptions sont également des indices de la place des unités à l'intérieur de la structure caractérisée par les flux.
Par exemple, le « retard » considéré en réception n'a pas exactement le même sens que celui caractérisé dans la situation de référence ou dans la situation d'énonciation. Ces différentes significations sont conciliables et contribuent chacune à caractériser la sous-spécification du terme.

Il y a un autre intérêt à ce passage par une sémantique : c'est le fait que les flux puissent rendre compte de phénomènes linguistiques, et que les phénomènes informationnels sont intrinsèquement liés au langage. Cette proposition, affirmée tout au long de ce travail, se doit d'être argumentée.

Ces constructions syntagmatiques ont un fondement sémantique et se caractérisent par une certaine généralité : elles sont donc identiques pour un nombre important d'items lexicaux. Ce sont donc des outils permettant de distinguer des emplois de mots fréquents (comme les connecteurs argumentatifs) ou pour constituer des classes de mots relativement à leurs similarités d'emploi. Pour nous, la question est un peu différente puisqu'il s'agit de construire des ensembles de mots relativement à une même place à l'intérieur de la structure d'information.
On pourrait penser qu'il s'agit là du retour des questions de syntaxe que nous avons précédemment mises de côté. Or, nous n'avions pas rejeté l'idée de structuration des unités symboliques par le biais de leur mise en relation au sein de la feuille ou de l'écran[210]. Les

[210] A ce titre, nous amorçons une mise en parallèle entre notre structuration de l'information et celle que l'on rencontre dans les expressions associées aux langages du web.

unités symboliques ne sont pas distribuées au hasard sur la feuille, mais selon des règles précises et strictes.

Dès lors, la sous-spécification sert à répondre aux deux questions suivantes :

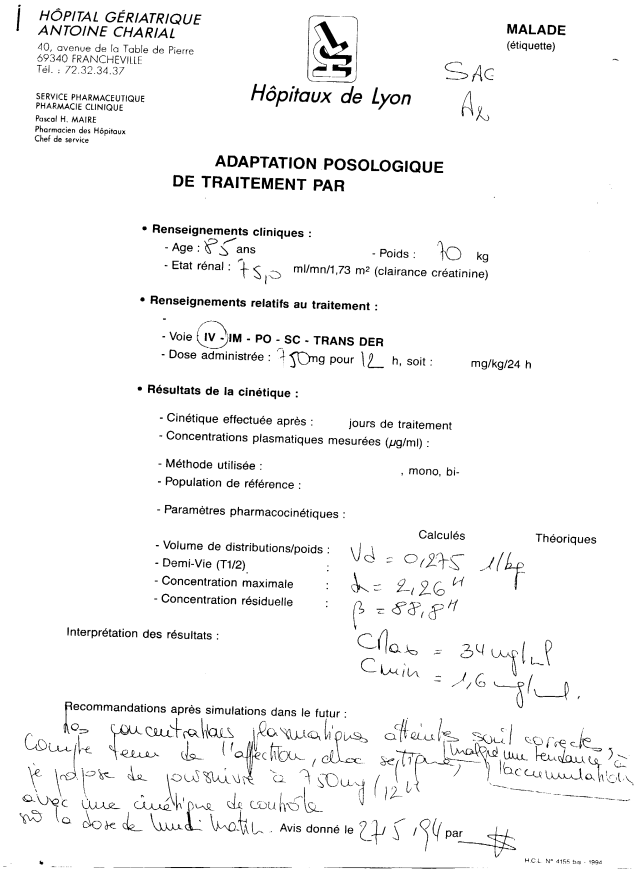
- En quoi la structure d'information sert à spécifier la signification des unités symboliques
- En quoi, en observant que les termes peuvent avoir des significations complémentaires en fonction des contextes, on peut obtenir une représentation cohérente et complète du sens.

Pour ce résultat, nous considérons d'abord le cadre de l'analyse des unités symboliques, qui diffère quelque peu de celui des opérations parce qu'il prend plus en compte la dimension discursive, à savoir la structure des feuilles pré-imprimées.

Cadre d'observation des phénomènes de sous-spécification.

Il est méthodologiquement nécessaire d'identifier le cadre de discours dans lequel l'analyse s'intègre. Pour cela, il apparait fondamental de caractériser la façon dont ces discours réfèrent dans l'espace de l'activité. Enfin, nous pourrons caractériser l'analyse linguistique.

Afin d'illustrer la mise en discours, nous proposons le fac-similé suivant et une schématisation du processus d'alimentation de ces feuilles :



HÔPITAL GÉRIATRIQUE
ANTOINE CHARIAL
40, avenue de la Table de Pierre
69340 FRANCHEVILLE
Tél. : 72.32.34.37

SERVICE PHARMACEUTIQUE
PHARMACIE CLINIQUE
Pascal H. MAIRE
Pharmacien des Hôpitaux
Chef de service

Hôpitaux de Lyon

MALADE
(étiquette)

SAC
Ax

ADAPTATION POSOLOGIQUE
DE TRAITEMENT PAR

• **Renseignements cliniques :**
- Age : 85 ans - Poids : 70 kg
- Etat rénal : 75,10 ml/mn/1,73 m² (clairance créatinine)

• **Renseignements relatifs au traitement :**
-
- Voie (IV) - IM - PO - SC - TRANS DER
- Dose administrée : 750 mg pour 12 h, soit : mg/kg/24 h

• **Résultats de la cinétique :**
- Cinétique effectuée après : jours de traitement
- Concentrations plasmatiques mesurées (µg/ml) :
- Méthode utilisée : , mono, bi-
- Population de référence :
- Paramétres pharmacocinétiques :

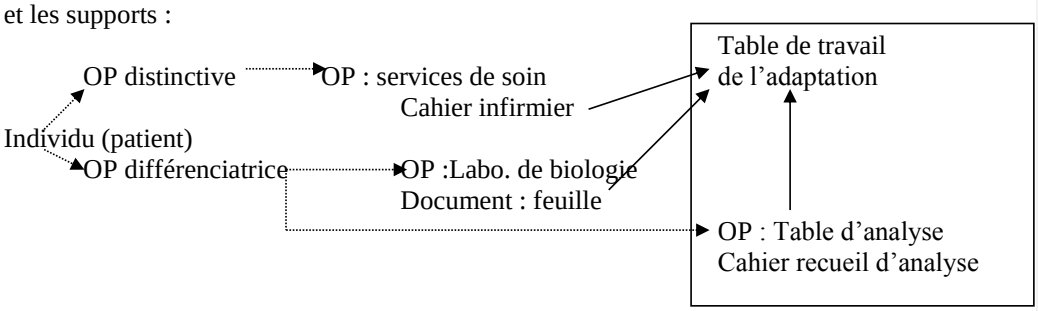
	Calculés	Théoriques
- Volume de distributions/poids :	Vd = 0,275 l/kg	
- Demi-Vie (T1/2) :	k = 2,26 h	
- Concentration maximale :	β = 88,8 h	
- Concentration résiduelle :		

Interprétation des résultats :
Cmax = 34 µg/ml
Cmin = 1,6 µg/ml.

Recommandations après simulations dans le futur :
Les concentrations plasmatiques atteints sont correctes, compte tenu de l'affection, chez ce patient traité une fenêtre d'accumulation je propose de poursuivre à 750 mg/12 h avec une cinétique de contrôle au 10e jour du traitement. Avis donné le 27/5/94 par

HCL N° 4155 bis - 1994

On schématise cette circulation de la façon suivante, en ne retenant que les distinctions de lieu et les supports :

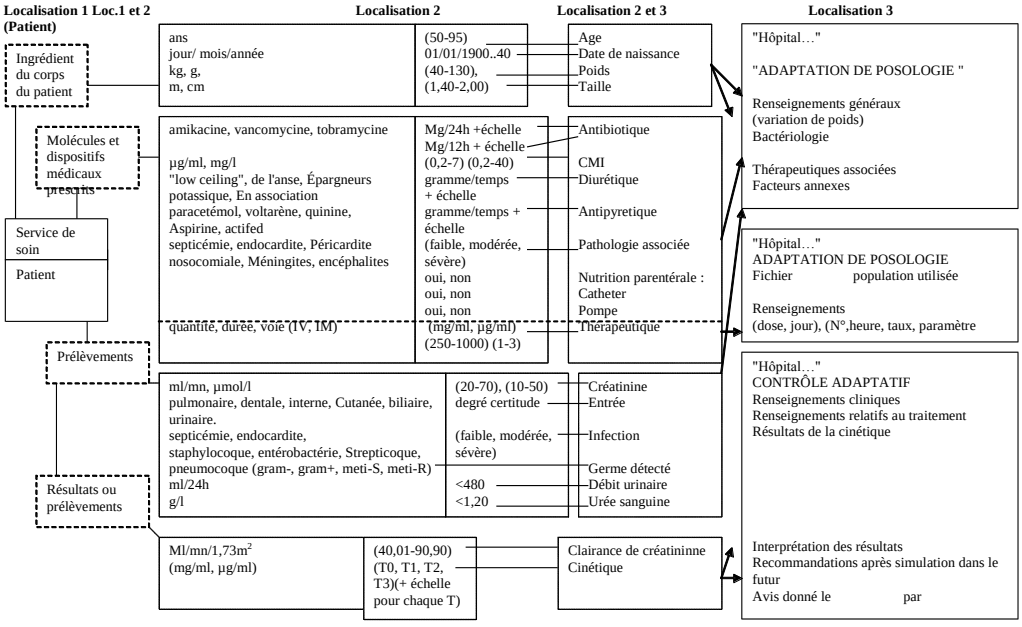


Espace de la pharmacie

→ : Circulation d'informations écrites.
⇢ : Circulation d'objets.

OP : opération (nous reprenons la définition des opérations présentée plus haut). Les opérations distinctives sont des mesures de propriétés de patient. Les opérations différenciatrices impliquent des prélèvements et des analyses de différenciation interne dans ces prélèvements de produits afin de permettre une mesure. Une opération distinctive consiste à associer une valeur individuelle à une propriété du patient alors qu'une opération différenciatrice est une opération distinctive mais requérant que les propriétés soient associées à plusieurs individus : le patient, mais également la molécule ou la bactérie.

Nous présentons maintenant, toujours à propos de la pharmacie, un tableau d'ensemble de la distribution des unités linguistiques dans l'espace de l'activité.



Identification de l'ensemble des entités circulant dans le système d'information.

Cette schématisation identifie le corpus linguistique global dont on dispose pour mener l'analyse. Néanmoins, on n'identifie pas exactement ici les spécificités des différentes opérations présentées plus haut.

La description fait apparaître quatre formes d'expression distinctes, associées à quatre sortes d'opérations distinctes.
O1 : opération simple.
O2 : opération avec objet.
O3 : absence de représentation d'opération
O4 : Validation d'existence d'objet.

O1 : opérations appliquées sur l'individu directement avec sélection d'une propriété de l'individu dans le temps (marquée par le canal).
O2 : opérations sur la base d'un objet associé extensionnellement à ce patient (comme une molécule) ET représenté par son type.
(Paraphrase du type : Il y a l'AMIKACINE chez ce patient, et l'AMIKACINE est le type de cette AMIKACINE chez ce patient). (Cette paraphrase est validée empiriquement : une molécule chez un patient n'a pas le même comportement que cette molécule dans la littérature).
Les expressions de la forme O3 ne représentent pas l'opération :

- absence de dédoublement de l'opérateur
- ni de marque temporelle comme indice d'objet.

Il en va de même pour la forme O4 ("nutrition parentérale") pour laquelle l'opération est occultée. Le problème est alors celui de l'analyse du marqueur d'assertion.

Enfin ces quatre structures peuvent considérées comme la réduction d'une structure : O2

En respectant les différents paramètres déjà présentés, on peut caractériser de la façon suivante les différentes opérations :

Argument1 :	Opérateurs canaux : O	argument objet : E	Opérateurs d'opérations : D	arguments résultats : R
P (patients)				
	O1		DA	RA
	Ages		ans	(50-95),
	Poids		kg, g	(40-130),
	Taille		m, cm,	(1,40-2,00),
	CMI		mg/ml, µg/ml	(0,2-7), (0,2-40)
	Créatinine		ml/mn, µmol/l	(20-70), (10-50),
	clairance de créatine		Ml/mn/1,73m^2	(40,01-90,90),
	débit urinaire		ml/24h	(<480),
	Urée sanguine		g/l	(<1,20),
	O2	EA	DB	RB
	Antibiotiques	Amikacine…,	mg/24h, mg/12h, µmol/l,	échelles de chaque antibiotique
	Diurétiques	"low ceiling"…,	quantités/jour	échelles de chaque diurétique
	Antipyrétiques	paracétamol…,	g/t	échelles de chaque antipyrétique
	thérapeutique	Amikacine…,	mg/ml, µg/ml	250-1000
	cinétique	Amikacine…,	mg/ml, µg/ml (T1, T2, T3))	échelles des cinétiques en fonction des différents T
	O3	EB		
	Entrée	pulmonaire…		
	germe	staphylocoque… (gram -, gram +, meti-S, meti-R)		
	pathologies associées	septicémie…,(faible modérée sévère)		
	infection	septicémie…,(faible modérée sévère)		
	date de naissance	(01/01/1900-12/12/1940)		
	sexe	M/F		
	O4			RD
	nutrition parentérale			oui/non
	catheter			oui/non
	pompe			oui/non

Représentation des différentes classes d'opérateurs et d'arguments en fonction des structures qu'ils identifient.

Figure 9. Représentation des différentes réalisations des unités lexicales dans le cours de

l'activité

Néanmoins, cette représentation n'identifie pas clairement le positionnement des unités dans la structure des feuilles, et notamment les variations d'ordre.

Si l'on observe la structure linéaire, l'inversion de la place du résultat par rapport au paramètre met en valeur un changement systématique d'ordre.

Autrement dit, on passe des formes

O1 DA RA (Smith 18h24 *créatininémie* ml/mn 5,4)

Ou

O2 EA DB RB (Smith 18h24 *Amikacine* mg/24h 750)

à des formes

O1 RA DA (Smith (18h24) *créatininémie* 5,4 ml/mn)

Ou

O2 EA RB DB (Smith Antibiotique : Amikacine départ 750 mg/24h)

Pour les structures O3et O4 il n'y a pas de transformation : O3 EB dans les deux cas. *Néanmoins, entre O3 et EB, on peut voir apparaître des modifieurs, attachés à EB "certainement pulmonaire". Le modifieur peut être "?", et dès lors placé après EB.*

Cette présentation permet de situer le cadre de travail dans lequel se pose la question de la sous-spécification. En effet, la sous-spécification repose sur l'hypothèse selon laquelle les constituants linguistiques sont sous-déterminés, et donc, que le contexte permet de les spécifier. Justement, nous spécifions ce contexte, et nous intégrons le fait que ce contexte est décrit par des opérations de production des unités symboliques et de leur duplication dans des espaces dédiés.
En ce sens, il existe un cadre contenant des espaces dévolus à certaines unités symboliques, et ce cadre structure donc les expressions informationnelles.

Sous-spécification et rôle du contexte dans la formation de la structure d'information.
Avant d'aborder plus précisément la sous-spécification, il faut revenir un peu plus précisément sur quelques présupposés de la théorie des situations sur la théorie du langage lui-même, et qui font que la sous-spécification constitue un abord approprié des phénomènes linguistiques.
Le principal est la relativité du langage[168]. Le relativisme linguistique permet d'analyser des unités et des structures linguistiques en contexte, plus précisément considère que des modèles du contexte sont requis pour analyser l'interprétation des unités linguistiques. Les situations constituent des modèles contextuels du point de vue des unités linguistiques. Le relativisme s'appuie notamment sur le rôle des indexicaux dans l'interprétation sémantique. En effet, ceux-ci renvoient directement au contexte, à la différence des unités conceptuelles. L'interprétation des déictiques renvoie directement au contexte. L'hypothèse défendue entre autre par F. Récanati[169] (p. 7), c'est suivant R. Montague, que les phrases indexicales ne sont pas des propositions au sens standard. Une proposition se caractérise par une fonction depuis des mondes possibles vers des valeurs de vérité. Or, les phrases indexicales se caractérisent par une fonction depuis des points de référence (en temps, lieu et énonciation) vers des

valeurs de vérité. Au-delà des valeurs de vérité, on caractérise le contenu de l'expression, donc ce qui est apporté par l'expression à propos de l'objet désigné.
Ainsi, on peut considérer que les expressions sont sous-spécifiées tant qu'elles ne possèdent pas de marqueurs déictiques permettant de les faire référer à un contexte défini.

La question peut être tournée différemment ; toute proposition relativisée contient des constituants inarticulés. Ainsi par exemple, « il est 5 heure » est une proposition relativisée à une certaine zone temporelle. Mais cette zone temporelle n'est pas un aspect de la pensée qui serait exprimé ici : c'est un positionnement dans les conditions de vérité. C'est un aspect de la circonstance à partir de laquelle ce qui est dit ou cru est évalué. Dans ce cas, la pensée « concerne » une zone particulière de temps, mais n'est pas « à propos » d'elle. En fait, la pensée est effectivement contextualisée dans le cadre défini par les indexicaux.

Nous adoptons ici un point de vue interprétatif, c'est-à-dire que l'on étudie les conditions produites pour l'interprétation des expressions. Nous partirons de la structure transmise pour ensuite caractériser plus précisément la signification des unités la composant et formant la structure d'information. En ce sens, la sémantique que l'on définit ne caractérise pas explicitement le langage naturel, mais bien les relations entre les unités de la langue et les contextes. Rappelons que la sémantique distingue (1) les unités du langage naturel et (2) les discours, à savoir des formes réalisées. Elle recoupe l'articulation entre les propriétés linguistiques des unités et les usages.

C'est par cette méthode que l'on pourra au mieux préserver la spécificité de l'approche informationnelle : la structure d'information comme déterminant la signification des catégories qu'elle accepte. En ce sens, notre problématique intègre les flux, et puisque les catégories sont déduites des flux, ils sont au centre de la sémantique que l'on propose.

Déictiques (ou indexicaux).

Nous ne pouvons pas ignorer la dimension des déictiques, qui est essentielle pour la caractérisation linguistique des entités intervenant dans les flux. En effet, les unités du niveau des tokens sont des déictiques et les types sont des unités conceptuelles, mais d'un niveau d'abstraction moindre par rapport aux canaux.

Les positions de F. Récanati[170] l'amènent à proposer le concept de dossier mental pour caractériser les concepts relativement aux situations qui les accompagnent. Ces dossiers mentaux constituent des cadres servant à classer les expressions indexicales et donc de les reconnaitre. Ces dossiers caractérisent la dimension intensionnelle des expressions.

Les indexicaux ne constituent pas les seuls marqueurs de dépendance au contexte : les marqueurs de circonstances d'interprétation et la modulation (les expressions sont interprétées en utilisant des connaissances permettant de compléter l'information transmise, et donc affectant les contenus) constituent d'autres marqueurs de dépendance au contexte. Ces marqueurs permettent entre autre de distinguer un contenu linguistique interprété littéralement d'une proposition interprétée dans le cadre d'une activité. Cette distinction permet à F. Récanati de distinguer entre interprétation sémantique et pragmatique.

Le fait marquant est que les tokens des flux correspondent à des indexicaux, que leurs classes sont des unités distribuées, et que les canaux sont les termes caractérisant les concepts les plus

abstraits, et qui sont effectivement relatifs à des constructions intensionnelles permettant, par la sous-détermination, de caractériser des phénomènes différents et complémentaires dans les différents mondes.

Application : caractérisation sémantique des unités informationnelles.

Jusqu'à présent, nous avons essentiellement considéré la modélisation des opérations et la façon dont on pouvait caractériser différemment l'élaboration de la signification. Nous présentons maintenant de quelle façon il est possible de traduire ces différentes opérations dans le cadre de leurs spécificités de signification. Nous considérons ainsi de quelle façon les résultats des opérations prennent sens.

Nous reprenons une nouvelle fois l'exemple utilisé dans les parties précédentes, mais en proposant une autre sorte d'explication. Nous ne nous intéressons plus aux mécanismes de flux, mais à l'interprétation de l'information transmise. Nous centrons donc notre propos sur la dimension sémantique.

Nous reprenons donc les structures caractérisées dans le cadre des flux et nous les explicitons maintenant (où ≈ symbolise une traduction) :

<Object, nomination> ≈ a.A

Tout objet du monde est caractérise par une paire associant une dimension référentielle et une nomination. Cette paire constitue une construction du token.

<Part of (obj., nom.), token. (Par. 1)> ≈ a\b.B

La condition pour avoir deux structures hétérogènes liées par une fonction inférentielle est que la deuxième paire indexicale soit d'un grain plus fin que la première. La relation [part_of] n'est pas universelle. Néanmoins, elle permet d'identifier la relation entre les deux entités hétérogènes liées par une fonction inférentielle du niveau des tokens. (Par. 1) représente l'opération matérielle permettant de construire le token b.

Maintenant, nous pouvons passer aux classifications, à savoir le lien entre les opérations matérielles et symboliques. Ce lien se marque par des paires associant un objet et un paramètre, relativement au fait que le paramètre constitue un type et donc doit être représenté de façon autonome. On retrouve les trois composants des opérations dans chacune des parenthèses, sachant que [Inst.] caractérise une instance choisie acceptée par le paramètre, et constitue donc un résultat d'opération.

(Inst. Par.2, token. (Par. 2)), (Inst. Par.3, token. (Par. 3)) ≈ ⊨ [a,b], ($\alpha \in \Gamma$, $\beta \in \Delta$)

La traduction de la classification <⊨, a, α> est représentée dans la première partie de l'expression.

Les relations fonctionnelles entre opérations ou paires contra-variantes de fonctions : < $f^{\wedge}$, $f^{\vee}$>, sont marquées au travers de l'implication de chacune des opérations. Si nous avons déjà expliqué cela au niveau des tokens (par l'opération utilisant le paramètre 1), par contre la

seconde est réalisée au travers d'un outil qui associe au type de la première classification (une propriété de l'individu token *a*) le type de la seconde (une unité mesurée sur le token *b*).

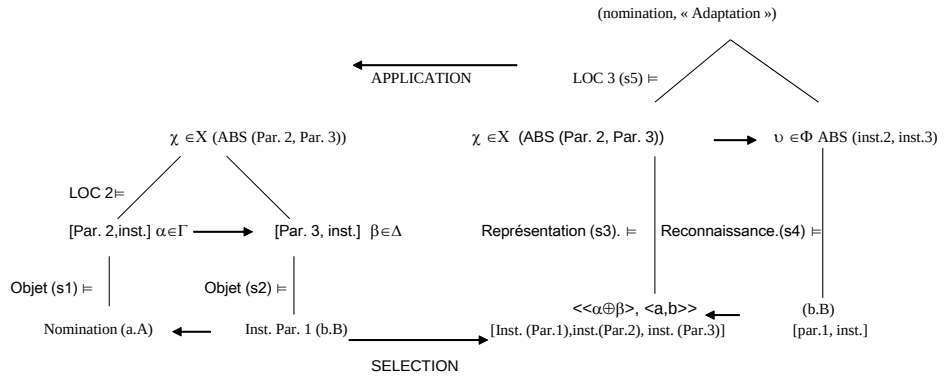
On représente les opérations ordonnées à l'aide des fonctions: un type $\alpha \in \Gamma$ peut être traduit dans un type plus large $\beta \in \Delta$ sous la condition que α classe *a*, donc β classe *b*. (Sit: situation ou cadre dans le monde ; il s'agit de la portion du monde qui n'est pas transmise par le flux d'information et qui concerne le protocole réalisé lors de l'opération).

$$\begin{array}{ccccc} \alpha\in\Gamma & — & f^{\wedge} & \rightarrow & \beta\in\Delta \\ s_1 \vDash \ | & & & & s_2 \vDash \ | \\ a.A & \leftarrow & f^{\vee} & — & b.B \end{array}$$

Cette représentation caractérise comment le résultat des opérations peut être transmis. Elle ne montre pas comment une telle construction peut être interprétée ni le canal par lequel elle peut être transmise.

Nous pouvons proposer la caractérisation suivante du canal. Le canal, rappelons-le, classe l'infomorphisme précédent et assure le transfert, autrement dit l'opération de duplication. Un canal est défini par sa capacité de transport, ici une certaine paire de classifications. Dans notre exemple, "Kinetic": est le concept permettant ce transfert. Il classe à la fois les types "*M. Smith, mg/ml*" et les tokens "*12h34, 5,56* ".

Si $\chi \in X$ (X est l'ensemble des canaux acceptés par le système), il y a une situation complète s_3 où χ classe n'importe quelle information contenant chaque token et type de l'information. Le canal classe les deux opérations précédemment identifiées dans un seul type. Ce type synthétise les deux opérations et assure le transfert de l'information pertinente par le biais d'un concept dont l'interprétation est partagée par l'ensemble des acteurs de la communauté d'émission et de réception de l'information. On obtient ainsi la représentation déjà présentée dans la partie précédente et dont on peut maintenant détailler certains aspects :



$s_3 \vDash < \chi, <<\alpha \oplus \beta>, <a,b>>>$.

Cette situation assemble l'ensemble des procédures liées à des opérations constitutives d'une information. Elle est dénommée « cinétique ». Elle permet d'interpréter de façon

complémentaire chaque situation identifiée précédemment ; s_1 et s_2, sachant que chacune de ces deux situations s'interprète au niveau des types et au niveau des tokens.

La seconde classification de la duplication $s_4 \vDash$ <<b>, < a,<α⊕β>, υ>>: représente la classification dans le media, à savoir la feuille pré-imprimée "*adaptation*" dans notre exemple. Il s'agit de la situation qui permet de faire le lien avec les espaces d'états et la dynamique discursive. Ainsi, l'information est inscrite dans un processus qui la dépasse et dans lequel elle s'interprète comme un état.

Tout canal permet la circulation de l'information :

SELECTION représente l'opération de duplication (et la disparition des traces symboliques non pertinentes). Il s'ensuit une représentation plus abstraite des objets du monde. Elle reprend le principe d'économie formulé par G. Chaitin.

APPLICATION caractérise la direction de l'interprétation (d'une représentation abstraite vers le monde).

APPLICATION et SELECTION caractérisent une paire de fonctions entre les deux espaces caractérisant le système d''information.

Conclusion.
La question de la structure et de son rôle sont posés ici : en quoi les relations entre les unités produisent une signification qui ne peut être caractérisée par les seules propositions composant la structure ? On a montré que les flux étaient associés à des dimensions sémantiques propres, à savoir des interprétations qui ne pouvaient être associées à l'interprétation d'une expression.

Ces propriétés d'interprétation ne peuvent être seulement associées à une forme abstraite et contextuelle. Elle se traduit dans l'usage du langage par des catégories d'entités linguistiques satisfaisant les contraintes des flux et permettant de fonder une structure d'information. On pourra ainsi définir des classes qui organisent des ensembles de termes selon des règles propres à l'information.

Ainsi, on pourra montrer que les flux ne constituent pas seulement une représentation de transferts d'information, mais qu'ils s'appuient sur une structure linguistique. Néanmoins, avant cela, nous devons préciser la structuration des univers de référence, de façon à disposer d'une représentation du contexte stable et régulière.

5.1.4. Pertinence de l'articulation situation/flux pour la caractérisation de la signification.

L'hypothèse que l'on développe est que chaque expression informationnelle a deux directions distinctes d'interprétation, l'interprétation au niveau des types et des structures intensionnelles, et une interprétation référentielle, associée au niveau des tokens. A ces deux interprétations, on en ajoute une troisième : la connexion entre ces deux mondes permet de spécifier le fait qu'une expression signifie relativement à un outil de représentation, qui appartient à un monde propre de référence, et relativement à des objets qui sont représentés. Ces deux contraintes (les deux dimensions interprétatives et la caractérisation d'un univers

composé des outils permettant de représenter l'information) limitent les possibilités expressives, donc le nombre des unités lexicales que l'on peut trouver à telle ou telle place de la structure. Tout se passe comme si la dualité extension/intension ou concept/référent était mise à mal par une analyse précise du fonctionnement de la référence dans un contexte ici relativement spécifique. Tout se passe comme si on devait présupposer un autre niveau de référence, intermédiaire entre le monde référence et la conceptualisation intensionnelle. Cette caractérisation tient dans le fait qu'il soit possible de structurer l'univers de référence et que par ailleurs il soit possible de représenter la façon dont l'information est élaborée, circule et est interprétée dans un cadre spécifique.

L'approche par la théorie des situations ne s'explique pas seulement par la familiarité théorique et philosophique avec la théorie des flux ; comme G. Cooper le précise, la théorie des flux visait au départ à préciser certains concepts de la théorie des situations, notamment celui de contrainte et de contrainte conditionnelle. Les situations caractérisent des structurations du monde fondées sur des dimensions linguistiques et conceptuelles ; ce que les flux montrent, c'est le lien qui peut être opéré entre ces différentes situations. Ces lien se marque par un transfert d'information et se caractérise par des contraintes entre les différentes situations. Ainsi, les flux permettent de caractériser une dynamique à l'intérieur même de l'interprétation.

Ces propriétés permettent de distinguer la théorie des situations d'autres perspectives de sémantique dynamique, qui se sont développées autour des DRT. La première différence avec la théorie des situations est liée au fait que la dynamique est considérée dans les DRT au sein même des discours, dans la prédication. Dans la théorie des situations, elle est liée aux univers d'interprétation et au monde de référence.

Cette perspective permet de mettre en place une sémantique qui ne viserait plus seulement à caractériser comment des états ou des événements du monde sont représentés, mais quels sont les moyens expressifs que l'on peut mobiliser pour représenter un état temporaire du monde, qui, de plus, aurait la possibilité d'assurer la transition vers un autre état à l'aide d'un flux d'information. Cela permet d'affirmer le statut propre de l'information, et de distinguer la dynamique propre que représente le discours de celle des entités du monde.
En effet, l'intérêt de l'approche référentielle que l'on a proposé réside en partie dans le fait que l'on peut poser la dynamique comme un phénomène extérieur au discours, existant chez le patient ou dans le trafic des trains, et pas seulement dans le discours.
Ces moyens expressifs sont contraints par les opérations qui construisent ces informations. En d'autres termes, la structure d'information possède sa forme canonique dans des dispositifs supportés par des outils et des opérations techniques, tels les résultats d'analyse médicale ou les annonces en gare. Si ces exemples peuvent apparaître limités, les principes qui peuvent être établis à partir de ces objets peuvent être étendus.

Caractérisation de l'interprétation des flux et modélisation générale.
L'observation de la complémentarité entre flux et situations passe par la mise en relation que l'on peut opérer entre d'une part les unités symboliques et leur structuration, et d'autre part les opérations matérielles et techniques qui permettent de produire ces entités et de les faire transiter. Nous n'abordons pas maintenant les implications théoriques de cette intégration d'opérations qui ne sont pas seulement symboliques mais également cognitives dans le cadre de la caractérisation de la signification : nous l'aborderons dans le 5.3.
Il s'agit ici de comprendre comment les flux peuvent être interprétés, et ce de façon spécifique par rapport à la structure d'information. Néanmoins, pour aborder cette question, il est

nécessaire au préalable de caractériser quelle interprétation est construite par les différentes situations construites par les flux.
En intension, l'expression "*M. Smith, 12h34, Kinetic, mg/ml, 5,56 »* est classée dans la représentation intensionnelle du patient en s_3 : cela signifie que le processus de cinétique (s_3) classe l'opération réalisée en s_1 et s_2 (et relative au tokens a et b) et permet d'interpréter les types transmis comme une information à propos d'un état de ce processus. On raisonne ici sur le déroulement d'un processus de cinétique.
En extension, le concept $\upsilon \in \Phi$ (où est Φ l'extension des concepts de X) est une interprétation au niveau des tokens (s_4) : cela signifie le token b, fonctionnellement lié au token a, représentent les objets dans lesquels les représentations d'état disponibles sont valides.
N'importe quel type χ de X est fonctionnellement spécifié dans le type extensionnel de υ de Φ où ν est l'extension de χ. On raisonne ici sur l'état d'un patient dans le cadre d'une cinétique.
Ces deux caractérisations de l'interprétation (celle du processus biochimique et celle du contrôle thérapeutique du patient) sont possibles au travers des deux niveaux de représentation, caractérisés au travers des flux.

Nous proposons la modélisation générale de l'interprétation. Ainsi le flux d'information devient un cadre explicite pour caractériser l'interprétation.
On interprète à partir d'une représentation contrainte par des opérations, caractérisant des symbolisations spécifiques, hétérogènes par rapport à l'objet qu'elles représentent. L'interprétation d'états de patient se fait non seulement sur la base de l'information représentée, mais également en considérant les contraintes formées par l'ensemble des opérations de symbolisation. L'interprétation des tokens requiert celle des opérations ayant permis la symbolisation de certains états –partiels- de ces tokens.

La caractérisation des interprétations va au-delà d'une application des flux. Elle concerne les propriétés du raisonnement à distance. Si raisonner à propos de l'état d'un patient sans ne jamais avoir accès à la matérialité de cet individu peut sembler un handicap pour le raisonnement, cela peut être aussi un avantage. En effet, le raisonnement à distance permet de produire des inférences vraies qui pourraient ne pas être validées dans le cadre d'une observation in praesentia de l'individu. Cela peut sembler assez trivial, mais c'est justement ce qui vient argumenter l'intérêt des théories et des logiques locales, et par-delà, le rôle de l'information dans la formation des raisonnements abstraits et leur application dans le monde.

Caractérisation des flux et interprétation de l'information : conclusion.
Nous verrons plus loin l'enjeu des opérations concernant la construction des mondes de l'interprétation, et l'intégration plus avancée des concepts de la cognition située et distribuée. En effet, tant que l'on n'a pas modélisé la construction de l'univers d'interprétation, il n'est pas possible de rendre compte de façon plus détaillée de l'interprétation des constituants de l'information. Jusqu'à présent, nous avons caractérisé comment et pourquoi l'information circulait et pouvait être interprétée, mais pas le fondement même de cette interprétation, à savoir la caractérisation de faits dans le monde et dans l'esprit.
Nous avons proposé une application de modèle des flux mais également une caractérisation des objets sur lesquels ils peuvent s'appliquer. En l'occurrence, il s'agit d'objets et d'opérations dans le monde qui peuvent sembler éloignées des structures de données que l'on trouve sur le web. Or, ces structures de données, qu'il s'agisse de thésaurus, d'ontologies, de lexiques, de terminologies ou encore de métadonnées, sont d'abord des structures d'objets

symboliques issues d'opérations techniques régulières. Ce sont des résultats d'opérations impliquant des paramètres, des outils et donc des connaissances spécifiques, différentes de celles qui sont utilisées par les flux.

La caractérisation des opérations est une étape essentielle parce qu'elle a permis d'isoler le processus de symbolisation, que nous avons choisi matériel, donc plus facilement observable que l'élaboration d'un thésaurus. Ensuite, nous avons pu considérer comment ces résultats d'opérations pouvaient être assemblés et transmis dans le cadre d'un flux d'information. Enfin, nous avons pu considérer de quelle façon les flux pouvaient contribuer à structurer une activité contenant des opérations (des raisonnements externalisés), des raisonnements de contrôle et d'estimation associés à d'autres facultés mentales que les opérations de mesure.

Tous ces phénomènes impliquent que l'assemblage des symboles transmis soit en lui-même structuré afin que l'interprétation ne donne pas lieu à une erreur d'interprétation. C'est ce que nous allons donc voir à propos de la structure d'information.

5.2. Classification et primitives de la structure d'information.

Afin de caractériser les constituants de la structure d'information, on explore en quoi les principes de classification présentés par les flux permettent de construire des primitives de la structure d'information. On a caractérisé la structure d'information comme un format (et non comme une expression véhiculant une intention comme dans la perspective discursive). Cela signifie qu'elle possède des constituants acceptant certaines valeurs. En ce sens, une structure d'information peut être structurée et représentée par un schéma RDFS tout autant que par une phrase du langage naturel ou par une structure comme celle-ci : "M. Smith, 12:24, Kinetic, 4.8 mg/ml".

La structure d'information n'a donc pas à être caractérisée par des catégories au sens linguistique puisqu'elle est le produit des flux, mais par des constituants fondamentaux qui prescrivent des propriétés sémantiques que doit satisfaire chaque constituant candidat. Telle entité symbolique peut être de tel constituant type de la structure d'information relativement à certaines propriétés de comportement sémantique qu'elle valide. Un comportement sémantique est à la fois une propriété sémantique et un certain rôle dans un flux d'information (attesté à la fois dans la description et la modélisation). Ainsi, la structure d'information réutilise les résultats des précédentes analyses.
Les constituants fondamentaux sont des types d'objets symboliques obtenus à partir de l'ensemble des opérations précédemment évoquées et qui se situent à une place précise dans le dispositif planaire de la feuille, leur permettant alors de contribuer à la structure permettant l'interprétation de l'information.

Les flux mettent en évidence des régularités concernant la production, la circulation et l'interprétation de certains phénomènes d'information. Néanmoins, ils ne permettent de représenter que le processus informationnel, et non la structure. Or, pour aborder cette structure, il convient de reprendre la modélisation en comprenant exactement comment les unités symboliques sont produites et surtout, de quelle façon on peut rendre compte des opérations de symbolisation présentes dans l'ensemble de l'espace couvert par l'activité. Ces opérations sont envisagées dans le sens où elles permettent de produire des unités symboliques qui se composent les unes par rapport aux autres.
C'est pour cela que l'on aura besoin de spécifier ce que l'on peut entendre par opération. La caractérisation de ces opérations régulières permettra de comprendre un peu mieux ce que

l'on entend par structure, considérant l'information. Rappelons qu'il s'agit ici d'expliquer pourquoi, en l'absence d'une syntaxe du langage naturel, on dispose d'une structure.

Au-delà de l'argumentation, l'enjeu de la structure d'information est l'élaboration d'un modèle d'extraction d'information qui permette de circonscrire une information à l'intérieur d'un texte.

Opérations et constitution de la structure d'information.
Les opérations que l'on a identifiées précédemment s'enchaînent de façon contraignante. Cela s'inscrit dans le principe des flux et des situations. L'ensemble limité de ces contraintes construit la structure d'information. On pourrait considérer que les règles de l'élaboration de l'information seraient suffisantes pour rendre compte de ces structures.
Or, il nous semble au contraire qu'elles doivent être complétées par une caractérisation syntaxique, en conformité avec les règles sémantiques que l'on vient de présenter. A cela, deux raisons essentielles :

- N'importe quelle combinaison de termes n'est pas possible, et il y a un certain ordonnancement à respecter ; cet ordonnancement reproduit (avec une certaine flexibilité d'ailleurs) l'ordre des opérations observé dans le cadre des flux.
- En matière d'extraction d'information, qui constitue effectivement l'objectif de ce travail, il est nécessaire de disposer d'une modélisation de la structure. En effet, dans ce cadre, puisque l'on ne dispose que du support et pas de l'ensemble de l'espace de l'activité, il est fondamental d'avoir à disposition cette représentation formelle de la structure. C'est elle qui permet de reconstituer les opérations.

En définitive, la caractérisation de la structure par une syntaxe et des primitives constitue un moyen pour disposer d'une modélisation autonome, qui n'implique pas une explicitation du contexte du support.

Nous construirons cette représentation de la structure d'information au travers de deux dimensions dont nous voudrions maintenant expliciter les rôles réciproques : les opérations syntaxiques et les primitives sémantiques.

Les flux d'information représentent comment les contenus sont obtenus et transmis, mais pas la structure linguistique qui clôt ces contenus. Comment se fait-il que l'expression peut disposer d'une certaine autonomie pour son interprétation ?
L'entrée syntaxique choisie est celle de grammaires catégorielles simples qui ont comme intérêt de reproduire, sur les relations entre les entités symboliques, les opérations identifiées dans le cadre de l'activité. On présente donc des règles applicatives. Les règles grammaticales sont contraintes par les règles des opérations. Le lien qui permet le passage de la syntaxe à la sémantique est le fondement prédicatif des deux outils d'analyse.

Néanmoins, cette représentation syntaxique n'a pas une valeur en soi ; elle a pour vocation de caractériser une dimension de la structure et sert d'outils pour construire le modèle de la structure, qui reste à fondements sémantiques. Nous présentons maintenant l'hypothèse fondamentale de notre travail relatif à la structure, à savoir la caractérisation de primitives.

Définition de primitives pour la structure d'information.
Les primitives représentent les conséquences des flux sur les choix et l'organisation des unités lexicales. Par conséquent les primitives caractérisent des catégories au sens linguistique et non des types. Cela évite toute confusion avec les types sémantiques associés aux flux.

En termes de théorie de l'information, ces primitives représentent des contenus, à savoir les ensembles de possibilités de choix dans une situation définie. Les primitives représentent donc des résultats d'opérations régulières, qui se décrivent à la fois comme des choix informationnels et des inférences contraintes entre deux situations.

Les primitives sont justifiées par un postulat cognitif lié à l'information et à la double structuration des mondes de référence (les objets du monde d'une part, les opérations permettant de produire de l'information symbolique à propos de ces objets). En effet, on caractérise l'information comme une fonction requérant la mise en relation d'ensembles lexicaux hétérogènes, par un flux et par des opérations réalisées dans le monde à propos d'un certain état d'une autre partie du monde. (Les ensembles lexicaux doivent satisfaire des propriétés sémantiques définies par les flux et qui portent sur les dimensions interprétatives). Les composants de la structure ne sont donc pas définis par de seuls éléments symboliques, ou tout du moins ceux-ci constituent les indices d'autres opérations que la référence à des objets du monde.
On peut caractériser ces lexiques en utilisant des outils terminologiques, ou des classes distributionnelles, voire encore des catégories comme celles utilisées par FrameNet[211]. On montre à ce moment leur distance sémantique : quel lien y a-t-il entre un processus corporel constant, une unité de temps, des intégrales, des relations entre deux substances organiques ? En soi, ces lexiques sont très éloignés. Or, ils sont tous présents dans les expressions qui nous concernent. C'est ce que l'on aimerait expliquer à propos de la structure d'information à l'aide des primitives.

Les primitives sont définies par un niveau d'abstraction supérieur par rapport aux unités linguistiques. Une telle définition permet d'envisager des composants de la structure qui ne seraient pas nécessairement des unités lexicales caractérisées à partir du langage naturel mais des unités lexicales inscrites dans des langages contrôlés. La structure d'information peut ainsi donner sens à des processus comme l'alignement d'ontologie ou de thésaurus. (En effet, si les langages du web (OWL, SKOS en premier lieu) constituent des règles d'écriture, ils n'indiquent pas le sens que peuvent porter les expressions qu'ils génèrent). Ce niveau d'abstraction a un autre enjeu : il permet d'intégrer à la fois les dimensions syntaxiques et sémantiques.
Nous pouvons schématiser ainsi cette construction :

Niveau de représentation	Flux d'information	Syntaxe	Prédication
Objet représenté	Contraintes informationnelles sur les entités linguistiques.	Trace des opérations informationnelles sur la grammaire	Structure d'information

Structure d'information et théorie des situations.
La caractérisation de la structure d'information que l'on propose s'inscrit à l'intérieur du cadre de la théorie de l'information et des situations. Reprenons F. Dretske : il considère l'information comme une succession d'opérations permettant d'obtenir des entités discrètes à

[211] FramNet constitue une base de données lexicale organisée selon les principes des « frames » ou cadres, définis par Charles Fillmore. Ces cadres sont représentés à l'aide de traits de signification spécifiques à ce cadre, Ces traits sont également associés à des contextes particuliers, lesquels sont également répertoriés dans le cadre. Enfin, on associe des catégories grammaticales à cette structuration sémantique.

partir de phénomènes continus. Dans ce cadre, les opérations sont récursives et ne sont pas limitées. Or, le principe d'économie présenté par G. Chaitin permet de caractériser la saturation (dans le sens d'une complétude de la prédication) de la suite de symboles.
C'est sur ce fait que se fonde par ailleurs K. Devlin (op. cit., pp. 17-18) pour distinguer une problématique de l'information d'une problématique des flux. C'est cette saturation que l'on explique maintenant.

Conformément aux principes définis par F. Dretske, la structure d'information constitue l'achèvement d'un processus de discrétisation, l'information possède un contenu limité par la situation décrite à laquelle elle fait référence.
En ce sens, augmenter l'information consistera à réintroduire d'autres dimensions continues, comme les dimensions visuelles du texte. Dans ce cadre, on peut rendre compte de l'image, à savoir le fait qu'une analogie permet, dans une situation identifiée, de véhiculer un contenu plus complexe parce que non discrétisé. Ce sera d'ailleurs un objectif des propositions de J. Barwise & G. Allwein[171].

L'explication de cette clôture reste néanmoins ouverte : constitue-t-elle un phénomène propre d'information, ou au contraire relatif au langage, ou enfin lié à l'activité, donc à la culture commune ?
La réponse à une telle question n'est pas directe, au sens où elle requiert différents niveaux de relation entre d'une part les questions d'opérations de flux et de propriétés propres des unités du langage. Surtout, elle requiert des outils théoriques et méthodologiques propres aux sciences cognitives.

Nous commencerons par présenter les formats et limites fixés à la structure d'information (sans néanmoins fournir une explication satisfaisante de ce fait). Ces formats ont pour objectif de montrer les choix opérés à l'intérieur des différentes approches de la structure d'information, et ensuite de valider en quoi ces choix permettent de caractériser des constituants. Enfin, nous présenterons les outils de représentation, à savoir l'adoption d'un dérivé de la théorie des types, la théorie des traits.

5.2.1. Format et limites pour la structure d'information.

On définit la structure d'information comme un ensemble de symboles aboutissant à la discrétisation d'un état dans le monde.
Rappelons que la structure d'information constitue une problématique récurrente à la fois en informatique et en linguistique (et pas simplement en linguistique formelle). Nous commencerons à présenter les travaux à fondement linguistique avant de spécifier comment dans notre cadre théorique cette question se pose en enfin, nous présenterons l'apport des analyses précédentes à notre propre caractérisation de la structure d'information.
Rappelons que l'enjeu de la caractérisation de ces structures consiste à proposer un modèle de référence pour l'extraction d'information. Nous dissocions les questions de modélisation des structures d'information de celles relatives aux outils d'extraction, et qui requièrent d'abord des méthodes et des connaissances statistiques. Ces derniers outils ne proposent pas de modèle de la structure d'information : ils sont en quelque sorte au service des modèles.

Définitions et perspectives à propos de la structure d'information.
Tout d'abord, on explore les définitions de la structure d'information, notamment en ce qui concerne la distinction et complémentarité avec la structure prédicative.

La définition de base serait la suivante : « il s'agirait du contenu nouveau véhiculée par une expression, à savoir comment elle est organisée de façon à faciliter l'addition de ce contenu à l'intérieur des connaissances de celui qui écoute ». [kim98-information]
C'est donc une problématique transversale, qui concerne à la fois la syntaxe, la sémantique et la connexion que l'on peut établir entre linguistique et représentation des connaissances. (C'est une question essentielle puisqu'elle permet d'entrevoir, l'automatisation de procédure d'extraction de connaissance et de structuration de connaissances sur une seule base textuelle).

La structure d'information se distingue fortement de la structure lexicale par l'opposition entre un lien ALTERNATIF et un lien CONSTANT. (Le lien alternatif caractérise des relations binaires de présence ou d'absence d'un trait ou d'une entité, le lien constant caractérisant des traits définitoires donc constants).

En premier lieu, la recherche d'un format de représentation pour l'information constitue en propre un champ de recherche, dont l'aboutissement est la caractérisation de la structure d'information dans ses dimensions syntaxiques et sémantiques. Cela en la distinguant et l'affiliant à la prédication et la structure argumentative. On aboutit alors à des modèles alternatifs à ceux fondées sur des catégories linguistiques pour identifier des structures linguistiques. Les travaux de M. Steedman (op. cit.) sont un excellent exemple de cette perspective.
En second lieu, il s'agit d'élaborer un certain nombre de concepts qui serviront alors à caractériser comment les discours sont construits afin de permettre la communication d'une information nouvelle dans des connaissances établies. A ce moment-là, il s'agit de stratégies de discours. Alors, la structure d'information est caractérisée dans le cadre des visées communicationnelles d'un discours et de la façon dont on envisage l'apport de connaissances auprès du lecteur. La synthèse présentée par Manfred Krifka[172] s'inscrit totalement dans cette perspective.

Cette distinction renvoie plus à des questions de priorité : la seconde est fondée sur l'articulation de la signification aux connaissances, alors que la première vise à associer un niveau d'analyse et de génération des discours à partir d'une dimension d'analyse distincte de la phrase.

La structure d'information permet surtout d'illustrer :

- la substitution de catégories grammaticales à des catégories structurantes de la signification dans le cadre d'une modélisation des expressions linguistiques (orales autant qu'écrites).
- L'élaboration de modèle de discours fondés sur des stratégies d'apport de connaissances sur des thèmes connus.

Suivant la première approche, les dimensions et constituants de la structure servent à la caractérisation d'articulations distinctes de la phrase (clause, expression notamment).
En syntaxe ce principe se traduit par des opérations génératives fondées sur les capacités de contraintes de lexiques (et non sur des catégories préétablies comme le nom, l'adjectif, le verbe…).
En sémantique cela se traduit par la distinction à la prédication et à la structure argumentative des phrases en établissant des structures d'univers complexes : on présuppose qu'une partie de l'expression est supposée connue, l'autre étant dérivée : renouvellement de la représentation de la prédication.

La structure d'information sert dans la seconde approche de modèle pour expliquer les discours, représenter la sémantique et la pragmatique des dynamiques textuelles.
Plus précisément, dans le discours, on caractérise des cadres pour représenter l'actualisation des connaissances.
Ainsi, les structures des lexiques apparaissent comme des ensembles d'éventualités, ce qui permet la représentation de situations et d'actualisations.
La pertinence pragmatique se marque essentiellement autour du marqueur intonatif : celui-ci spécifie ce qui est (1) présupposé, (2) supposé connu et (3) nouveau dans une expression (cela dans l'esprit du locuteur, pour un certain récepteur).
Le fondement de l'observation de la structure est le couple question-réponse. La structure possède alors une dimension inter-phrastique.

La structure d'information considérée comme un format d'expression. Marqueurs de surface.
Au départ, on n'a pas de format exact pour caractériser la structure d'information, sachant que le format de la phrase n'est pas pertinent. (Tous les formats peuvent être envisagés sous la condition d'un transfert effectif de l'information, donc d'une interprétation identifiant en propre un apport informationnel).
Dès lors, on peut effectivement considérer le texte comme une autre dimension possible.

Par ailleurs, pour étudier la structure d'information, il apparaît fondamental de disposer de critères d'identification spécifiques et incontestables. L'idée a consisté à considérer que l'intonation pouvait jouer ce rôle. Le fondement de la structure est non logique mais relatif à la perception : c'est une structure de surface.
L'intérêt de l'intonation consiste à disposer de marques non symboliques dans la caractérisation de la signification. Une même proposition peut avoir des significations distinctes en fonction de l'intonation ; cela entraîne l'étude de la distinction de focalisation, présuppositions et attitudes propositionnelles.
Exemple : « Les copies doivent être NOTEES ». « Les COPIES doivent être notées ».

L'intonation constitue le marqueur considéré comme incontestable de la structure et de ce qu'elle permet d'apporter à l'analyse de certains problèmes (distinctions de focalisation, présuppositions, attitudes propositionnelles entre entités dans le modèle de discours).

La dimension de l'intonation laisse voir une structure de la phrase qui ne correspond pas à la structure syntaxique, ou mieux, une structure syntaxique peut correspondre à des structures différentes.
Si l'on articule intonation et syntaxe, on obtient un cadre de travail unique : comment l'expression est diffusée. La structure d'information devient homomorphe à la structure intonative ET respecte la syntaxe de surface.
La sémantique associée s'inscrit dans la perception (R. Jackendoff), soit il s'agit de prédication (une écriture particulière de la prédication).
« JOHN a vu la pièce hier »
« HIER, John a vu la pièce »
« LA PIECE, John l'a vu hier »
(Le thème est à gauche, le commentaire à droite : on ne parle pas du même objet[212]).

[212][212] Cette présentation ne fonctionne que si l'on ne prend pas en compte certaines propriétés sémantiques et pragmatiques de « hier »…

Dans cet exemple, on a la même proposition, donc le même contenu sémantique. Ce qui diffère, c'est la construction de l'information. Le thème est à gauche, le commentaire à droite.

(On pourrait croire à une remise en cause des principes fondateurs de la sémantique ; en apparence, dans ces théories, il n'en est rien. Il s'agit bien plus de construire un autre niveau de représentation que celui qui est utilisé en sémantique).

L'intonation comme indice matériel de la structure a une influence importante sur la définition de l'information : alors que la prédication a un fondement logique, l'intonation, qui ne repose pas sur un raisonnement a au contraire un fondement psychologique.
C'est quelque chose qui serait plus proche de la perception que du raisonnement.

Par conséquent, la structure d'information se caractérise d'abord par la perception de différences intonatives. Cette dimension oriente le travail vers les questions de cognition (d'où par exemple les propositions de R. Jackendoff[173], pour lequel la sémantique est à fondement psychologique).

On caractérise de cette façon une structure d'information du discours, en considérant le couple question-réponse comme un moyen pour discerner les constituants de la structure. On se base donc sur la présupposition.
C'est en ce sens aussi que la structure d'information est un phénomène du discours : une structure est expliquée dans le couple question-réponse.
La dimension non linguistique de la structure d'information est essentielle pour la caractérisation de la dimension représentationnelle (s_3) associée aux flux. Elle est importante également dans le cadre de l'extraction parce qu'elle permet de définir des indices formels d'identification de la structure.

Articulations fondamentales de la structure d'information.
Les concepts fondamentaux qui caractérisent les structures sont des articulations distinctes des articulations linguistiques. Ce sont des primitives de l'information:[174]
[Ground-focus] (fond-focalisation) ou [topic-comment] (thème-commentaire).

L'articulation topic-comment (thème-commentaire) a un fondement prédicatif ; il s'agit d'une part de caractériser le cadre de référence de la phrase mais aussi du discours, et ensuite de mettre en place les commentaires, considérés comme des constituants immédiats.

« E est le thème de la phrase S, ssi l'énonciateur cherche à augmenter la connaissance du récepteur à propos de, répond à une question à propos de, demande d'agir en lien à E ».

Le point de départ de la clause est un message, qui est un cadre de référence la phrase ; le thème comporte alors un rôle d'ancrage fondamental.

Néanmoins, certaines phrases peuvent fonctionner sans thème (seulement en commentaire) ; ce sont les phrases « présentationnelles » ou nouvelles phrases.
« L'écran est MORT ». (La phrase contient un rapport thème-commentaire).
« L'ECRAN est mort ». (Celle-ci ne contient qu'un commentaire).

Généralement, la position thème est associée à la position gauche dans la phrase.
L'élément marqué (ou rhème) est celui qui est choisi à l'intérieur d'un ensemble de possibles.

Dans l'articulation ground-focus (fond-focalisation), le focus contient un élément saillant intonatif, à un moment donné du discours. (La saillance intonative s'explique aussi par des causes purement linguistiques). Cette articulation est fondée sur les thèmes non-marqués ou phrases à seul commentaire. Le focus est ce dont on parle dans le discours (et qui est saillant). Le focus n'est pas l'information nouvelle (pronominalisation). Ce sont par exemple « les copies » ou « elles » dans une seconde expression.
Dans une acception propositionnelle et discursive, être en focus constitue un statut référentiel concernant les entités de discours et non une notion relative à l'informativité des propositions. Le référent en focus est un référent connu de l'interlocuteur et peut être référé par une forme pronominale.

A ce moment-là, le « ground » est ce qui est présupposé, donc à un niveau antérieur de la structuration. L'arrière fond constitue un certain nombre de propositions présupposées (ou de situations communes) « les x doivent être notées ». A ce moment-là ce qui permet d'analyser la distinction entre le focus et le « ground », c'est le lien entre ce qui est asserté (au sens de R. Stalnecker[175] : réduction de l'ensemble des contextes) et ce qui ne l'est pas (et donc qui constitue le « ground »). Le « ground » renvoie systématiquement à un construit antérieur dans le discours. Ainsi, la focalisation constitue un marqueur d'une restriction. Elle identifie un objet ou une action saillants (dans l'arrière-fond).

Exemple : « Ma voiture n'a pas démarré ce matin. Elle a passé la nuit dehors ».
Le « focus » est la voiture, le « ground » est alors la proposition selon laquelle « j'ai une voiture. ».
L'hypothèse à la base est que seul ce qui est nouveau est exprimé (principe d'informativité).

Thème/rhème constituent une structure binaire de l'expression. Arrière-fond / focalisation sont des opérateurs binaires (ce à quoi le locuteur nous demande de nous intéresser dans un certain cadre –l'arrière fond).

(Marie	ADMIRE)	(la femme qui	DIRIGEAIT	l'orchestre)
Arr.fd.	Focus	Arr.fd.	Focus	Arr.fd.
Thème		Rhème		

L'articulation thème/rhème permet d'utiliser une logique modale : le thème introduit un ensemble contextuel d'objets possibles ; le rhème est le choix d'un objet de l'ensemble.
L'articulation arrière-fond/focus est une caractérisation de la référence : parmi ce qui peut se réaliser (ou est réalisé) dans un univers, on opère le choix d'une entité pour être focalisée.
Ces articulations permettent une augmentation de la description de la signification par rapport au modèle traditionnel de la prédication.

L'école de Prague (et en particulier J. Peregrin[176]) propose une intégration maximale de la distinction thème - rhème dans le système grammatical. Il s'agit d'une approche fonctionnelle : le langage est un outil de communication et la structure d'information est essentielle à la fois pour le système de la langue et le processus de communication.

A la base, la grammaire a comme vocation de caractériser de la façon la plus précise la distinction entre topic et focus, en considérant un lien entre l'intonation et l'ordre des mots.

1. Le premier outil utilisé pour cela est un ensemble d'arbres non terminaux caractérisant des structures de dépendance fondées sur le verbe. (Structure tecto-grammaticale)

2. Dans ce cadre, chaque item est à la fois contextuellement limité (l'expression n'apparaît qu'avec sa précédente dans le discours) ou contextuellement non-limité. La limitation contextuelle ne caractérise pas seulement les éléments utilisés immédiatement avant, mais ceux qui sont impliqués par le co-texte verbal, la situation de discours et les pré-requis culturels portés par l'interaction.
3. L'ordre par défaut des items autour du verbe (l'ordre des rôles thématiques et des adverbes) est considéré comme fixe pour un langage donné. C'est l'ordre systémique. Cet ordre, combiné avec l'ordre d'autres items ne dépendant pas du verbe, est modifié dans une expression concrète, donc l'ordre résultant des items de la structure tecto-grammaticale est celui de la dynamique communicationnelle. Cet ordre des items contextuellement limités, d'un même niveau, est déterminé par la stratégie de discours plus que par la grammaire. Pour les items non contextuellement limités (dépendant de la même tête), cet ordre est en accord avec l'ordre systématique. De cette façon, un item est moins dynamique que sa tête ssi l'item dépendant est limité.
4. Le principal élément dynamique de la phrase est le thème propre.
5. Tous les items contextuellement limités dépendant du verbe central et tous ceux qui dépendent d'aux et dépendant du verbe central si celui-ci est contextuellement limité constituent le focus. Alors, la classification thème/focus devient exhaustive : tout élément de la phrase est soit thème, soit focus.

L'outil qui permet de formaliser ces propositions est une logique intensionnelle (Transparent Intensional Logic).

Dans ce cadre, « John marche » et « JOHN marche » ne présentent pas la même signification, ce que l'on peut représenter respectivement de la façon suivante :
MARCHE(JOHN) , λf.f(JOHN)(MARCHE).

On peut tirer quelques conclusions d'étape à propos des propositions de J. Peregrin (op. cit.) :

- la distinction de la sémantique de la structure d'information avec la prédication traditionnelle serait marquée par l'inversion possible du rapport sujet –prédicat, vu alors comme objet – concept.
- Une proposition n'est alors plus associée à une phrase : chacune possède deux propositions : l'une, présupposée, renvoie au sujet, la seconde, posée, à l'articulation du sujet et du prédicat.

Interprétation de la structure d'information.
Dans le cadre d'une sémantique, l'idée consiste à traiter de façon spécifique le sujet (considéré comme une proposition présupposée). Parce qu'il constitue une présupposition, le sujet n'est pas soumis aux valeurs de vérité ; il est considéré comme un objet préexistant.

On caractérise alors trois dimensions pour l'expression « X » ; ||X|| désigne l'extension de l'expression « X » :

||X|| : valeurs de vérité si X : phrase.

||X|| : individu si X : un terme.

||X|| : classe d'individus si X : prédicat unaire.

Toute expression X est associée à une proposition dont l'extension est $|X|$ (comprise comme une présupposition associée à X).

$|X|$ = ||X|| si X est une phrase

= ||∃y.y = X|| si X est un terme

= ||∃y.X(y)|| si X est un prédicat unaire

De telles règles permettent ensuite de représenter une formule P{S} où P représente une prédication sur le sujet S.

||P{S}|| = T ssi |S| = T & ||P{S}|| = T

= F ssi |S| = T & ||P{S}|| = F

= 0 ssi |S| = F

Cette représentation donne la possibilité de caractériser plusieurs niveaux de proposition dans l'expression.
Le tour de force consiste à considérer que chaque niveau est le résultat d'opérations prédicatives particulières et donc, qu'il est possible de caractériser l'information comme étant associée à chacun de ces niveaux.
Les différentes formes d'opérations permettent de montrer exactement où se situe la focalisation, à savoir de quelle partie de l'expression pourra être considérée comme subissant un traitement particulier de choix préalable à l'expression globale.
Ainsi : « John marche » : marche{John}.

« JOHN marche » : λf.f(John){marche}.

Le problème à résoudre, c'est de rendre compte non pas de : « un homme marche », mais de « l'homme qui marche est l'unique entité qui marche ». L'unique est caractérisé parmi l'ensemble des possibilités déterminées par le contexte.
Toute expression peut être caractérisée par un ensemble d'alternatives (ALT(X)) et P !(T) où P ! désigne le prédicat unique P de T.

||X||i : extension de X sous l'interprétation I.

I[P/p] : interprétation selon laquelle on assigne p à P. [||P||i = p]

Les alternatives ALT(p) ⊆ D sont inclues dans le domaine.
ALT(p) = D ; l'ensemble des alternatives est équivalent au domaine.
λf.f(S) = P : P est la seule propriété instanciée par S.
P !S signifie que P est la seule des classes restreintes de propriétés à être instanciée par S.

Comme n'importe quelle propriété appartient à plus d'une seule classe de propriétés, on caractérise ce qui compte dans une telle alternative : n'importe quelle classe de toutes les propriétés d'un individu.
N'importe quelle classe incluant un individu est considérée comme une propriété de l'individu. Donc, n'importe quel individu instancie une vaste quantité de propriétés.

Ainsi, ALT peut être considéré comme un nouvel élément qui peut être modifié par les expressions à venir. Cette notion permet deux directions :

- la problématique des stocks de connaissances[177],
- les sémantiques dynamiques.

Les questions de falsification et de d'attribution inappropriée ;
P !(S) est faux alors que P(S) ne l'est pas : erreur d'exhaustivité : « l'allemand est parlé en Autriche ».
P(S) est faux alors que P !(S) ne l'est pas : « Chirac est l'actuel roi de France ».

C'est une erreur de présupposition.
P(S) est faux : erreur de thème.

Perspectives ouvertes vers les lexiques et les connaissances.
On a vu que l'information constituait un objet qui pouvait particulièrement bien relayer les problématiques dynamiques développées en sémantique, puisqu'elle propose d'identifier des articulations et des processus distincts des modèles traditionnels de la phrase. Ces articulations peuvent alors servir à construire des objets d'analyse fonctionnels et intégrant les dimensions du discours, voire même du contexte.
La question de l'information renouvelle les approches sémantiques en considérant trois angles distincts :

- soit on considère que l'information s'inscrit dans les dynamique par le biais d'opérateurs de discours, et donc on étudie comment ces entités lexicales structurent les enchaînements entre phrases, ce qui aboutit alors aux questions de discours, voire même de rhétorique. C'est grossièrement la position des TAG (Tree Adjunct Grammar).
- Soit on considère d'un point de vue plus sémantique, que la segmentation opérée par la structure, fournit un cadre pour isoler des unités ou des ensembles d'unités décrivant un état. L'information caractérise alors la succession de ces états. C'est le cadre des sémantiques dynamiques.
- Soit enfin on considère que les processus argumentatifs inscrits dans la génération des discours se décrivent par des opérations informationnelles. Ces dernières assurent la dynamique propre du discours. (C'est le cadre par exemple des grammaires catégorielles).

Par ailleurs, ces travaux sont orientés vers la considération du lexique comme unité de base.

Néanmoins, chacune des deux approches précédentes sont incapables de capturer certains des articulations fondamentales de l'information. Par exemple, si « Marie donne quelque chose à Harry » constitue le fond, et le focus serait « une disquette », on peut rendre compte de « Mary donne [$_F$ une disquette] à Harry ».
Par compte, on ne peut pas rendre compte de « A Harry, Mary donne [$_F$ une disquette].

Par ailleurs, thème -rhème ne rend pas compte de la spécification suivante :
« Maire donne une DISQUETTE à Harry »
« Marie donne une disquette à HARRY ».

Ce genre de problème est résolu de deux façons :

- soit on accumule les deux perspectives, ce que l'on verra avec M.A.K. Halliday (op. cit.).
- soit on considère une tripartition qui aurait comme objectif de montrer où se situe l'information dans la phrase et comment elle s'inscrit dans les connaissances de l'interlocuteur. Ce sera la perspective de E. Valduvi (op. cit.).

Evidemment, les différents niveaux sont complémentaires. Simplement, l'une oriente le travail vers les modèles de discours, l'autre vers la représentation des connaissances.

Néanmoins, le premier et principal intérêt de ces travaux, c'est de caractériser l'information que porte une expression comme un autre niveau de représentation de la signification que la prédication et la proposition.

Modèles discursifs de la structure d'information.
Une seconde entrée pour caractériser la structure d'information est de partir des fonctions du discours ; la structure d'information serait une unité de réalisation des fonctions du discours. On se libère de cette façon des questions de syntaxe et de sémantique associées à l'expression. Par ailleurs, ces questions permettent d'accentuer la dimension dynamique associée à la structure d'information.

En ce sens, les distinctions présentées sont d'abord à **fondement rhétorique**. Ensuite, elles sont **fonctionnelles**, notamment dans la perspective de M.A.K. Halliday.
Enfin, elles permettent de construire un **modèle multi-entrées**, fondé sur des types et correspondant à différents niveaux de description, comme celui de E. Valduvi.

Les notions de clauses permettent, entre autre parce que chaque clause est associée à une autre via un connecteur, d'observer des constructions discursives : à chaque connecteur est associée une règle permettant de lier une clause à la suivante. Ce qui est alors essentiel, c'est le fait que l'information constitue une unité à l'intérieur d'un processus rhétorique plus vaste. De cette façon, il devient cohérent d'envisager la structure du texte à partir d'une dimension rhétorique fonde sur l'information.

On propose les schématisations suivantes, pour rendre compte de la construction dynamique des discours.
C'est l'idée de progression thématique :T = thème, R = rhème
(1, .., n) : expressions successives dans un discours.

$T_1 \rightarrow R_1$
$\downarrow$
$T_2 (= R_1) \rightarrow R_2$
$\downarrow$
$T_3 (= R_2) \rightarrow R_3$

Exemple : « [T1]Atos Origin [R1] a signé un accord de partenariat avec Inkra Networks en France. [T_2 (= R_1)] Aux termes de cet accord, [R2] la SSII française intégrera à son portefeuille de solutions la technologie Virtual Service Switch (commutateur de services virtualisés) d'Inkra Networks ».

(La structure d'un discours peut être déterminée par la structure d'information précédant l'expression en cours et la structure d'information de l'expression en cours).
Elles expliquent par exemple :
(1) « Les chats mangent les souris ; or les souris sont rapides et agiles. »
(2) « Les chats mangent les souris ; (or) les chats sont rapides et agiles ».
Il n'y a pas de dynamique de discours dans le second cas.

Or, comment différentes structures sont impliquées ?
On considère qu'une structure de discours inclut ou pas une structure d'information :
- elle l'inclut dans le cas 1,
- elle le rejette dans le cas 2.

Les choses deviennent plus compliquées avec l'exemple suivant :
« Certes, Toto a épousé Bertha. Néanmoins, il n'a pas hérité d'un euro ».

Deux analyses sont possibles :
- soit la structure de discours n'inclut pas l'IS2, alors quel est le rôle du subordinateur ?
- soit elle l'inclut, mais alors quel est le lien entre les deux IS ?

L'hypothèse alors, c'est de distinguer les domaines de la structure de discours de ceux de la structure d'information. Les premiers constitueraient des représentations sémantiques (permettant notamment la possibilité de reformulations de surface), les seconds des précisions à l'intérieur du domaine, notamment par le biais des choix alternatifs. Sinon, on est obligé de considérer que les domaines de structure d'information sont limités par ceux de la structure de discours. (On aurait d'abord la clause, et dans cette clause, le développement d'une structure d'information (correspondant donc aux aspects intonatifs).
Autrement dit, les choix alternatifs sont fondés sur les cadres limités fournis par la sémantique de l'expression.
En fait, deux arguments militent en faveur de la distinction entre les deux procédures :
- la possibilité de reformulation (qui entraîne une seule expression et la disparition du connecteur)
- le fait que l'on puisse produire des allégations à partir d'une seule structure de discours, que vient contredire la proposition suivante (marquée par un connecteur).

En fait, ce sont les différentes IS qui distinguent les différentes présuppositions, et non les clauses. La signification d'une clause peut être identique ; ce qui la distingue, c'est le présupposé qui lui est attaché, donc la façon dont il peut être déduit par les marqueurs informationnels.

La question est alors fondamentalement liée à la pragmatique de la structure : quels sont les contextes dans lesquels une expression avec une focalisation particulière est satisfaite.
La structure d'information est alors une structure fondée sur l'intention (en fonction des objectifs du discours et non des expressions ou séquences d'expressions utilisées pour la présenter).
Craige Roberts[178] se fonde alors sur la caractérisation du contexte proposée par R. Stalnecker, ce qui permet alors de motiver pragmatiquement la structure.

Théories fonctionnelles du discours et structure d'information.
La dichotomie de la structure d'information, fondée sur une base psychologique et associée à la linéarité de l'expression, peut également être considérée comme la dimension la plus réduite de l'observation des textes. En ce sens, en présupposé, le texte constitue un outil de médiation dans le système d'information, susceptible d'accueillir des informations. L'analyse permet de déduire qu'il s'agit d'un format permettant de spécifier les contenus informationnels qu'il peut accueillir (donc qu'il limite).

Que l'on soit chez M.A.K. Halliday et l'Ecole de PRAGUE[179], et plus loin les DRT et SDRT, la définition de la structure d'information implique de caractériser des niveaux de contexte qui peuvent être sommairement représentées ainsi :
- limitations contextuelles,
- dynamisme communicationnel,
- ordre des mots.

Avant de présenter la structure d'information telle que définie par M.A.K. Halliday[180], on essaiera de caractériser d'abord l'approche:

- L'approche de M.A.K. Halliday est systémique et non structurale. La signification est caractérisée à partir du contexte. A l'opposé, la définition structurale implique de caractériser le langage comme l'ensemble de toutes les phrases possibles.
- Le langage est considéré comme une ressource, donc sa description est celle de choix (les choix que font les utilisateurs en contexte et avec des objectifs).
- Ces objectifs permettent de caractériser une stratégie entraînant une certaine flexibilité entre les objectifs recherchés et les ressources linguistiques pour un certain potentiel de signification.

Pour M.A.K. Halliday, il n'y aurait qu'un seul niveau d'analyse: la lexico-grammaire, qui groupe à la fois la morphologie, le lexique et la syntaxe.
Une grammaire ne peut être modélisée comme de nouvelles phrases faites sur un stock limité de mots dans lequel il n'y aurait pas de répétitions. Au contraire, on va caractériser des "chunks", comme "comment allez-vous", qui constituent des ensembles complets, comme des unités. Plus globalement, il propose de considérer qu'il existe un certain nombre de structures fondées sur des propriétés lexicales, et dont la dimension excède celle du mot. On constitue ainsi des patrons. Ces patrons apparaissent avec plus ou moins de probabilité selon les contextes.

La structure d'information est définie par M.A.K. Halliday est un niveau de représentation de la phrase spécifié par un bloc où un message est isolé par l'intonation ou l'ordre des mots (la partie la + informative suit la partie la - informative). Il emprunte cette définition à l'école de Prague (dichotomie "thème/rhème"). Cette dichotomie peut être observée autant au niveau de l'intonation que des structures de texte plus complexes. (Cf. "The Prague school approach of the IS")

Le thème chez M.A.K. Halliday est une fonction textuelle: c'est le point de départ du message, la limite à partir de laquelle le propos peut être tenu.

M.A.K. Halliday propose en fait de considérer la structure d'information comme la corrélation entre des unités d'information et un phrasé intonatif.
La partition est fondée sur le contraste d'informativité. Néanmoins, il y a contradiction dans la définition de la structure saillante :

- expression : le phénomène nouveau est le plus informatif, et donc constitue l'élément saillant.
- Dans le discours l'ensemble ce qui est toujours établi et activé.

Les paires de la structure désignent des phénomènes très différents :

- soit "à propos de " : thème - rhème ; topique - commentaire.
- Soit ancrage discursif : présupposition - focus ; background - focus ; proposition ouverte - corpus ; ancien/donné - nouveau.

Systématiquement, la structure d'information ne réfère jamais à la structure syntaxique. (Les premiers modèles de la structure sujet - prédicat ne sont pas présents dans toutes les langues, et elle ne possède pas de définition grammaticale claire (si ce n'est P : SN, SV)).

A contrario des précédentes propositions, M.A.K. Halliday postule l'indépendance de l'information au système grammatical. Il postule la partition interne de la phrase par une structure thématique (thème, rhème). La structure définit l'ordre linéaire des unités informationnelles.

L'organisation interne de chaque partition est composée de données qui ne sont donc pas étudiées.

La structure thématique est organisée par "à propos de…" ; thème : l'objet à propos duquel l'objet est "à propos de…". Rhème : ce qui est dit.
La structure interne d'une unité d'information ; ses éléments sont marqués par leur ancrage discursif, à savoir que le focus d'information est l'information nouvelle apportée. (L'informativité d'une information est marquée dans le focus).
L'informativité : nouveau matériel apporté, et qui n'est pas déjà observable dans le discours.

La structure proposée par M.A.K. Halliday est la suivante :
INTONATION corrélée UNITES D'INFORMATION laquelle est organisée par une STRUCTURE THEMATIQUE.
La nouvelle information est marquée soit :

- une énonciation particulière (intonation)
- contraire à une alternative posée
- recomposée par une question présupposée (QUE.. ?)

Elle peut être présentée de la façon suivante : "A piece of discourse consists of a linear succession of message blocks, the information units, realized by tonality. Each information unit is the point of origin for the choice of information focus, by which one element is selected as focal. In the unmarked case the new is, or includes, the final lexical item, but the focus can appear at any point in the information unit. New information is non-derivable, whereas given information is recoverable anaphorically or situationally. The information structure provides the framework within which all of these choices are exercised."[181]

M.A.K. Halliday distingue trois types de focalisation (en lien à la précédente définition de la nouveauté) :

1. Informativité : assertion, addition
2. Contrastivité : assertion, contradiction
3. Focus par question : question, paire de réponse.

Enfin, on doit mentionner L. W. Chafe[182] : il s'intéresse à la façon dont le discours est structuré en fonction des croyances de celui qui parle bien plus que du contenu sémantique des expressions. Le "package" d'information caractérise comment le message est envoyé et non le message lui-même.
L. W. Chafe (repris dans E. Valduvi, *Information packaging, a survey*, op. cit.) considère que l'information est une suite d'instructions sur une expression commune. Ce sont ces instructions qui font la structure d'information. Cette conception est opposée à celle de J. Barwise & J. Perry.
Par contre, L. W. Chafe considère les structures comme des structurations des différentes propositions contenues dans les expressions, structurations gouvernées par le locuteur (en fonction de ses croyances à propos des connaissances de son interlocuteur). (E. Valduvi[183], *The Informational Component*, p.12/22)

A propos du nom, il constate les différents usages suivants :
(a) donné ou nouveau, (b), focalisation de contraste, (c) défini ou indéfini, (d) sujet (e) thème, (f) représente le point de vue de celui qui parle.

L. W. Chafe utilise la notion de donnée comme étant ce que celui qui parle assume comme étant présent dans la conscience de celui auquel il s'adresse au moment où il parle.
L'information nouvelle est ce qui est assumé comme n'étant pas dans la conscience de celui auquel on s'adresse.
Le lien à aux connaissances est alors essentiel.

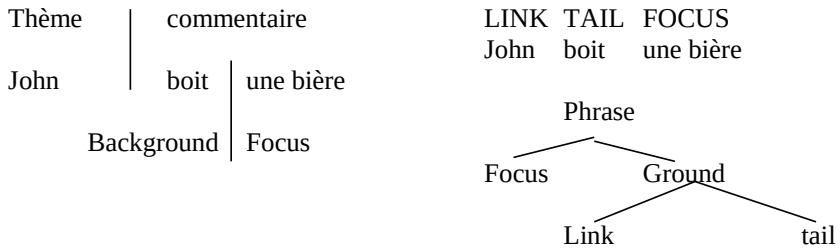
Vers des modèles tripartites.
E. Valduvi (*The Informational Component*, op. cit.) a particulièrement développé la trajectoire depuis les constructions rhétoriques et de connaissance initiales vers la façon dont les entités sont distribuées à l'intérieur d'une structure syntaxique.

Le propos de E. Valduvi consiste en partie à expliciter une trajectoire depuis des phénomènes qui ne peuvent formalisés (comme l'intention de transmettre une information à un locuteur) vers leur traduction dans des opérations et des représentations qui peuvent l'être. Certaines opérations à un certain niveau sont traduites dans d'autres, à un autre niveau, et qui elles, pourront être plus formellement décrites. Ces opérations reposent sur le postulat que les intentions, et en règle générale les phénomènes pragmatiques, s'observent dans des opérations observables. Autrement dit, il serait possible de caractériser des phénomènes pragmatiques par leurs traductions dans des règles syntaxiques.

E. Valduvi va connecter la structure d'information à un traitement informatique au niveau de la syntaxe. Il part (p.147/232) du fait qu'une syntaxe de surface peut par "mapping" permettre l'identification des "packages d'information".
Chaque phrase (le postulat de E. Valduvi est que l'information est associée à la phrase) encode une proposition logico-sémantique mais également une instruction de packaging d'information. Un type d'instruction est conçu pour indiquer quelle part de la phrase constitue l'information et où et comment l'information s'inscrit dans le stock de connaissances de l'auditeur.

Suivant H.W. Choi[184], l'information est un niveau de représentation de la phrase ; reprenant E. Valduvi, ce niveau de représentation est contextuel et s'inscrit dans le déroulement linéaire du discours.
L'information comme mode de partition de la phrase est soit bi nominale soit tri nominale. Dans les deux cas, on a une focalisation et un arrière-plan. Dans le deuxième cas, l'arrière-plan se décompose en une direction et un arrière fond.

La tri- partition de E. Valduvi se structure à partir de FOCUS BACKGROUND et THEME COMMENT :



Cette définition de la structure autour des trois entités est également proposée par B. Partee, reprise par J. Peregrin.
On aurait les marqueurs suivants :
DET : opérateur
NP : restricteur
VP structure nucléaire de portée (« scope nuclear »)

Mais le principal intérêt des propositions de E. Valduvi réside dans l'intégration de la question des packages à l'intérieur de la caractérisation de la structure d'information.
Les boxes et packages sont des modèles de représentation des connaissances, donc sont caractérisés comme des outils informatiques.
Ils font partie selon I. Kruijff-Korbayová, & M. Steedman. "Discourse and information structure (op. cit.) des sémantiques d'actualisation, donc des modélisations permettant de représenter le renouvellement des données.
On considère donc qu'ils s'appliquent à la première dichotomie (thème/rhème), mais également à la trichotomie.

On a pu s'apercevoir que les structures d'informations recouvraient des figures syntaxiques relativement différentes (E. Valduvi (*The Informational Component*, op. cit., p. 12) donc que l'on ne pouvait pas lier fonctionnellement un phénomène syntaxique à l'identification de la structure d'information.
L'idée a donc été de connecter la structure d'information à une activité particulière du locuteur, et qui serait PRAGMATIQUE, et qui consisterait à caractériser l'information que contiendrait une structure de phrase. Les **instructions** sont des directions pour rechercher une information encodée dans une phrase.

Le packaging est cette activité qui consiste à traiter une phrase considérée comme préexistante. (Les boxes désignent quant à eux une modélisation des apports d'information dans un cadre référentiel).

Les questions de « package » d'information apparaissent à partir du moment où l'on va s'intéresser au contenu véhiculé, donc les questions de communication en traitant essentiellement de la façon dont un contenu nouveau est avancé relativement à certaines croyances relatives au contenu antérieur, supposé connu.
Le « package » concerne donc l'optimisation de l'entrée de l'information dans un cadre donné.
Mais en même temps, la construction de l'information dans le discours permet de structurer les différentes propositions qui s'y trouvent (E. Valduvi (*The Informational Component*, op. cit., p.12) en fonction des représentations que le locuteur a des connaissances de son interlocuteur.

A la différence des outils présentés précédemment, qui servent à associer des réalisations linguistiques à des conceptualisations fondées sur les fonctions rhétoriques (ou plus vaguement, informationnelles) du texte, la perspective des « packages » est fondamentalement orientée par des objectifs de représentation dynamique de l'apport de connaissances.
Cette perspective se situe à un niveau d'abstraction plus élaboré que les précédentes.

Pour E.F. Prince, citée par E. Valduvi (*The Informational Component*, op. cit., pp.12-13), le modèle du package d'information est un marquage formel des phrases pour indiquer les

croyances de celui qui parle à propos de l'état d'attention de son interlocuteur [structure d'information] ET le marquage formel de NP pour indiquer le statut d'entités en lien au modèle de discours [statut d'information].
Les deux niveaux précédents sont mutuellement indépendants ; l'information a donc deux niveaux : NP et phrase.

L'**information** est entendue comme un contenu propositionnel qui peut signifier différemment en fonction de la quantité d'information nouvelle véhiculée. Cette quantité varie en fonction du stock de connaissances de l'interlocuteur.
Ainsi, l'information contenue dans une phrase (Is) est la partie du contenu propositionnel (Ps)qui constitue la contribution des connaissances au stock du récepteur (Kh) :
Is = Ps – Kh

L'information dont il est question ici est évaluée à l'intérieur d'une dynamique communicationnelle et est vue du côté de celui qui reçoit l'information. (Cela justifie le recours méthodologique aux questions).

Dans cette définition de l'information, celle-ci devient indépendante de la proposition qui l'intègre pour devenir un fait d'interprétation. Dans ce cadre aussi la structure constitue un médium, entre d'une part la symbolisation des états du monde et d'autre part les modèles de discours dans lesquels les expressions informationnelles sont interprétées.

A ce moment-là, les modèles de discours sont associés à des modèles informatiques. Cela sur une base argumentée sur le fait que les dimensions de la langue et celles de la connaissance (comme d'ailleurs de l'action sociale), sont distinctes. Mais ces processus sont observables dans le discours parce qu'ils requièrent des marqueurs linguistiques comme les anaphores, pronominalisations, connecteurs argumentatifs, etc.
Ces outils langagiers permettent à la fois de comprendre comment le discours est structuré et quelles hypothèses on peut produire sur la structure de connaissances parallèle à ce discours.

Ce courant de travail donne lieu à une modélisation par « box »[185]. Cette modélisation visait au départ à traiter le lien entre la prédication, les structures du langage (notamment les composants de la phrase) et enfin l'informativité des expressions. (cf. E. Valduvi, *The Informational Component*, op. cit, p. 67).
Les informations présupposent que l'interprète dispose d'un dossier dans lequel un mot est associé à une structure et que cette structure accepte certaines variables.
Les référents de discours sont considérés comme des cartes qui sont rangées dans des dossiers. En début de discours, F0, il n'y a pas de carte dans le dossier. Les cartes sont additionnées au fur et à mesure de la progression du discours. Par exemple, « Paul m'a raconté aujourd'hui une histoire terrible ».

On représente ainsi le dossier et ses deux cartes :

ADDRESSE : Pat
Raconte z à Narrateur (Pat)

ADDRESSE : histoire (z)
terrible (z) Racontée par Pat (z)

Le récepteur du discours se dirige depuis F0 vers F1. Tant que le discours continue, toute expression va causer un changement de dossier depuis Fn vers Fn + 1.

L'histoire peut être poursuivie : « elle a vu un homme avec un sac de brocolis, et lorsque l'homme a commencé à les manger, il a été immédiatement arrêté ».

.

ADDRESSE : Pat	ADDRESSE : man(x)	ADDRESSE : sac de brocolis
Raconte z à Narrateur (Pat) Voit x (Pat)	Homme (x) Vu par Pat (x) A y (x) Mange y (x) Arrête (x)	Possédé par x (y) Mangé par x (y)

On travaille ici essentiellement sur la présupposition en considérant une connaissance préexistante et une actualisation.

Un **référent de discours** est défini comme construction mentale assurant la médiation entre les expressions référentes et les référents dans le monde réel. (E. Valduvi, *The Informational Component*, op. cit , p. 35).
Une entité de discours n'est pas une construction mentale isolée mais s'inscrit dans un modèle de discours.

Une entité de discours ne constitue pas une construction mentale isolée mais s'inscrit dans un modèle plus global. La représentation mentale n'est pas seulement celle des entités impliquées mais de leurs attributs et des liens entre eux. Cette perspective amène à H. Kamp.

Enfin, au terme de cette présentation, nous aborderons quelques aspects d'une caractérisation discursive de la structure d'information.
Le point de départ ici consiste à dire que l'information permet de dériver des interprétations autres que celles qui sont déterminées par la structure de la phrase.
Le point de départ est celui des connecteurs de discours. Ceux-ci ont des propriétés liées au discours, ce qui permet de construire à la fois une syntaxe et une sémantique du discours[186].
Par ailleurs, **les inférences** (basées sur la connaissance du monde, les conventions d'usage) contribuent à la production de propositions non fondées, construites dans l'interprétation du discours. Les inférences non fondées sont déduites des propositions fondées et sont permises par la sémantique compositionnelle. Ces inférences ne cherchent pas à lier les composants, à la différence des présuppositions anaphoriques.
Dans l'exemple « John doit être dans son bureau. Sinon, la lumière serait éteinte », l'inférence de sémantique compositionnelle considère que la deuxième clause constitue une preuve pour la première, alors que l'implicature contribue à l'interprétation non démontrable que « la lumière est allumée ». (Or comme les implicatures ne présupposent pas le lien entre les propositions, elles ne caractérisent pas l'information).
L'hypothèse consiste à dire que certains items lexicaux (à savoir les connecteurs) présupposent une éventualité ou un ensemble d'éventualités. L'hypothèse est alors que le modèle du discours contient une entité qui « décharge » (ou plus exactement indique) la présupposition anaphorique. Cette entité, qui est un connecteur de discours, possède une source. La **source** de cette éventualité, c'est une clause dans le discours ou une présupposition ayant un rôle similaire aux relations entre clauses. (Une clause désigne une expression non-nominale, telle qu'une subordonnée, une incise, etc.).

Dans une sémantique compositionnelle « classique », la deuxième clause continue la description de la situation partiellement décrite dans la première.
Dans une théorie inférentielle fondée sur les implicatures, les deux propositions contribuent à la conclusion non démontrable que la lumière du bureau est allumée.
L'interprétation est traitée ici par des ensembles alternatifs :

- Un thème présuppose un ensemble alternatif de rhèmes.
- Un focus dans un rhème réduit l'ensemble rhématique à un singleton.
- Un thème présuppose un ensemble thématique alternatif.
- Un focus dans un thème restreint l'ensemble rhématique à un singleton correspondant au thème.
- L'ensemble alternatif de rhème : propositions acceptées par le contexte (donc possibles).
- L'ensemble alternatif de thème : objets possibles dans le contexte.

On peut reprendre le schéma suivant :
Contexte $\rightarrow$ Ensemble alternatif de thèmes $\rightarrow$ Thème $\rightarrow$Ensemble alternatif de rhèmes $\rightarrow$ Rhème.

On peut également déduire de cette construction des propositions sur la structuration des discours, et notamment le passage de rhèmes vers des thèmes (voir Nobo Komagata: *Entangled Information Structure: Analysis of Complex Sentence Structures,pp.53-67)*[187].

Bilan.
Les propositions de M. Steedman, qui utilisent les grammaires catégorielles, seront traitées ultérieurement, parce que nous serons amenés à les interroger dans le cadre de la modélisation de la structure d'information. Nous nous en distinguerons parce que, à la suite des grammaires génératives, M. Steedman postule le caractère intrinsèquement génératif des structures d'information. Au contraire, pour nous, les dynamiques sont externes, et on se positionne sur l'interprétation. Les structures d'information ne constituent que des représentations d'états, d'actions et d'événements dans cette dynamique.
En définitive, dans leur version la plus élaborée, les structures d'information sont des phénomènes associés à la réception. Nous avons dû passer par cette longue présentation afin de présenter le fait que les structures d'information se situent à l'interface entre les questions linguistiques de discours, de représentation des connaissances et des raisonnements et de dynamique cognitive dans un cadre de pragmatique conversationnelle.
Nous garderons donc ce principe d'une caractérisation de la structure comme un phénomène d'interprétation, ce qui permet d'articuler les dimensions de flux à celles de structures. Néanmoins, nous nous distinguons de ces perspectives par l'importance que nous donnons à la dimension référentielle.
Les fondements de la structure d'information sont donc soit prédicatifs, soit rhétoriques. Concernant ces derniers, le problème est qu'il est nécessaire de disposer d'une théorie de la communication afin d'aboutir à une caractérisation formelle des échanges. En dépit des travaux effectués par J. Van Benthen, il n'existe pas véritablement un cadre permettant de modéliser ce phénomène sur la durée d'une activité. Par ailleurs, comme l'origine rhétorique le rappelle, les logiques fondées sur le dialogue traitent d'un problème de communication, et

non, comme dans notre cas, d'un transfert d'information associé à la capacité des outils à produire et transmettre des contenus symboliques.

Si la structure d'information ne constitue pas un concept qui puisse être intégré aisément dans une théorie du langage mais seulement dans une théorie du discours, on peut se poser légitimement la question de sa formation, de sa représentation et de son interprétation. Comme on ne peut la caractériser comme un phénomène seulement linguistique, et d'ailleurs les modélisations par l'intonation en sont un bon indice, on pourra proposer une explication en utilisant d'autres outils et méthodes. C'est justement ce à quoi nous nous engageons maintenant.

Un autre point problématique pour nous est bien que les théories du discours, en tant qu'elles s'intéressent à la communication d'un contenu (considéré comme nouveau dans le cadre de l'échange), considèrent ce contenu comme étant déjà une connaissance pour un des membres de cet échange. En ce sens, une structure d'information ne transmet un phénomène nouveau que pour un membre de l'interaction. On ne résout donc pas ici le problème de la construction de l'information en tant que production d'un contenu inédit.

A ces critiques sur les fondements de la modélisation de la structure d'information, on peut adjoindre celles qui concernent son efficacité en termes d'extraction et d'annotation. Dans les travaux de P. Cook & F. Bildhauer [188], la structure d'information est utilisée pour l'annotation de corpus (en prévision d'une extraction des structures d'information). C'est une annotation manuelle. Les problèmes qui y sont posés, même s'ils montrent par ailleurs l'intérêt du modèle pour la caractérisation de problèmes d'analyse linguistique, caractérisent les limites de la structure pour une exploitation automatique. En effet, la structure d'information, dans ce type d'acception, ne repose sur aucune structure lexicale. Ce qui ne permet pas l'automatisation. En proposant de corréler la structure d'information et des structures sémantiques lexicales, nous pouvons lever ce verrou. Au fondement de la structure d'information, on trouve un processus informationnel et non un phénomène rhétorique. C'est cette corrélation qui permet de lier les types lexicaux et les structures phrastiques à condition évidemment de déterminer les types lexicaux conformes aux contraintes de cette structure.

Enfin, les perspectives de l'analyse discursive permettent de mettre en évidence des phénomènes, comme la nouveauté de l'information dans le discours[213], qu'il ne nous est pas possible de prendre en compte autrement que dans le cadre de la temporalité associée à chaque structure d'information. C'est la conséquence d'une sémantique essentiellement référentielle.
Néanmoins, on peut retenir une caractérisation précise du phénomène que représente la structure d'information dans les discours. Il s'agit de la caractérisation d'un fait nouveau à l'intérieur d'un contexte connu et reconnu par le récepteur. Ce fait transforme la connaissance que l'on a de ce contexte.

5.2.2. Caractérisation des constituants de la structure d'information.

Nous avons vu que les approches de la structure d'information se fondaient sur des principes discursifs mais intégrant de plus en plus des modèles de connaissances. Dans cette

[213] La nouveauté de l'information peut être essentielle dans le cadre de certaines problématiques d'extraction d'information.

présentation, nous nous sommes limité à des travaux à fondement linguistiques parce qu'ils nous ont apparu comme étant les plus fondés sur des entités observables. De nombreux travaux existent dans un cadre informatique, mais ils considèrent la structure d'information de façon totalement ad hoc, liée à des stratégies et des outils d'extraction propres. Ainsi, conformément à notre principe de corrélation entre les Sciences Humaines et Sociales et les Sciences de l'Ingénieur, ces propositions n'ont pas été retenues. Conformément aussi aux principes de relations entre structures lexicales et information, on proposera une entrée dans les questions de structure d'information en reprenant les théories de l'information et des flux que nous avons développées précédemment et que l'on intègre dans ce domaine.

Nous appelons les constituants de la structure d'information des primitives. A la manière des relations ontologiques, ces primitives constituent des unités indécomposables de la structure. Après avoir précisé comment il est possible d'entendre ces primitives, nous en donnerons une justification syntaxique. Celle-ci sera suivie d'une sémantique, et comme cela on pourra envisager les primitives comme les unités conceptuelles qui assemblent des comportements d'unités symboliques relevant de dimensions différentes. Ainsi, les primitives assemblent les comportements des unités symboliques à la fois dans le cadre des opérations, des flux, de la syntaxe des expressions symboliques et de leur interprétation.
Nous commencerons par définir ces primitives, puis nous caractériserons la dimension syntaxique de la structure d'information. Auparavant, il nous faudra donner une explication sommaire sur ce que sont les grammaires catégorielles.

Définition des primitives.
Il convient de distinguer nos primitives d'un certain nombre d'autres catégorisations des constituants sémantiques de l'expression, et notamment des cas de C. Fillmore[189], ou des primitives de RJ Jackendoff[190].
Les cas de C. Lilllmore constituent un ensemble de relations sémantique entre les différents constituants d'une expression. Dan Moldovan & alii[191] donnent une vision globale de ces cas et de leur utilité dans le cadre de rapports question-réponse.
C. Fillmore définit une catégorisation générale des relations sémantiques, sans au préalable caractériser de fonctionnalité.

Notre point de vue est différent puisque l'on ne postule pas une légitimité linguistique mais fonctionnelle à nos primitives. (Cette position n'empêche pas de reprendre à notre compte des travaux de linguistique pertinents dans notre problématique).

Prenons un exemple simple :
"M. Smith, Kinetic, 4.8 mg/ml, 12:24",

Un INDIVIDU est alors n'importe quel indexical (déictique) qui isole un objet composé dans le monde. ("*M. Smith*")**.** Cette primitive répond à la définition de l'individu proposée par STRAWSON.
Une PROPRIETE est n'importe quel concept qui constitue un type en relation à un individu et ayant un rôle fonctionnel vis-à-vis du spécificateur. ("*mg/ml*"). On reprend ici la définition des propriétés proposée par la théorie des situations.
Un EVENEMENT est le concept type qui classe à la fois la classification par propriété et la classification par spécification fonctionnelle. Dans la théorie de l'information, l'événement est associé au canal. ("*Kinetic*"). Nous nous référons également à la définition de l'événement proposé par la théorie des situations.

PARTIE DE [INDIVIDU] accepte n'importe quel indexical qui spécifie la partie de l'individu satisfaisant la spécification de la propriété. ("12:56").
SPECIFICATION accepte n'importe quelle valeur fonctionne associée à la propriété et relative à une partie distincte de l'individu. ("*5,56*"). Cette spécification est valable dans le contexte défini par la propriété, mais caractérise en fait un phénomène plus général (une unité sur une échelle) que celui proposé par la propriété.

Nos exemples montrent ainsi des unités lexicales fortement distinctes entre elles mais qui constituent une expression ; ces unités lexicales sont substituables au sein d'une même position dans l'expression. On peut ainsi construire des classes à partir de ces noms propres, marqueurs temporels, de spécification ou d'événement. Ces classes sont décrites parles primitives ; celles-ci servent à caractériser les classes de lexiques acceptables pour des positions dans la structure d'information.
Ces primitives sont définies par leur hétérogénéité. Un intérêt primordial de la caractérisation de primitives pour la structure d'information réside dans la capacité d'intégration de lexiques relativement étendus et aux contours souples. En particulier, on pourrait les représenter sous forme de clusters.

Structure d'information et syntaxe.
Nous avons présenté les constituants de la structure en insistant sur leur articulation aux flux et aux situations. On ne peut pas se contenter d'une telle présentation pour affirmer le fait que l'on a affaire à une structure d'unités symboliques. Pour cela, on doit associer à cette construction sémantique un niveau syntaxique.

Nous schématisons les différents niveaux d'analyse qui contribuent à l'élaboration des primitives (et de leur relations contraintes) comme suit :

	Processus d'information	**Structure linguistique**
Niveau abstrait	Flux d'information	Sémantique
Niveau des observables	Opérations	Syntaxe

La syntaxe ne permet pas de caractériser d'autres dimensions de la structure, notamment la dimension spatiale de sa répartition sur la feuille.
Le choix d'une représentation syntaxique abstraite répond au besoin de disposer d'une grammaire utilisable dans des contextes très différents, mais surtout qui puisse s'articuler aux propositions des flux. Elles sont à fondement lexical et reposent sur le principe d'opération depuis un type d'entité vers un autre.

Les primitives que l'on présente sont de nature sémantique. Si nous développons d'abord leur dimension syntaxique, nous les définissons néanmoins d'abord comme types sémantiques.

Principes théoriques et méthodologiques à la caractérisation sémantique des primitives.
Nous nous référons à un cadre conceptuel qui est celui de la théorie des types, mais tel qu'il est considéré dans le cadre de l'analyse linguistique. Ca cadre permet de justifier les primitives comme objets scientifiques. Cette perspective n'est pas à confondre avec une modélisation de l'interprétation utilisant les types tels que définis dans la théorie des situations.
Une théorie de types (dont les grammaires de trait font partie) est une construction par laquelle la classification d'une occurrence dans une catégorie est enrichie par le type qui la décrit. En somme, il s'agit de l'activité contraire d'une classification par un générique : on ne

vise pas à éliminer des traits (qui seraient occurrents dans un objet, afin de l'associer à d'autres objets), mais au contraire à caractériser une propriété (prédéfinie) que l'entité posséderait et qui serait attestée sous forme d'une valeur prise.
Autrement dit, "chèvres" serait de type nominal parce que le mot contient le trait NOMBRE ayant une valeur [pluriel] (parmi d'autres possibles), le trait GENRE avec la valeur [féminin] et la fonction DETERMINANT qui est caractérisée par la valeur [pluriel]. Cette fonction DETERMINANT part de « chèvres » vers l'ensemble des déterminants [féminin, pluriel]. Néanmoins, ces traits caractérisent seulement la catégorie : la dimension type est donnée par le fait que [chèvre] désigne un certain type d'objet dans un certain univers et que chacun des traits permet de spécifier la quantité ou la précision des objets satisfaits. A ce moment, on a effectivement décrit un type). Autrement dit, caractériser une entité (un mot, une phrase) ne consiste pas à la classer dans un objet plus général mais au contraire à caractériser ses différentes propriétés (qui sont des traits ne décrivant que partiellement l'objet).

En linguistique, et plus particulièrement en sémantique formelle, les traits ou les types sont des outils qui sont construits pour représenter un certain nombre de faits (qui peuvent être partiels par rapport au langage : la passivation, la question, la négation). Ils ne postulent pas de théorie relativement à ce qu'est la langue, son rôle, etc. Ils visent seulement à expliciter des faits sélectionnés par celui qui analyse.
Par conséquent, ces outils ont des usages qui peuvent être extrêmement variés, et notamment, par exemple, la distinction entre les différentes significations d'un mot en fonction de son contexte syntaxique et discursif. "Les types distributionnels sont des propriétés des items lexicaux en tant que tels" (D. Goddard & J. Jayez, « Quels sont les faits ? », p.2)[192].
Cela dit, pour D. Goddard & J. Jayez, certains types ne sont pas lexicaux mais bien phrastiques[214]. Il s'ensuit que les types distributionnels servent à construire des structures de traits ayant des propriétés sémantiques variées. Ils citent les rôles thématiques, qui permettent de construire les relations entre prédicats. De cette façon, le typage sert à caractériser des propriétés de prédicats mais également à caractériser des relations que peuvent établir différents prédicats.

Nous reviendrons sur ces traits lorsque nous caractériserons la structure d'information dans le cadre de l'extraction d'information. Néanmoins, on peut d'ores et déjà distinguer la structure telle qu'on l'envisage et la structure d'information telle qu'elle est envisagée en linguistique. Si nous conservons la même définition de ce que représente la structure d'information, par contre les outils de représentation et de spécification de cette structure sont différents. Nous ne nous situons pas dans une perspective fondamentalement discursive et ne considérons pas la structure comme un objet linéaire.
En effet, dans notre approche, les unités composant la structure d'information peuvent être disséminées dans un discours. L'exemple de l'information pharmaceutique l'illustre ; le média impose une dissémination spatiale de la structure, ce qui constitue aussi un principe de représentation de l'information. En effet, la lecture et l'interprétation de l'information peuvent se réaliser suivant différentes entrées

Quelques notions de grammaires catégorielles.
Les grammaires catégorielles constituent un ensemble de travaux reprenant en partie les principes posés par la théorie des types. Pour M. Mootgat[193], les deux questions centrales des grammaires catégorielles sont :

- quels sont les invariants de la composition grammaticale ?

[214] En effet, considérant les principes de Vendler, l'interprétation n'est pas compositionnelle mais fondée sur des structures accolées à des unités lexicales, et en premier lieu les connecteurs.

- Comment les uniformités de correspondance entre forme et signification peuvent être capturées entre les langages ?

On dispose d'invariants grammaticaux et d'une variation structurale.
La catégorisation fonctionne comme déduction. On assigne des types à des entités élémentaires dans un lexique. Ces opérations de formations de types sont des connecteurs logiques.
Les types basiques jouent un rôle similaire aux constituants majeurs de la phrase: ils catégorisent les expressions que l'on considère comme complètes (*np* pour nom propre, *s* pour phrase, *n* pour phrase nominale notamment).
Les opérations unaires et binaires constituent un vocabulaire permettant de catégoriser les expressions dans les termes de leurs parties constituantes. (A B) catégorise une expression qui peut être décomposée dans un constituant de type A et suivi par un constituant de type B.
Une expression avec une fraction de type A/B (ou B\A) combine une expression de type A avec une expression de type B. Les opérations unaires sont un ajout récent qui peut être considéré comme une attribution de traits.

Les grammaires catégorielles constituent des outils de calcul permettant de représenter les enchaînements et les choix d'unités à partir de certaines autres entités. Ce sont des grammaires fondées sur des raisonnements. Elles s'adaptent à des objectifs distincts, qui peuvent être soit de caractériser des capacités génératives, soit au contraire de caractériser des ordonnancements.
Une entité se caractérise par les opérations qu'elle réalise, à savoir les entités qu'elle actualise nécessairement. Ce qui implique de définir une catégorie pour un mot relativement aux entités qu'il implique.
Ainsi : « chien » : n
« le » : gn/n

Les catégories caractérisées sont fondées sur des lexiques, ce qui signifie que les propriétés grammaticales sont associées à des mots.

Caractérisation syntaxique de la structure d'information.
Nous présentons maintenant le modèle global de la structure d'information, caractérisé à la fois par une écriture utilisant les flux et une autre, utilisant les grammaires catégorielles.

Nous nous sommes pour cela largement inspiré des travaux de M. Steedman, non seulement parce qu'ils portent sur la structure d'information, mais aussi parce que les outils qu'il mobilise ont la capacité de prendre en charge les relations marquées par les flux. Cela est dû entre autre à l'absence d'une théorie puissante concernant le langage[215]. Par ailleurs, les grammaires catégorielles de M. Steedman sont génératives, ce qui explique leur cohérence par rapport aux flux, si l'on considère que l'interprétation d'une structure repose en partie sur la reconnaissance des opérations de production.

On s'intéresse à l'expression considérée au moment de sa duplication et classification dans un texte. Cela nous permet d'éviter de confondre les questions de discours et celles d'information. Les opérations syntaxiques sont, comme nous l'avons considéré, les traces des opérations dans le monde, mais également le miroir des opérations d'écriture et de

[215] Ce sera la raison pour laquelle on n'utilisera pas les HPSG, qui constituent une grammaire intégrant la théorie des situations pour sa sémantique, mais qui ne permet pas d'intégrer d'autres relations et contraintes que celles fixées par la langue.

renseignement de la base. (Pour mémoire, voici un énoncé exemplaire : " M. Smith, kinetic, concentrations, 5,56 µg/ml, T0, 12h32 ")
On isole les règles simples suivant le principe selon lequel les opérations et les flux utilisent des catégories linguistiques très larges, que l'on peut comparer à des clusters. Nous illustrons sur l'exemple la conversion en utilisant à la fois le vocabulaire des opérations et celui des flux. On propose la formulation suivante :

Entité dénommée (ou token d'un ensemble de tokens) : "M. Smith": NP
Paramètre : type de la classification de l'entité dénommée : "mg/ml": (NP)\N
Canal permettant la duplication du résultat de l'opération et élément fondateur de la structure : "Kinetic": NP\S/(NP)\N
Segmentation de l'entité dénommée, ici en l'occurrence un marqueur temporel, considéré comme token : "12:56": NP\NP
Résultat typique de l'application du paramètre sur l'individu dans la segmentation présentée : NP\NP : "5,56": (NP\NP)\N\N

On peut maintenant préciser quelque peu la représentation syntaxique. Nous présentons cette caractérisation en reprenant les différentes étapes de l'intégration du résultat de l'opération à l'intérieur du texte.

La condition à une classification est l'existence d'une feuille dévolue et donc le nom du patient en position d'argument :

M. Smith	mg/ml	Kinetic
NP	(NP)\N	(NP\S)/N

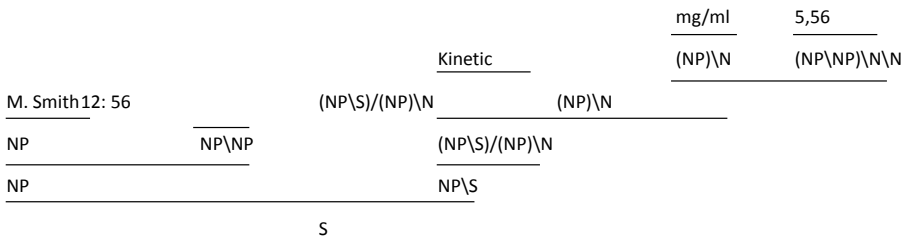
Par une opération de substitution, on obtient cette première forme, qui correspond au contexte linguistique de l'expression du résultat de l'opération :

	Kinetic	mg/ml
	(NP\S)/(NP)\N	(NP)\N
M. Smith	(NP\S)/(NP)\N	
NP	NP\S	
S		

Cette structure est pérenne au sens où elle accepte l'ensemble les résultats possibles dans le cadre qu'elle fixe.
Une structure d'information spécifique requiert une spécification en temps du résultat. Cette insertion qui individualise la structure relie le temps au NP et la spécification au N :

12:56	5,56
NP\NP	(NP\NP)\N\N

Les deux derniers composants sont intégrés par adjonction :



La grammaire applicative caractérise comment le niveau syntaxique de description hérite de la structure des opérations fondatrices des flux. Ainsi, chaque opération de choix est considérée comme une inférence grammaticale, depuis des entités sélectionnées de façon arbitraire vers les résultats des opérations.

Cette catégorisation permet les différentes dérivations possibles suivantes sans changement de signification :
"M. Smith, Kinetic, 4.8 mg/ml, 12:24",
"M. Smith, 12:24, Kinetic, 4.8 mg/ml".

Les autres formulations ("Kinetic, M. Smith, 12:24, 4.8 mg/ml" ou "Kinetic, M. Smith, 4.8 mg/ml, 12:24") ont des conséquences sur la classification de l'expression dans le texte et donc ont des conséquences sur l'interprétation.

Les structures **(NP)\N** et **(NP\NP)\N\N** représentent la duplication elle-même ; cela inclut le fait que NP et NP\NP apparaissent à la fois dans le support imprimé et dans l'information transmise.

Ainsi, le niveau linguistique permet la caractérisation du système d'information comme une construction structurale et non plus seulement comme une succession linéaire d'opérations.

Cette construction fonde la prédication et il est donc possible de passer de la grammaire à la sémantique, considérant alors la structure d'information dans sa dimension interprétative.
Considérant les contraintes présentées précédemment, la représentation prédicative de la structure hérite de l'ordre des opérations et de l'élaboration progressive de l'information.

Formulation prédicative de la structure.
On justifie une structure de prédicats par une traduction des opérations depuis les flux. Ainsi, on fait correspondre :

- Chaque relation : une prédication simple ;
  exemple : <M. $Smith_1$ (x, 2), $Kinetic_2$ (x, y, 3, 4)>
- Chaque fonction : une structure prédicative marquée ;
  exemple : (<M. $Smith_1$ (x, 2), $Kinetic_2$ (x, y, 3, 4), mg/ml_4 (x, 3)>
- Le canal : une prédication complète ;
  exemple : (x,y) [(12:24 (y, 1)), (<M. $Smith_1$ (x, 2), $Kinetic_2$ (x, y, 3, 4), mg/ml_4 (x, 3)>, 4.83_3 (y))]

La structure d'information détermine ce qui peut être dit à propos du monde distinctivement de ce que l'on sait. On représente maintenant la structure en intégrant l'apport de connaissances nouvelles, en considérant que la structure d'information représente ce qui fait

la différence. En ce sens nous caractérisons en quoi nous avons bien une structure d'information.
Nous considérons les individus comme des objets dénommés ; ils sont donc caractérisés comme des variables (x, y).
Les constructions parenthétiques (<,>) caractérisent les parties thématiques (ou *x*) de l'information ;
Les unités nouvelles adjointes de la structure représentent la partie rhématique (or *y*) de la structure.
Ainsi, on obtient la représentation suivante :
(x,y), [(12:24 (y, 1)), (<M. Smith$_1$ (x, 2), Kinetic$_2$ (x, y, 3, 4), mg/ml$_4$ (x, 3)>, 4.83$_3$ (y))]

Le processus général de flux d'information peut être considéré comme une explication des régularités de la structure d'information. Ainsi, les opérations constituent des contraintes de flux qui ensuite se traduisent sur des règles grammaticales. Le flux permet l'interprétation de chaque composant comme étant un cadre de connaissances, et donc comme un composant de la prédication. Ces types caractérisent un cadre particulier de connaissances pour chaque catégorie et chaque composant de la prédication. Nous proposons la traduction suivante, que l'on argumentera dans la partie suivante par la sémantique propre de chacune des primitives. Les équivalences entre les catégories et les types pour la formation des primitives seront formulées ainsi :

- A propos de la dénomination de l'entité individuelle : "M. Smith": **NP: INDIVIDU**
- A propos de la segmentation en temps de l'individu (paramètre 1) : "12:56" **NP\NP: PARTIE DE IND.**
- A propos du paramètre 2 (type de la première classification des flux : "mg/ml": **(NP)\N: PROPRIETE.**
- A propos du paramètre 3 (type de la seconde classification et résultat de l'opération : "5,56": **(NP\NP)\N\N: SPECIFICATION**.
- A propos du canal ou paramètre 4 : "Kinetic": **NP\S/(NP)\N: EVENEMENT.**

Nous pouvons maintenant présenter une représentation de la prédication intégrant les primitives. Le fondement prédicatif étant pour nous défini par les flux, nous pouvons maintenant caractériser comment les flux peuvent se traduire dans l'écriture des structures informationnelle :
" M. Smith" : [a.A] : NP
"T0, 12h32 " : [$f^{\vee}$: b.B $\rightarrow$ a.A] : NP\NP
"µg/ml" : [$\vDash$ <a, α>] : (NP)\N
"5,56" : [$\vDash$ <b, β >, ($f^{\wedge}$: $\alpha \rightarrow \beta$)] : (NP\NP)\N\N
"kinetic" : [$\vDash$ <<a, b>, χ >, < <$\alpha \oplus \beta$>, χ >>] : (NP\S)/(NP)\N

On pourra conclure notre caractérisation de la prédication en spécifiant ce qu'elle représente. Dans le cours de l'activité, l'interprétation est une activité générique pour n'importe quel agent. Elle consiste à inférer un fait à partir d'une structure d'information, et donc une évolution de l'état du monde décrit par le discours. Dans ce cadre, un fait est un état dans le cours de chaque événement. L'événement est marqué par le canal et indexé sur n'importe quel individu.
A partir de la succession des états, on peut reconstituer la trajectoire de l'événement et se risquer à quelques prévisions ou anticipations. On peut ainsi proposer quelques inférences à

propos de l'évolution du monde et plus particulièrement ici, du patient. Ce processus est connecté au raisonnement du pharmacien, et à ses outils, notamment le logiciel USC*PACK, fondé sur des probabilités conditionnelles.

Conclusion.
La présentation que nous venons de développer ne concerne que l'élaboration de la structure d'information. Nous avons parcouru jusqu'à présent la dérivation suivante, depuis des opérations dans le monde vers des flux d'information et des structures linguistiques :

Opérations matérielles dans l'activité	Régularités des flux d'information	Formation des structures d'information

La partie suivante concerne elle l'interprétation, et impliquera notamment la théorie des situations. Nous formulerons les primitives de façon plus précise encore.

D'ors et déjà, on peut considérer que l'on aboutit à la caractérisation d'un certain type d'information. Il s'agit bien de caractériser ici comment un certain état du monde, original, peut être transmis dans le cadre d'un système symbolique.
On a défini ici un certain format pour l'information. Il répond à la question de ce que l'on transmet du monde. Il distingue par ailleurs clairement, et c'est ce à quoi a servi l'état de l'art sur la structure d'information, ce qui est une information de ce qui constitue un partage de connaissances.

On obtient ainsi une représentation qui peut être appliquée à des données structurées de niveaux d'abstraction différents. Si l'enjeu initial consiste pour nous à caractériser une structure d'information qui puisse servir de modèle dans le cadre de l'extraction d'information, par contre, on pourra utiliser cette structure pour caractériser les expressions que l'on obtient en liant les différentes données structurées hétérogènes dans le cadre des flux.

5.2.3. Caractérisation de l'interprétation des constituants de la structure d'information.

Nous nous sommes intéressé précédemment à une définition par les flux des constituants de la structure d'information et à une caractérisation de l'interprétation qui ne prend pas en compte la dimension proprement linguistique des expressions informationnelles. En effet, ce n'est pas le propre des flux que d'expliquer l'interprétation des unités et des structures symboliques. Maintenant nous pouvons présenter l'interprétation de ces structures.
On définit les composants de la structure comme des types primitifs dans lesquels les unités lexicales candidates doivent être classées pour intégrer la structure à une certaine place.

En choisissant de caractériser ces composants comme des types, on se situe dans une continuité théorique par rapport à la modélisation par les flux. Pour cela, nous présentons une approche relativement générale en sémantique de l'utilisation du typage dans la modélisation des phénomènes linguistiques.
L'interprétation que l'on cherche maintenant à représenter hérite de l'ensemble des phénomènes caractérisés dans la démarche de production et de transfert de l'information. Par conséquent, l'interprétation des constituants de l'information reprend l'ensemble des opérations présentées plus haut. Elle utilise la théorie des situations comme vocabulaire d'expression et vise à clarifier les questions de logiques locales laissées en suspens précédemment.

Cette sémantique associée à la structure d'information nous permet de commencer à répondre à la question de la structuration de l'univers de référence. Nous avions émis l'hypothèse que les mondes de référence des structures d'information (et donc les flux) ne pouvaient se résumer ni dans les principes d'un seul et unique univers de référence, ni dans celui de mondes possibles. On pourra à l'aide de l'interprétation de la structure d'information, argumenter plus finement cette hypothèse. La partie qui suivra sera consacrée à la modélisation cognitive de cet état en utilisant des ressources théoriques supplémentaires, issues de travaux sur la cognition.

Interprétation de la structure d'information et distinction avec les flux.
Dans le cours d'une activité, l'interprétation est un raisonnement produit par n'importe quel agent. Elle consiste à inférer un fait à partir d'une structure d'information, et donc un état temporaire dans le cours des événements relatés par le discours.
A propos de chaque individu, un fait permet d'inférer un état ; l'inférence d'une évolution est fondée sur la succession des états (fondés sur des faits) dans le discours. Ces faits sont actualisés par le canal. C'est ainsi aussi que l'on peut prévoir les états futurs dans le cours d'un certain événement : un évènement caractérisant une transformation dans une certaine continuité temporelle, l'ensemble des états réalisés permet de prévoir (à l'aide d'un modèle) les états suivants. Ce processus interprétatif est directement en relation aux outils de calcul probabilistes utilisant les probabilités conditionnelles mis en œuvre par les pharmaciens.

Dans la théorie des situations, un fait est défini par une structure d'information identifiant une conformité entre une situation et une proposition. C'est la raison pour laquelle il n'y a pas d'ambigüités dans la structure d'information.

Afin de caractériser cela, nous posons le vocabulaire suivant, issu des propositions de K. Devlin (op. cit.) :
S : situation
T : type de situation
σ : expression informationnelle
Les situations sont des instances de situations types. Les situations-types sont des propositions qui correspondent à des cadres : ce sont des segmentations du monde.

Le fait introduit une dualité acceptée par les flux : chaque proposition peut être interprétée à la fois au niveau des types et des instances. Les classifications sont donc interprétées à deux niveaux de façon systématique. Ainsi, à la fois les types et les tokens ont un contenu. L'interprétation d'une proposition peut avoir la double formulation suivante :

- « cette proposition représente comment ce type a ce token comme instance » : $s \vDash T : \sigma$
- « cette proposition représente comment ce token satisfait ce type » : $s \vDash \sigma : T$.

Cette interprétation bi-dimensionnelle est conditionnée par la reconnaissance de la classification comme un ensemble de contraintes.

Les contraintes des flux sont des paramètres permettant l'interprétation d'une situation dans le monde. Ainsi :
"*M. Smith, mg/ml"* :: T : [par. (OP) | s $\vDash$ << R(α), a>>]

Traduction : « une situation *s* est acceptée par cette proposition (ou cette proposition signifie dans cette situation) si les paramètres définis par les flux sont satisfaits. Ces paramètres sont

définis par un type de situation ». Le type de situation définissant les paramètres correspond à une opération. Ce type de situation paramétrique est similaire à la situation d'énonciation, telle que posée par la théorie des situations.

On considère que *a* et *b* représentent des tokens qui constituent les signifiants d'objets du monde distincts symbolisés par *x* et *y*. Ce sont des objets de référence hors de la symbolisation mais auxquels cette dernière réfère. Chaque dépendance de la représentation prédicative est interprétée comme suit au niveau des tokens :
T1 : [par. (OP_1) | s (y, x) $\vDash$ << a, b>>]
T2 : [par. (OP_2) | s (x) $\vDash$ << R(α), a>>]
T3 : [par. (OP_3) | s (y$\rhd$x) $\vDash$ << [R(α), a] $\rightarrow$ [R(β), b]>>]
T4 : [par. (OP_4) | s (y$\rhd$x) $\vDash$ << R(χ), [R(α), a] $\rightleftarrows$ [R(β), b]>>]

Ici, chaque type de situation correspond à une situation particulière.
Pour chaque proposition, un paramètre (depuis le niveau de l'opération) est requis. (Les opérations sont des situations dans le cours de la production de l'information. Un paramètre dénote une situation dans l'univers des opérations paramétriques des flux). Ainsi, la signification est contrainte par les flux parce que le flux produit l'information.

Maintenant, afin de progresser dans la caractérisation de l'interprétation des structures d'information, il nous faut passer par la modélisation des trois mondes que nous avons précédemment évoqués : le monde des individus, le monde des opérations et celui de l'interprétation. Nous nous contentons maintenant de les définir de façon à ce qu'ils soient utilisables dans le cadre d'une sémantique. Nous verrons leur justification théorique, en termes de théorie de la cognition, dans la partie suivante.

Dimensions des trois mondes.
Maintenant, nous revenons à une caractérisation sémantique des univers que nous venons de proposer de façon à ce que l'interprétation puisse être modélisée en conformité avec le cadre dans lequel les flux ont été définis.

Nous proposons maintenant une première modélisation de ces trois mondes, considérant que l'on se situe toujours à l'intérieur du cadre théorique posé par J. Barwise. La sémantique considère qu'il existe un monde unique (et non des mondes possibles) et des facultés cognitives, d'abstraction notamment. Mais cette dualité est insuffisante pour caractériser une interprétation intégrant les contenus et le discours. (Rappelons que par contenu on considère l'ensemble des termes pouvant se réaliser dans le cadre d'une prédication). C'est pour cela que nous proposons une troisième composante : le monde des opérations mentales externalisées.
Le monde est ainsi segmenté par les flux en trois mondes : un monde en extension (ou monde des tokens), un monde des représentations externalisées de ce monde en extension, qui lui-même peut être également en extension, et la corrélation de ces deux mondes dans le monde en intension (ou le monde des types).

Le monde extensionnel est caractérisé par l'ensemble des individus qui vont pouvoir être soumis à la procédure caractérisée par les flux. (Par exemple, le patient, la bactérie, la molécule).
Le monde intensionnel caractérise les propriétés associées à des individus.
L'univers des opérations caractérise l'association, d'une propriété, d'objets du monde réalisés en série et d'unités symboliques structurées (comme les échelles de valeurs).

Le rôle des flux dans la structuration de ces trois univers est essentiel puisque c'est le flux qui permet de les mettre en relation, et notamment permet l'existence même du monde des opérations externalisées. La distribution des opérations par rapport au monde intensionnel constitue un point essentiel de l'apport des flux.
Une seconde qualité est la capacité à inférer des propositions depuis les propriétés associées au monde intensionnel, vers celui des opérations et in fine le monde en extension. Cette inférence correspond à la capacité des flux à mettre en relation de façon signifiante des mondes hétérogènes par des trajectoires informationnelles depuis le monde en extension vers le monde intensionnel (au travers notamment de propositions de doses médicamenteuses).

On peut schématiser le flux d'information entre les différents mondes de la façon suivante, considérant que U signifie le monde en extension, U' le monde en intension, et U'' le monde des opérations, qui hérite de caractéristiques des deux précédents :

U' -> U''
| |
U <- Ut

On note que l'univers extensionnel est structuré par une paire <individu, temps>, ce qui rappelle la structuration de l'extension proposée par R. MONTAGUE. Nous reviendrons sur la façon dont la temporalité est caractérisée dans l'interprétation.

La structure d'information est interprétée différemment dans chacun de ces mondes, ce qui oblige à proposer une interprétation différente pour chacun. En d'autres termes, chaque primitive signifie différemment dans chacun de ces mondes.
De façon informelle, on peut présenter les liens entre opérations matérielles et opérations réalisées dans un univers de la façon suivante :

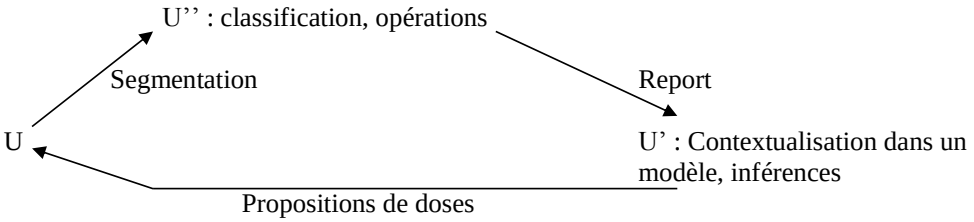


Schéma des relations entre les différents mondes de l'activité.

Si l'on regarde par rapport au contexte (qui n'est plus simplement celui de données structurées mais d'une activité humaine), on peut facilement comprendre que les flux rendent compte de la production d'outils et d'artifices de représentation. Cette portée des flux représente un aspect essentiel du rôle de l'information dans les pratiques scientifiques.

Nous pouvons maintenant poser le modèle général des mondes et la façon dont notre vocabulaire s'insère dans ce cadre.

Monde U: objets, parties d'objet	Monde U': concepts, relations entre concepts	Monde U'': élaboration d'information et transfert
Interprétation extensionelle	Interprétation intensionelle	Interprétation extensionelle
Référence	Signification attributive	Connaissance commune de l'utilisation du langage dans l'activité.

Propriétés de chacun des mondes.
Nous présentons maintenant quelques propriétés intrinsèques de ces mondes avant d'aborder les phénomènes d'interprétation[216]. Ceci afin de caractériser les différents objets dans leur monde propre de référence.

U'' : monde défini par des contraintes (et notamment par les opérations). Limitation du nombre d'informations possibles et donc des unités lexicales. Ce monde est une externalisation des compétences cognitives humaines.
Explication de l'infinitude des phénomènes du monde. $(s_1) \oplus (s_2),(s_3) \oplus (s_4)$

U' : monde caractérisé en intension par des types associés à des propriétés et permettant de proposer des inférences. Ce monde est constitué de représentations mentales et constitue l'univers d'interprétation des acteurs. [T : <<σ>>]

U : monde composé d'individus ; certains individus peuvent être explicites et dénommés, d'autres peuvent rester implicites. Ce monde est constitué d'objets dénommés par un nom propre. (Les distinctions que l'on peut établir entre les noms de personnes humaines et de bactéries ou de molécules ne sont pas opérantes dans notre cadre). Ainsi, on obtient, pour un individu x, son expansion y (généralement marquée temporellement) : x,y, y⊳x

Modélisation de l'interprétation intégrant les trois mondes.
La caractérisation du cadre de l'activité a impliqué la définition de trois mondes ayant chacun leurs spécificités référentielles et intensionnelles. On peut maintenant reformuler les propositions de la théorie des situations intégrant les propriétés de ces trois mondes.

Tout d'abord, caractérisons précisément la façon dont les trois mondes peuvent être caractérisés dans une interprétation intégrant les flux. Nous présentons donc une interprétation intégrant la duplication (défini par F. Dretske). On a vu le rôle essentiel des contraintes dans la théorie des situations et les flux. Nous appliquons maintenant ces principes à l'articulation entre les trois univers, de façon à rendre compte d'inférences depuis l'interprétation effectuée dans un univers vers un autre.
Les flux permettront de comprendre comment une interprétation issue d'un univers est transmise vers une interprétation dans un autre.

Si la classification (U': $s_3 \vDash <\chi, <<\alpha \oplus \beta>, <a,b>>>$) interprète le résultat d'opération
(U'': $s_1. \lhd s_2 \vDash < <s_1 \vDash <\alpha, a>>, s_2 \vDash <\beta, b >>$) et que cette information interprète

[216] Nous reviendrons nettement plus en détail sur cette caractérisation des mondes, lorsque nous aurons réintégré la question de l'activité (et donc la modélisation cognitive) dans le cadre de travail.

(U: $s_{<ind \lhd part\ of\ ind>} \vDash < a(x) \lhd b(y) >$),
alors la classification [s_3] permet d'interpréter [$s_{<ind \lhd part\ of\ ind>}$].

Ainsi, pour n'importe quelle proposition dans le texte, on interprète à la fois une situation dans le monde, et une autre dans le monde des opérations. Ce dernier est une situation paramétrique. On exprime de la façon suivante cette interprétation :

Dans U':
" M. Smith, kinetic, concentrations, 5,56 µg/ml, T0, 12h32 " est interprété de la façon suivante ; T : [par. ($OP_{1,2,3,4}$) | s $\vDash$ << R(χ), ([R(α), a] $\rightleftarrows$ [R(β), b])>>]

Dans U'':
L'opération paramétrique est interprétée comme une situation :
T : [par. (OP_n)]: s $\vDash$ <R(lex.op. $_n$), State1, State2, t >

Par exemple : s $\vDash$ <<R (G, {α_1, …, α_n}), State (a), State (a, α_1), t >>
Pour
< Part of (Obj.,Nom.), inst. (Par.1)> $\rightarrow$ <inst. (Par.2), <inst. (Par.1), part of (Obj.,Nom.)>>

Dans U :
s $\vDash$ x, y, y$\rhd$ x

On peut maintenant, à l'aide de cette caractérisation de l'interprétation, spécifier comment sont interprétées différentes dimensions exprimées par des structures d'information.

L'interprétation d'un fait depuis la situation d'interprétation vers les individus (ou classification de l'information vers les individus) peut se représenter ainsi :
U': s_3 $\vDash$<Kinetic, <<mg/ml $\oplus$ 4.83 >, < M. Smith(x), 12:24(y) >>>

L'interprétation d'un état depuis les tokens vers les individus (depuis les résultats des opérations vers les individus) :
U: $s_{<ind \lhd part\ of\ ind>}$ $\vDash$ <M. Smith (x) $\lhd$ 12: 24 (y)>

L'interprétation d'un processus depuis le support vers les opérations (depuis la situation de représentation vers les résultats d'opérations) :
U'': $s_1 \lhd s_2 \vDash$ < < mg/ml, M. Smith(x) > $\lhd$ < 4.83, 12:24 (y)>>

Une dernière interprétation considère la spécificité de l'interprétation de la structure d'information. Il s'agit de l'inférence d'une conclusion à partir des trois interprétations précédentes :
s_4 $\vDash$<Kinetic, [$s_1 \vDash$ < mg/ml, M. Smith (x) >] $\lhd$ [$s_2 \vDash$ < 4.83, 12:24 (y)>]>

Cette formulation permet l'addition de n'importe quelle autre information (s_2) à droite avec les mêmes fonctions depuis s_3 vers s_4.

Nous reprendrons ici la caractérisation dynamique de l'interprétation, que nous avions associée aux flux, dans le cadre de l'activité.
Une distinction fondamentale entre l'interprétation de la structure d'information et de l'information dans l'activité est bien un rapport au temps : le temps associé à la structure d'information (et marqué par une primitive correspond à celui du prélèvement. Il réfère à un

instant de l'activité et à un état temporaire et partiel de l'individu de référence. Il est systématiquement étiqueté comme passé dans le temps de l'interprétation. La distinction relève alors de l'articulation entre valeur de vérité à un instant *t* et validité de cette valeur à un instant *t+n*.
En d'autres termes, un paramètre fondamental de l'activité est la distinction entre les dimensions temporelles associées à chacun des univers.

Dans le cadre de l'activité, il devient effectivement possible de réinterpréter les logiques dynamiques et les états que l'on a précédemment présentés. On peut également s'aider pour cela des logiques temporelles

Représentation de chacune des primitives de la structure d'information comme un dispositif particulier de référence.
Le schéma précédent (*Schéma des relations entre les différents mondes de l'activité)* ne représente que la distribution des différents mondes au sein du système d'information de l'activité.
Il ne caractérise pas la relation entre les unités symboliques, les univers et les objets construits de référence. (Par exemple, l'intensionnalité des échelles temporelles ne peut être décrite précisément dans le précédent schéma). Nous proposons donc maintenant une représentation de l'interprétation de chaque primitive à l'intérieur de chacun des mondes de l'activité.

A l'aide des schémas suivants, on représente :
- Pour chaque monde (colonne de gauche)
- La description des entités interprétées (collonne centrale)
- Les marques matérielles sur lesquelles se fondent cette interprétation (colonne de droite). Ces marques sont attestées dans les feuilles d'adaptation.

Cette représentation informelle permet de représenter sous forme planaire l'ensemble des entités réalisées empiriquement et mentales intervenant dans l'interprétation de la primitive constituante de l'information.
Enfin, cette représentation a comme objectif d'assembler différentes formes d'existence des primitives. Elle fournit en outre un cadre pour aborder les questions de cognition.

INDIVIDU :

U : (x)	x: ind. (x)	< "M. Smith" , espace dévolu>
U': s_1, $s_3 \vDash$	x [ind. (x)] *C*: x.A $\Rightarrow$ INDIVIDU	NP
U'': $s_{ind.}$	OP: < Obj.,Nom.>	(noms propres, noms d'obj

INDIVIDU et PARTIE DE sont des concepts incorporés ou « concevables (cf. D. Chalmers, voir infra).
Ils sont organisés comme une méréologie dans la dimension extensionnelle. Pour caractériser leurs relations, nous adoptons un vocabulaire adapté : LOC IN. LOC IN représente une relation d'inclusion régionale.

PARTIE D'INDIVIDU :

U: (x,y) :	x,y: part de ind. (y) ◁ ind. (x)	< “M. Smith” , « 12:24 »>
U’: $s_2 \vDash$	(x,y) [(part of ind. (y)), ind.(x)] *C*: y.B ⇒ PART DE [INDIVIDU] x.A ⇒ INDIVIDU *f*: PART DE [IND.] ⇒ INDIVIDU	NP\NP
U’’ : $s_{\text{<part de ind ◁ind >}}$	OP: <inst (Par.1), part of (Obj.,Nom.)>	(marqueurs temporels ou spatiaux)

PROPRIETE :

U: (x, y) :	x,y: Prop LOC_IN [part de ind. (y) ◁ ind. (x)]	< “M. Smith”, « mg/ml », espace dévolu>
U’: S1, S3 ⊨	x [<ind (x), prop (x)>] *C*: [(α∈Γ) ∧ (x.A)] ⇒ PROPRIETE x.A ⇒ INDIVIDU	(NP)\N
U’’: S_1	OP: <inst. (Par.2), <inst. Par.1, part de (Obj.,Nom.)>>	(concepts fonctionnels)

SPECIFICATION :

U: (x,y) :	x, y: [Spe LOC_IN part de ind. (y)] ◁ ind. (x)	<[“M. Smith, mg/ml, 4.83, 12:24 »], ligne>
U’: S1, S2 ⊨	xy [event (x, y), (<ind. (x), prop (x)>, <part of (y)), spe. (y)>] *C*: [(β∈Δ) ∧ (y.B)] ⇒ SPECIFICATION *f*: PROPRIETE ⇒ SPECIFICATION [(α∈Γ) ∧ (x.A)] ⇒ PROPRIETE *f*: PART DE [IND.] ⇒ INDIVIDU y.B ⇒ PART DE [INDIVIDU] x.A ⇒ INDIVIDU	(NP\NP)\N\N
U’’: $S_1 \rhd S_2$	OP: <inst. (Par.2), <inst. Par.1, part of (Obj.,Nom.)>>	(intégrales positives-négative,numéraux,…)

EVENEMENT :

U: (x,y)	: x, y: Event LOC_IN [part de ind.(y) ⊳ ind (x)]	<[“M. Smith”, Kinetic, mg/ml, 4.83, 12:24], ligne + espace sur la feuille>
U’: S1, S2, S3 ⊨	xy [(part de (y)), (<ind (x), event (x, y), prop. (x), spe. (y)>] *C*: χ ∈X ⇒ EVENT ((PROPRIETE ∧ SPECIFICATION) ∧ (PART DE [IND.] ⇒ INDIVIDU)) ⇒ EVENT [(β∈Δ) ∧ (y.B)] ⇒ SPECIFICATION *f*: PROPRIETE ⇒ SPECIFICATION [(α∈Γ) ∧ (x.A)] ⇒ PROPRIETE *f*: PART OF [IND.] ⇒ INDIVIDU y.B ⇒ PART OF [INDIVIDU] x.A ⇒ INDIVIDU	(NP\S)/N
U’’: [S_1 ⊳ S_2] S_3	OP: <structure d’ information, duplication>	(nom d’événement)

Cette représentation permet uniquement la reconnaissance des traits de chaque primitive, mais pas la représentation de la structure d’information elle-même. En effet, elle n’intègre pas la situation s_4 et les règles de dépendance.

Vers une nouvelle formulation des logiques locales.

Nous avons laissé en suspens la question des logiques locales tout simplement parce que nous ne disposions pas de la modélisation des mondes, qui rend effectivement possible la caractérisation des raisonnements en ce que leurs conclusions caractérisent des tokens.

Nous obtenons des régularités de comportement à partir des différents niveaux d’analyse précédents. Ces régularités sont expliquées dans le cadre des primitives. Nous présentons donc maintenant les inférences qui peuvent être réalisées dans le cadre de l’activité et considérant les flux.

Nous suivons le principe de l’accessibilité : une forme linguistique ou une inférence dans une activité ne peuvent être considérées directement comme un fait cognitif, mais comme une façon d’accéder aux ressources cognitives.

L’articulation des dimensions matérielles des opérations et des dimensions linguistiques nous a permis d’accéder à un niveau d’analyse plus abstrait, dans lequel l’information est une composition particulière de primitives.

Ces primitives représentent des facultés humaines ontologiques. Nous décrivons maintenant comment elles contribuent à la caractérisation du raisonnement à distance, par l’inférence d’une représentation augmentée de l’état référentiel.

L’inférence à distance signifie que le raisonnement se fait sans accès physique au monde sur lequel il s’exerce, considérant également un cadre spatial et temporel différents. Le type de situation et les formules de l’univers intensionnel permettent d’inférer une situation plus large en intension que celle caractérisée par l’interprétation de la structure d’information.

Nous reprenons donc maintenant le cadre fourni par la théorie des flux : théories et logiques locales. Nous l’appliquons aux primitives, de façon à rendre compte de l’élargissement du monde de la référence à partir du moment où un raisonnement impliquant les théories et les

logiques locales est développé. C'est à cette condition que l'on peut envisager la mise en œuvre de raisonnement professionnels impliquant des concepts et des règles inférentielles qui ne sont pas présentes dans les théories et logiques locales. Rappelons que ces théories et logiques locales ne concernent qu'un raisonnement interprétatif[217].

Nous reprenons donc les primitives comme unités d'analyse. Nous considérons que les théories et logiques locales permettent de caractériser de la façon la plus exhaustive possible ces primitives.
A gauche, nous présentons le raisonnement de façon informelle, et à droite sa caractérisation utilisant les théories et logiques locales.

INDIVIDU :

Régularités de réalisation	Contribution aux théories et logiques locales
NP: A = [a | s ⊨<< R(χ), (a)>>] *AND* (x: a): [<Objet, nomination> → <(a), X>] *THEN* x : IND(x)	NP : (x:a) ⊢ [a.A] : x

PARTIE DE :

Régularités de réalisation	Contribution aux théories et logiques locales
NP\NP: B = [b| s ⊨ <<R(χ), [(a,b)]>>] *AND* (y: b (x: a)) : Opération 1 = Lex(B): [<(a), X> → < (a, b), X>] *THEN* x,y: PART DE [IND] (y) ◁ IND (x)	NP\NP : (y:b (x:a)) ⊢ [b.B] : y ⊳ x

PROPRIETE:
Régularités de réalisation

(NP)\N: Γ = [s ⊨ << <R(α), a>>]
AND
(Γ: {α_1, …, α_n}, [a.A]) :Opération 2 = Lex (Γ): <(a), X> : < [α (a)] , X >
THEN
[x : a] : PROP [IND(x)]

Contribution aux logiques locales :

(NP)\N :
Γ: {α_1, …, α_n}, [a.A]) ⊢ Δ: {β_1, …, β_n}, B
: x (OP2 (a(x)))

[217] Nous verrons, lorsque nous aborderons notre projet, comment il est possible d'articuler des connaissances professionnelles et des flux, à savoir des raisonnements limités à l'interprétation de l'information. Nous verrons alors à quoi peut servir une ontologie de domaine.

SPECIFICATION
Régularités de réalisation

(NP\NP)\N\N : Δ = [s ⊨ <<[R(α), a] → [R(β), b]>>]
AND
((α), (Δ: {β 1, …, β n}, [b.B]) : Opération 3 = Lex (Δ): < [α (a)] > → <[α (a)] ⊕ [β(b)]>
THEN
[[x : a], [y : b]] : SPE [PART DE [IND] (y)] ⊳ PROP [IND(x)]

Contribution aux théories :

(NP\NP)\N\N :
Δ: {β 1, …, β n}, [b.B] ⊢ X: {χ 1, …, χ n})
: y (OP3 (b(y))) ◁ x (OP2 (a(x)))

EVENEMENT :
Régularités de réalisation

(NP\S)/N : X = [s ⊨ << R(χ), [R(α), a] ⇄ [R(β), b]>>]
AND
X: {χ 1, …, χ n}) :
((χ), ((Γ: {α1, …, αn}), (Δ: {β 1, …, β n}), ([a.A], [b.B]))) : Opération 4 (duplication) = Lex(X):
<[α (a)] ⊕ [β(b)]> → [χ(α (a)) ⊕ (β(b))]
THEN
x : EVENT(x), [SPE [PART DE [IND]] ◁ PROP [IND(x)]]]

Contribution aux théories :

(NP\NP)\N\N : X: {χ 1, …, χ n}) ⊢ X: {χ 1, …, χ n})
: x, [y 1, …yn]

L'intérêt d'une représentation utilisant les primitives réside d'abord dans le fait qu'une interprétation peut ne pas concerner l'ensemble d'une expression, mais certains de ses composants.
Egalement, la mémoire des opérations est intégrée à l'intérieur des théories, ce qui permet d'argumenter l'hypothèse d'une dimension cognitive des opérations.
Enfin, les théories permettent de rendre compte des contenus informationnels au sens de la théorie de l'information. Sur notre exemple, ils permettent au mieux de rendre compte des raisonnements probabilistes mis en œuvre dans l'activité.

Utilisation de la théorie des situations pour caractériser une grammaire de traits.
Les HPSG constituent une grammaire de traits unifiée. Dans son volet sémantique, elle utilise la théorie des situations. On pourrait envisager d'utiliser ce modèle pour extraire des structures d'information relativement à certaines situations prédéfinies. Nous y reviendrons lorsque nous présenterons notre projet d'application.

Conclusion.
La structure d'information, telle que l'on vient de la caractériser, permet de mettre en relation des réalisations linguistiques et des connaissances. La structure ne représente pas les connaissances mais indique de quelle façon ces dernières peuvent être liées à des structures

caractérisant des connaissances. En ce sens, nous nous distinguons fortement des propositions de E. Valduvi.
Telle que l'on vient de le voir, la structure d'information permet de représenter la façon dont des connaissances s'appliquent sur des phénomènes du monde, accessibles au travers d'opérations de symbolisation.
En ce sens, notre travail propose une réponse à la question du lien entre les connaissances et les objets du monde[194].
La structure d'information telle qu'on vient de la définir fixe la quantité et l'organisation des données de façon à ce que soit représentée et interprétée une information. L'application alors serait l'extraction d'information. (On entend ici par extraction d'information l'ensemble des travaux opérant à partir de textes semi-structurés, utilisant notamment des balises[195]. Ces travaux reposent sur l'identification d'un objet et de relations syntaxiques associées à cette structure. On construit ainsi des patterns dont l'un des intérêts consiste à permettre une extraction multi-entrées. Cette méthode a été raffinée dans l'élaboration de « wrappers »[196]. (Un wrapper est une procédure permettant d'extraire le contenu d'une ressource[197]. L'induction de wrapper est une méthode pour construire automatiquement des wrappers. De telles propositions, d'abord associées aux textes semi-structurés, peuvent être étendues à des textes non structurées comme les textes annotés[198]).

Sans entrer plus précisément dans ces questions, notons tout de même que l'extraction d'information utilise des concepts relativement proches de ceux qui sont familiers à ce travail. Historiquement, la problématique de l'extraction d'information[218] est posée par HARRIS et est liée directement à celle de la prédication, donc de l'informativité[199].

L'extraction d'information est d'abord une problématique liée à une tâche. Par conséquent, les choix qui sont faits sont d'abord liés à des questions techniques, situées dans le cadre de l'informatique : modèles supervisés ou pas, apprentissage automatique, utilisation de modèles vectoriels ou pas, etc. Les propositions sont évaluées dans des campagnes.

On pourrait poursuivre l'hypothèse d'une modélisation de la structure d'information comme outil d'actualisation de connaissances dans le cadre d'une activité. Dans ce cadre, on peut effectivement rendre compte de la diversité de points de vue sur le monde en s'aidant de l'information. Ainsi, on répond à la question posée par la deuxième situation présentée en introduction, à savoir le fait qu'une information pouvait donner lieu à des interprétations différentes et complémentaires tout en ayant une seule signification.

Nous examinerons la modélisation de la structure d'information dans le cadre de notre projet dans la dernière partie du travail.

5.3. Hypothèses sur les fondements des flux dans les activités ordinaires.

Nous traitons maintenant d'une question qui peut apparaître relativement éloignée des bibliothèques numériques. Or, comme nous l'avons présenté en entrée, l'intégration des bibliothèques numériques dans le cadre du web de données entraîne la mise en relation

[218] Nous empruntons à Ralph Grishman la définition de l'extraction d'information que nous utiliserons : "Information extraction (IE) is the process of identifying within text instances of specied classes of entities and of predications involving these entities » (p.3).

d'objets et de structures hétérogènes. Nous présenterons ici les principes théoriques de la cognition située et distribuée, puis la façon dont ils peuvent contribuer à élaborer un modèle de l'activité pertinent pour l'intégration des flux. Nous terminerons cette partie par une présentation des enjeux de cette approche par rapport à la question des bibliothèques numériques.

Pour comprendre le problème dont il est question ici, à savoir comment conceptualiser cette hétérogénéité, il apparait particulièrement important de se référer à des ressources permettant de comprendre ce qu'est la distribution et l'hétérogénéité. C'est pour cela que nous faisons ce détour par les théories de la cognition située et distribuée.
La notion d'activité aura une autre fonction. Elle servira à caractériser les métadonnées que l'on produit à partir de la caractérisation d'une activité (et qui ont trait au contenu des documents) distinctement de celles qui sont liées à la seule description d'un objet comme le document.

Si les flux sont des contraintes sur le langage, comment alors ces contraintes peuvent être expliquées ? Comment peut-on caractériser les flux dans le monde ? (Ces questions peuvent sembler non-pertinentes si l'on ne considère que la capacité du modèle à relier des structures hétérogènes, mais deviennent primordiales lorsque l'on envisage de donner aux flux une dimension descriptive et un fondement théorique hors de la seule logique mathématique).
Notre hypothèse consiste alors à présenter les flux comme la représentation d'un phénomène relatif à l'activité humaine, et donc pouvant être étudié en considérant que l'activité cognitive n'est pas seulement un phénomène mental, mais se caractérise par son externalisation dans des outils, des objets du monde et des interactions.

Le choix des exemples et de la pharmacie permet une construction de données qui ne soit pas seulement linguistique mais qui repose sur une modélisation du contexte de formation, de diffusion d'interprétation des structures d'information. Un tel programme n'est peut-être pas suffisant pour construire un objet satisfaisant en anthropologie cognitive. Néanmoins, notre caractérisation de l'activité dans laquelle s'insère le système d'information apparaît suffisamment élaborée pour permettre un certain nombre d'emprunt théoriques à cette discipline.

En d'autres termes, peut-on donner une explication des flux qui ne soit pas limitée à la productivité du modèle? Cette question reprend les propositions formulées au début de ce travail et qui fondent l'appréhension pluridisciplinaire de l'information. Comme l'information dépend de flux, d'opérations « dans le monde » et de manipulations de symboles, on peut raisonnablement faire l'hypothèse d'un fondement cognitif à l'information. On peut alors considérer qu'il s'agit d'un raisonnement se manifestant de façon solidaire entre des dispositifs et des processus hétérogènes. On rejoint alors les principales hypothèses de la cognition située et distribuée.
Il ne faut pas considérer l'apparition de références à l'anthropologie cognitive et à la cognition située et distribuée comme une rupture théorique et méthodologique, mais comme une façon de fonder les domaines et les mondes de la référence. En effet, dans le cadre d'une activité, ils apparaissent beaucoup plus complexes et structurés que pour l'analyse d'expressions considérées uniquement dans leur dimension linguistique.

Par ailleurs, on pourra distinguer les raisonnements qui fondent les flux de ceux qui se situent à un niveau d'abstraction plus élevé. En effet, les flux proposent des inférences de moindre niveau d'abstraction que ceux que propose la représentation des connaissances. Cette

distinction se retrouve également dans les langages du web, entre d'une part les possibilités de RDF et celles d'OWL. Nous aimerions également ici proposer quelques pistes pour fonder cette distinction.

Pour expliquer plus précisément la démarche, on peut revenir sur une question essentielle : que représentent les flux ? Une réponse a été fournie par J. Jayez et A. Mari[200], qui considèrent que les flux représentent une relation de causalité entre deux ensembles hétérogènes de données, considérant que cette causalité se caractérise par l'action d'un constituant sur l'autre. J.Jayez et A. Mari appliquent ce principe à des entités nominales liées par le connecteur « avec ».
Enfin, ce qui nous intéresse plus particulièrement, c'est le rapport entre flux et action, autrement dit information et action. Pour cela, nous allons revenir plus précisément sur la question des opérations, dont on ne sait pas jusqu'à présent ce qu'elles représentent. Ces opérations caractérisent les unités minimales de l'activité cognitive pour G. Hutchins. Ce qui signifie que l'on caractérise la production de l'information au travers d'un raisonnement, constitutif des maniements de symboles, mais qui ne peut être expliqué seulement par des mécanismes linguistiques. Nous allons donc pouvoir caractériser plus précisément comment les flux sont intégrés dans une activité.

Caractérisation de l'interprétation et des connaissances.
Cette présentation a un but précis relativement aux sciences de l'Information et plus généralement aux théories de l'information. Les travaux de P. Ingwersen [201], comme ceux de M. Burgin, reposent plus ou moins directement sur l'ontologie de C. Popper[202]. Or celle-ci est fondée sur des caractérisations de la connaissance qui ne prennent pas en compte l'utilisation de cette connaissance dans le cadre des cadres d'activité et donc d'usage.
Néanmoins, les propositions de M. Burgin [203](pp. 20-21) distinguent clairement la cognition considérée comme un phénomène de l'esprit de l'information et de la connaissance (considérées comme constituant un monde extérieur à celui de la pensée individuelle).
Nous préciserons donc ces modèles en utilisant le cadre théorique de la cognition située et distribuée : à partir du moment où l'on considère que la cognition s'appuie sur des artifices extérieurs à l'esprit pour s'exercer, cette dichotomie a quelques difficultés à être maintenue

Une autre distinction importante par rapport au cadre proposé par M. Burgin est le fait que l'on spécifie le contexte dans lequel l'information s'insère. M. Burgin parle de système d'information. Dans son cadre, le lien entre l'information et le monde réel dans lequel celle-ci trouve un usage est inexistant. Notre point de vue a consisté tout au long de ce travail à associer l'information à des processus dans lesquels elle prend place. De cette insertion, on a émis l'hypothèse que c'était bien l'activité qui était porteuse de dynamique, et qu'il était possible de la considérer comme une dynamique de succession d'état. Cette dynamique est alimentée par l'information. Nous avons utilisé les successions d'état de S. Kripke pour représenter cette dynamique. Maintenant, nous pouvons argumenter plus précisément cette hypothèse en caractérisant ce que l'on entend par activité.
Enfin, cette caractérisation de l'activité a comme autre enjeu de situer les questions d'information dans un cadre socialisé. On considère habituellement les modèles de l'information, voire plus précisément ceux de la recherche d'information (comme par exemple ceux de P. Ingwersen (op. cit.) en dehors d'autres processus concrets comme ceux de la vie professionnelle. Or, ce sont eux qui déclenchent la recherche d'information (ou dans notre cas qui la requièrent). (Nous avons évoqué plus haut la mutation que constitue l'introduction d'appareils mobiles et portables permettant de procéder à des recherches d'information dans le déroulement des activités professionnelles). . Notre proposition s'inscrit

dans le cadre d'une intégration des processus informationnels dans les processus de la vie sociale.

Caractérisation de l'interprétation.
Jusqu'à présent, nous avons considéré l'interprétation dans le cadre d'une sémantique formelle. En effet, la théorie des situations et les flux limitent l'interprétation à une position et des propriétés de raisonnement.
Or, l'un des intérêts de la cognition située et distribuée est bien de structurer cette position. Néanmoins, comme nous le verrons, il n'est pas aisé d'utiliser directement ces outils anthropologiques pour spécifier l'interprétation telle qu'on la conçoit et la fonder dans le cadre d'une activité.
Justement, un certain nombre de travaux ont émergé afin de systématiser les principes de la cognition située et distribuée. Nous présenterons dans cet objectif les travaux de Chalmers & Clarke[204]. Les paires de Chalmers ne constituent pas simplement une extension des principes de cognition située et distribuée. L'implication qu'elles peuvent avoir sur la caractérisation des modèles de structuration des données est essentielle parce qu'elle permet de dédoubler ce qui présenté et ce qui est effectivement interprété.

5.3.1. Cognition située et distribuée : présentation générale.

Nous avons précédemment évoqué les travaux de G. Hutchins pour caractériser notre objet de travail. Ce qui nous intéressait chez G. Hutchins, c'était la précision de la modélisation des opérations distribuées. Le partage entre l'activité humaine réalisée dans l'esprit et celle externalisée dans des outils ou par des repères situés dans le monde constitue un point crucial de cette théorie. Enfin, G. Hutchins recourt à des niveaux d'abstraction différents pour caractériser l'activité cognitive, suivant en cela D. Marr. Cette caractérisation de l'activité en niveaux d'abstractions est cohérente par rapport à notre démarche générale.

Comme on a pu le voir, la cognition située et distribuée constitue un programme encore largement en recherche de ses fondations, et qui reste fondé sur des analyses d'anthropologie culturelle. Il s'agit de l'une des seules sciences humaines et sociales à construire des concepts à partir des objets matériels et à étudier leurs relations à la vie sociale et économique d'une communauté humaine.
Comment peut-on caractériser les flux dans le cadre de la cognition située et distribuée ?
Une telle question pourrait apparaître relativement secondaire si l'on n'aurait pas posé comme fondement méthodologique letransfert de connaissances et de méthodes entre des domaines d'application différents. Théoriquement, on propose que des théories de la cognition, fondées sur des observations, peuvent aider à concevoir et formuler des modèles. Plus empiriquement, on avait illustré cela avec la possibilité d'utiliser les résultats d'analyses menés en pharmacie hospitalière vers les problématiques du web de données (ou web sémantique).
On a postulé également que nos exemples constituent des illustrations d'un raisonnement beaucoup plus général concernant l'information. De la même façon qu'E. Hutchins valide ses hypothèses sur deux exemples éloignées de navigation (pirogues en Micronésie et porte-avion dans la rade de San Diego), une modélisation déduite d'un exemple concret permet de fonder une certaine généralité au propos.

De façon très sommaire, on distingue plusieurs approches de la cognition :
- L'approche symbolique. Le traitement de l'information opère sur des données symboliques à l'intérieur de la structure cognitive. Ces données symboliques représentent des concepts.

- L'approche réaliste fonctionnelle. L'activité cognitive réside d'abord en un travail relationnel relatif à l'environnement. Il s'agit de traiter (notamment par une mise en relation) des phénomènes environnementaux. La perception constitue alors un élément fondamental de l'activité cognitive. Le traitement de l'information consiste alors à discriminer et abstraire ce qui est perçu.
J Barwise et auparavant F. Dretske s'inscrivent dans cette hypothèse et c'est sur la base de travaux à propos de la perception (comme les Affordances de Gibson) comme organisation de la vision que naissent les classifications et les mises en relation d'expériences. (Bien évidemment, l'opération classificatrice de ces auteurs n'a pas la même désignation en anthropologie cognitive, où l'on identifierait une catégorisation).

Nous empruntons à P. Sallembier[205] la distinction entre cognition située et distribuée : « Une conception "située" de la notion d'action […] insiste sur la détermination de l'action par différentes variables situationnelles, limite le rôle fonctionnel des plans et remet en cause l'existence de représentations symboliques internes comme support des activités cognitives. (…) La finalité [du] cadre général [de la cognition distribuée] est de dépasser le niveau d'analyse classiquement adopté en science cognitive (l'individu) et de parvenir à la caractérisation d'une cognition située et incarnée (« embodied ») dans son contexte d'occurrence, ce qui signifie considérer la cognition en ce qu'elle a de distribué entre agents et éléments de la situation. » (pp. 2-4/14)

Emergence de modèles distribués utilisables dans le cadre de la conception.
Le principe de la distribution stipule qu'il existe un environnement matériel socialement organisé afin de permettre l'effectuation d'inférences justes. Il permet de caractériser l'hétérogénéité des objets entrant dans la structuration d'un cadre de raisonnement.

La distribution repose sur une modification du cadre fonctionnel associé au raisonnement, par une caractérisation relationnelle des liens entre les entités. De telles propositions initiales modifient considérablement la façon dont on peut définir l'activité cognitive, notamment celle du cerveau, mais également le rôle et la structuration des discours.
Un fondement du modèle est la caractérisation spatiale des processus inférentiels : ils se déroulent dans l'espace et font intervenir des unités hétérogènes. Ils sont inscrits à l'intérieur de l'interaction entre des agents et opérateurs distincts, d'une part, et dans leurs appuis mutuels pour la réalisation d'une inférence. Par conséquent, ces agents et opérateurs étant hétérogènes (machines, médias, individus), l'explication par leur structure interne ne peut être la seule pertinente.
Par ailleurs, ces modèles s'inscrivent à l'intérieur de processus caractérisés par des transformations d'états expliqués par des opérations. Ces modèles fonctionnels sont donc dynamiques.

Comme le remarque G. Hutchins[206], l'origine du courant de recherche est une convergence entre les sciences humaines et sociales et un certain nombre de modèles de la cognition destinés à l'IA : la « société de l'esprit » de M. Minsky, les modèles connexionnistes de Rumelhardt.
La cognition située et distribuée retient de ces travaux trois concepts essentiels :

- la temporalité,
- l'interaction entre les systèmes,
- les systèmes localisés et fonctionnellement limités.

Néanmoins, dans la perspective de G. Hutchins, ces modèles offrent éventuellement quelques outils conceptuels, mais leur rôle s'arrête face au postulat théorique et méthodologique fondamental : dès lors que l'on cherche à intégrer dans la caractérisation de la cognition les dimensions de l'environnement, de la culture et de l'histoire, il est nécessaire d'opérer une démarche depuis l'environnement vers l'intérieur de l'individu, en considérant non celui-ci comme le lieu du traitement de l'information, mais comme un composant de celui-ci. Il considère la distribution comme un outil méthodologique permettant de traiter un objet extrêmement complexe en isolant différents lieux et modes de traitement de l'information.

La cognition située et distribuée est organisée autour de postulats dont A. Clark & D. Chalmers, dans leur article « The Extended Mind »[207], et également plus récemment dont R. Wilson & A. Clark[208] ont essayé de donner une unité.
Pour l'essentiel, cet article constitue un ensemble de propositions méthodologiques pour les sciences cognitives et l'amorce d'une réflexion sur l'esprit, lequel n'abrite pas tant des croyances que la capacité de produire des relations entre des entités internes et externes, sans que celles-ci aient à être soit spécifiquement internes ou spécifiquement externes : une mémoire peut être externe (notée dans un calepin) ou interne. Ce qui importe est la caractérisation de la relation entre les dimensions externes et internes.
La position est alors phénoménologique, dans la mesure où le travail de l'esprit se caractérise d'abord par la production d'une expérience en reliant une mémoire à un programme. La théorie de l'esprit étendu est d'abord un travail philosophique ; il est donc dégagé des limites disciplinaires de travaux comme ceux d'anthropologie cognitive mais également ceux qui vont essayer de fonder sur une base perceptuelle la cognition située et distribuée[209].

Ces travaux fondent un programme de recherches cognitives visant à définir l'esprit en intégrant l'ensemble des objets produits par l'activité humaine intégrés à la réalisation d'actions individuelles et collectives. Le premier point défini est celui d'externalisme, le second étant l'aspect actif de cette cognition externe, à savoir le fait que l'on ne s'appuierait pas seulement sur des éléments externes pour réaliser un raisonnement, mais bien que ces éléments externes réalisent le raisonnement en question.

Cette caractérisation a un certain nombre de conséquences sur la nature de l'explication de la cognition.
Dans le cadre, des Sciences Cognitives, une théorie doit être évaluée relativement à la dimension neuronale de la cognition ; celle-ci requiert une certaine plasticité des mécanismes neuronaux, tels qu'ils puissent s'adapter à des situations de partage de tâche ou de raisonnement.

Cognition distribuée et intelligence distribuée.
Enfin, on reviendra sur un certain contexte des sciences cognitives qui a amené l'émergence de ces problématiques de cognition située et distribuée. Nous en citerons deux plus particulièrement : la conception d'outils informatiques intelligents et l'apprentissage.

Suivant W. Clancey (op. cit.), il convient de ne pas oublier l'importance que joue le développement de nouveaux outils intelligents dans l'émergence de telles méthodes. Outre les questions de dialogue homme-machine, l'émergence d'outils de CSCW amène à penser plus précisément la menée collective d'activité, et notamment le partage des raisonnements.

W. Clancey (op. cit.) fait remonter la question de la cognition située et distribuée en l'associant à une remise en cause de l'association ambiguë entre caractérisation des

connaissances d'une part et langage. Le premier intérêt de ce travail, c'est de déplacer la représentation des connaissances depuis des modèles abstraits (comme les modèles de domaines) vers des raisonnements incorporés et socialisés. Parallèlement, la réflexion se déplace depuis des questions de représentations abstraites vers des outils manipulés dans le cadre de communications. Cette position sera importante pour la définition des ontologies et de leur rapport aux représentations de connaissances. Pour N. Guarino (*Formal Ontology*, op.cit.), la position de W. Clancey permet de caractériser l'ontologie comme une représentation partielle du domaine, en lien à une finalité précise, extérieure à l'esprit. On ne peut donc limiter l'ontologie au seul domaine conceptuel d'un acteur mais à la totalité des acteurs en interaction dans le cadre de l'activité.
Cette position permet de fonder les ontologies dans un certain rapport au monde réel, ce qui permet donc de les distinguer des représentations de connaissances.
Néanmoins, les ontologies ne reprennent pas l'ensemble des principes de l'IA distribuée. N. Guarino se sert de la critique de W. Clancey pour caractériser sa conception du domaine par opposition à la représentation des connaissances (qui se veut à la fois exhaustive et indépendante des faits sociaux et du monde).

Positionnement de l'hypothèse de la cognition étendue.
En premier lieu, la cognition située et distribuée ne constitue pas nécessairement une alternative à la cognition symbolique, mais une augmentation des objets impliqués, donc des dimensions de cette cognition. Il s'agit non d'intégrer dans les raisonnements des éléments externes, mais d'intégrer dans la cognition des objets externes. On pourra alors fonder la cognition non sur de seuls postulats psychologiques, mais sur des faits ancrés.

Le modèle psychologique n'est plus au centre du modèle de la cognition ; bien plus, il s'agirait d'un modèle interactionniste, assez proche de celui proposé par H. Garfinkel.

La cognition étendue repose sur l'idée que des expansions de l'esprit humain peuvent réaliser des opérations cognitives ; en effet, l'esprit déchargerait certains de ses raisonnements dans des machines. Ces machines ont alors une compétence cognitive, qui faciliterait largement le travail humain. Les exemples se situent en partie en robotique, en partie aussi, en anthropologie (voir G. Hutchins notamment).

Cognition située et distribuée et validation d'hypothèses.
La base d'un travail d'anthropologie est la caractérisation de la façon dont un groupe humain construit son rapport au monde. Cette anthropologie devient cognitive dès lors que l'on s'intéresse à la façon dont ce rapport mobilise ou élabore des connaissances. R. D'Andrade[210] envisage ce champ d'activité de l'anthropologie sous deux angles communs à toute approche classique en sciences cognitives : la construction collective de la perception et les raisonnements effectués.
Ces recherches reposent sur l'élaboration de modèles conceptuels devant rendre compte à la fois de la dimension interne et externe de la cognition. Citons G. Hutchins : « There are two principal ways to achieve stability in conceptual models. First, the conceptual models that anthropologists call cultural models achieve representational stability via a combination of intrapersonal and interpersonal processes. Second, the association of conceptual structure with material structure can stabilize conceptual representations. This is an old and pervasive cognitive strategy"[211].
La validation de propositions en anthropologie se fonde sur des critères de validité propres :

- transparence collective,

- confirmation matérielle,
- reproductibilité,
- identification d'objets.

De tels critères de validité impliquent que chaque hypothèse relative au système cognitif ou à la menée d'un raisonnement soit assortie d'un observable dans le monde.

Contestation de l'individualisme méthodologique.

Les théories de la cognition située et distribuée reposent sur la contestation d'un fondement de la cognition classique, à savoir l'individualisme méthodologique, auquel est opposé l'externalisme socialisé. Nous développons cette articulation avant de présenter ses conséquences sur certains concepts fondamentaux des sciences cognitives.

On ne discutera guère les arguments biologiques, physiologiques et neurologiques de ces théories. Notre but consiste seulement à expliciter les fondements d'une théorie qui sert de cadre à notre objet d'études.

Globalement, la cognition située et distribuée repose sur une transformation de la dimension de l'objet des sciences cognitives. Le raisonnement, les connaissances ne sont plus contenus dans le cerveau individuel, mais à la fois en lui et dans ses liens avec le monde extérieur ; celui-ci intervient activement dans la transformation d'états.

L'individualisme méthodologique, à savoir le fait que l'on considère l'esprit individuel comme objet d'étude de la cognition est la principale remise en question effectuée par la cognition située et distribuée. Néanmoins, ces travaux restent dans le cadre des sciences cognitives, et reformulent les principaux concepts du domaine, notamment les états mentaux.

Or, on peut contester la pertinence même de la notion d'état mental et envisager les représentations comme des processus -(Havelange, V., Lenay, C., & Stewart, J. (2002). *Les représentations: mémoire externe et objets techniques*, op. cit.).]. De telles propositions restent problématiques si l'on envisage la question du langage (et des discours) et de leur place dans l'activité cognitive.

Envisager le langage comme un seul usage restreint l'analyse à celle de processus purement occurrents.

Questions de langue et d'activité. Adéquation des différents niveaux d'analyse.

La navigation et l'adaptation de posologie constituent des activités humaines. Or la langue (ou tout du moins les phénomènes qui nous concernent) requiert une autre dimension de généralité. Il existe donc un risque de conflit entre des dimensions hétérogènes. Par ailleurs, considérant que les flux reposent sur des constructions symboliques, comment se situe l'information, et plus précisément le flux par rapport à l'activité ?

L'activité pourrait être considérée comme un modèle contextuel dans lequel se réalise l'information. Or, les modèles de l'activité dont on dispose représentent une l'émergence d'un objet[212], et se situent à un niveau d'abstraction trop élevé pour rendre compte des flux.

Ainsi, on ne propose pas un modèle général de l'activité. Comme pour le langage, on préférera une problématisation partielle. On peut tout simplement faire l'hypothèse que les flux traduisent la façon dont les professionnels de la pharmacie construisent leur regard sur le monde. Nous suivrons donc une perspective proche de celle de C. Goodwin. Nous développons le cadre et la portée de ces travaux.

Les modèles de l'activité reposent sur une critique essentielle de l'ethnométhodologie : celle-ci ignore les dimensions constantes des connaissances déposées. En se focalisant sur l'interaction, et la résolution locale des conflits, les règles structurelles restent totalement ignorées.
Or, à partir du moment où l'on accepte le postulat que la cognition est externalisée, ou même, plus simplement, que les objets sont dépositaires d'une certaine connaissance, les objets de l'analyse des interactions sont considérablement élargis. En effet, les outils et autres objets sont marqués par une constance, à savoir qu'ils ont des fonctions régulières. Dès lors, l'analyse interactionniste peut s'appuyer sur des objets stables pour la menée de ses propres travaux. C'est de ce couplage entre interactionnisme et dispositifs d'objets et outils que naissent les travaux de G. Hutchins, C. Goodwin, notamment.
Ainsi, une analyse anthropologique (intégrant notamment les objets, la construction sociale de phénomènes naturels) a pu montrer que les interactions, considérées au travers du lien entre les hommes, les outils et les objets du monde, pouvaient caractériser des compétences autres que celles de la communication sociale.

Ce travail sur le regard professionnel[213] (on peut également citer A. Cicourel, pour un point de vue plus sociologique[214]) vise aussi à spécifier ce qui ressort de la construction du monde par une profession, et qui se caractérise par l'extraction et le traitement de certains événements afin de les placer à l'intérieur de certains objets phénoménaux autour desquels les discours sont construits, et les activités sont menées. Ainsi, on peut analyser la perception d'un événement comme n'étant pas le fruit d'un processus psychologique transparent, mais le résultat d'une activité construite socialement et finalisée. C. Goodwin (p. 606) étudie plus particulièrement les différentes étapes de la formation de cette vision :

- codage (ou transformation des phénomènes observés dans des objets de connaissance qui vont alimenter le discours de la profession,
- soulignement de certains phénomènes afin de les mettre en relation de façon privilégiée,
- production de représentations matérielles permettant de synthétiser les perceptions du monde et de produire un objet sur lequel travailler.

Ces trois pratiques sont préalables à l'activité elle-même, et constituent des objets de connaissances assurant la médiation entre un domaine d'exercice et une activité spécifique. C. Goodwin les dénomment des pratiques discursives. Elles articulent le langage aux pratiques d'action.

Revenons sur ces trois types de pratique :
Le **codage** consiste en une pratique systématique permettant de transformer le monde en catégories et événements pertinents pour le travail professionnel. Il s'agit d'un travail de dé linéarisation et de circonscription des faits de façon à constituer un ensemble de faits comparables. Le codage utilise des outils de normalisation ; ainsi, en archéologie, toute pièce trouvée est classée selon sa forme, sa couleur, sa texture, sa consistance. Un de ces outils est la table de MUNSELL, qui encapsulent fonctionnellement les pièces dans un seule couleur. Ainsi, la perception d'un objet est construite par sa satisfaction d'un critère distinctif unique.

Le **soulignement** consiste à établir une figure à partir d'un fond. Il s'agit de segmenter un espace de façon à marquer les événements pertinents pour une activité. Il ne s'agit pas de transformation des objets de travail, mais seulement de structurer l'environnement matériel en fonction de sa pertinence.

Ces marques mettent en évidence un trait commun entre des objets marqués, qui par ailleurs pourraient sembler disparates. Les liens entre ces éléments disparates construisent des pratiques ultérieures. Le soulignement est une opération préalable à la mise en œuvre d'une action.

La **représentation graphique comme incarnation d'une pratique** consiste à établir une représentation simplifiée du monde en y inscrivant ses propres activités, et notamment certains commentaires sur son propre travail. Ainsi, par exemple, un plan de coupe constitue une représentation graphique à l'intérieur de laquelle un archéologue place ses propres conclusions à propos des phénomènes représentés.
La connaissance est inscrite à l'intérieur des représentations graphiques de façon à ce que les résultats d'un raisonnement soient directement associés à des objets. Cette association permet d'engager un autre raisonnement, notamment des demandes d'information à propos du monde naturel.
La représentation graphique permet de construire un environnement interprétatif, tel que d'autres analyses puissent être menées. Elle constitue donc à la fois une mémoire des opérations réalisées et une condition à la réalisation d'autres opérations.
C. Goodwin insiste sur le rôle de la représentation graphique pour la construction d'une représentation commune, y compris la dimension technique de la production de la représentation (pp. 614-615). Il met en évidence un certain nombre d'échanges dans lesquels la construction de la représentation constitue une mise en œuvre des compétences professionnelles.
Ainsi, le regard professionnel se caractérise par la construction d'une représentation du monde naturel qui elle-même permet de réaliser d'autres activités (y compris des jugements) sur le monde naturel. Ainsi, les discours et les représentations dirigent les opérations réalisées sur le monde naturel.

On retiendra qu'un regard professionnel consiste d'abord à construire des objets dans le monde, et cette construction précède et accompagne l'activité elle-même.
Dans les exemples choisis par C. Goodwin, l'univers sur lequel s'exerce l'activité est remarquablement immobile et circonscrit, et évite la dynamique même des phénomènes du monde. Par ailleurs, on a chez C. Goodwin la façon dont une profession s'approprie les objets perçus, mais sans ne jamais considérer comment des représentations symboliques peuvent transiter à l'intérieur d'ensembles de professions plus larges. A ce moment-là, on peut difficilement éviter le recours au langage verbal, et donc une symbolisation qui dépasse largement le cadre d'une seule pratique professionnelle.
Dans les propositions de C. Goodwin, le discours est déterministe, au sens où il est construit afin de résoudre certains problèmes, et donc d'engager ou de conduire une certaine activité. Néanmoins, il n'explique pas le fait qu'une information n'est pas systématiquement établie à propos du monde. Autrement dit, le discours reste clôt sur lui-même, et n'est pas invalidé par de quelconques informations provenant du monde. Or, justement, l'information propose une autre caractérisation du rapport entre discours et monde. La neutralité informationnelle (telle que définie par exemple par L. Floridi) permet d'enrichir le discours d'éléments inédits, y compris lorsque la demande de ces contenus est le fait même de discours.
La pertinence de la problématique de l'information se situe dans la remise en cause de la clôture du monde par des discours. Nous allons donc maintenant caractériser plus précisément comment l'information peut s'inscrire dans les dispositifs d'activité.

Les questions de référence et d'inférence.

Les propositions (entendues au sens sémantique) pourraient très bien être validées non dans des univers postulés, mais dans des univers peuplés d'objets, d'individus et d'opérations observables. Les propositions de E. Hutchins pourraient être utilisées pour valider des hypothèses sur la signification. Cette hypothèse est séduisante, surtout dans un cadre circonscrit d'une activité.
Néanmoins, son éventuelle satisfaction repose sur un modèle de l'activité, que nous nous employons maintenant à proposer. Pour cela, il convient maintenant de revenir sur la question des opérations et de l'action, telle que nous avons pu l'envisager précédemment, afin de montrer de quelle façon les flux et les situations s'inscrivent dans un processus matériel d'action, tel que la production et la circulation des unités composant l'information.

G. Hutchins caractérise les opérations par un changement (création, transformation, propagation) d'état représentationnel. Les états représentationnels sont des réalisations physiques et manipulatoires ayant une valeur symbolique.
Les opérations et les états sont culturels : ils sont transparents et caractérisent une appréhension fonctionnelle. En effet, dès lors que l'activité est considérée au travers des processus qui s'y déroulent et des états produits par ces opérations, la culture est caractérisée par une dimension fonctionnelle.
Par ailleurs, cette activité se déroule au sein d'un environnement qui constitue « l'habitat naturel » de cette activité. Le lien entre l'activité et le monde naturel est alors caractérisé par les circonstances.

Les opérations sont au fondement de la dynamique de l'activité. Pour une situation donnée, la dynamique se caractérise par une altération de cette situation due à une action (ou comportement). Celle-ci produit une nouvelle situation.

On a vu en 4.3.6une représentation simple d'opérations. Elle permet ensuite de caractériser comment des opérations complexes sont inscrites et transcrites dans des outils. (Les outils étant caractérisés par des opérations fonctionnelles, toute analyse d'outil consiste à identifier l'ensemble des changements d'états qui sont réalisés au travers des opérations qui lui sont propres). L'outil peut être représenté par des dimensions fonctionnelles. Il possède une panoplie de fonctionnalités lui permettant de réaliser une opération en prenant en compte les spécificités de la situation de départ.

Par exemple, le calcul d'une vitesse sur une embarcation requiert un sablier et une corde contenant des marques de longueur, lestée d'un objet. Dès lors que l'objet a atteint le fond, à l'aide du sablier, on observe la longueur de corde déployée à la minute. Elle donne une vision analogique de la vitesse. Cette mise en relation distance/minute constitue la base de la transcription de la représentation analogique en représentation digitale.

Il est nécessaire d'utiliser d'autres outils pour transcrire ces résultats en heures. C'est alors que l'on utilise les tables logarithmiques. En effet, si la transcription des minutes en heures est facile, il n'en va pas de même pour les mètres. Ces tables permettent de remplacer les opérations typographiquement complexes de multiplication et de division par des opérations limitées d'addition et de soustraction. C'est un intérêt de la représentation spatiale que de permettre de substituer aux régularités physiques du monde les régularités du monde symbolique des nombres.

, L'identification d'opérations se fait au niveau computationnel. Ces opérations peuvent être de deux ordres :

- soit elles concernent l'élaboration des circonstances : ce sont les opérations d'abstractions et les sources.
- Soit elles sont réalisées par l'acteur dans la région manipulatoire et sont transparentes pour quiconque observe l'activité.

Par exemple, la mesure de la planche nécessite trois objets physiques (planche, mètre décimal, crayon) dans leurs dimensions symboliques : la planche en tant que représentation particulière d'une longueur, le mètre, comme représentation prédéfinie d'une longueur particulière, et enfin le crayon par ses capacités d'usage dans une mémorisation.
La mesure constitue la création d'un état représentationnel par l'ouverture du mètre sur la longueur recherchée (1 mètre 50). Cet état est propagé sur la planche par apposition du mètre sur la longueur. Le marquage constitue une transformation de l'état représentationnel de la planche : la longueur qu'elle représente est segmentée en deux parties, dont l'une mesure 1 mètre 50.

Enjeu d'une explicitation des méthodes de calcul utilisées.
Les opérations servent de grammaire de base pour la description des phénomènes observables. Généralement, dans le cadre de l'étude de l'information, les modèles d'opération utilisés sont d'abord probabilistes, à la suite de Carnap & Bar-Hillel, puis utilisent les logiques propositionnelles et les grammaires catégorielles.
Indépendamment de l'application que l'on proposera dans le cadre du web de données, on présente ici un certain nombre d'observations et de modélisations informelles qui permettent de présenter le type de phénomènes dont les flux peuvent rendre compte. Nous nous garderons d'interpréter ces phénomènes réguliers. Nous reviendrons sur le cadre théorique dans lequel ils prennent sens dans la partie 5.3., relative aux hypothèses de la cognition située et distribuée. En effet, nous ne voulons pas d'interférence entre notre explication des phénomènes dont les flux rendent compte et la façon dont une théorie peut expliquer ces modèles.

5.3.2. Modèle de l'activité.

Comme on a pu le noter précédemment, les flux d'information reposent sur une articulation entre dispositifs symboliques et matériels. Notre hypothèse initiale est que c'est justement dans une relation précise entre référence des symboles et objet des opérations matérielles que s'établit la pertinence de l'apport des flux dans le cadre de l'étude de la cognition en situation naturelle (par opposition aux protocoles de laboratoire).
Par exemple, une « cinétique » désigne à la fois un métabolisme propre à un patient et un protocole et des outils d'analyse de ce métabolisme.
Nous voudrions donc présenter tout d'abord l'amorce d'un modèle de l'activité, puis une présentation de la façon dont il est possible de caractériser de façon plus formelle la cognition située et distribuée. Cette caractérisation nous servira à envisager comment les principes de la cognition située et distribuée peuvent être envisagés dans le cadre du web de données.

5.3.2.1.Elaboration d'un modèle de l'activité.

Il existe de nombreuses définitions de l'activité. Nous tenterons une très rapide synthèse avant de revenir aux questions qui nous préoccupent.

Nous avons déjà évoqué la notion d'activité, essentiellement pour caractériser un processus marquant soit un changement de valeurs associées à une variable (selon le modèle des logiques dynamiques), soit l'émergence d'un objet.
Un tel modèle s'inscrit dans le cadre des sciences sociales au travers des problématiques de l'anthropologie cognitive, et plus particulièrement de la caractérisation cognitive d'une activité professionnelle. La théorie de la cognition sert alors d'outil d'analyse pour caractériser la façon dont une communauté traite d'un matériau pour produire un objet fonctionnel. Par exemple, C & J. Keller (op.cit.) proposent un modèle d'émergence pour caractériser un travail artisanal : quel système cognitif les ferronniers d'art ont-ils élaboré de façon à produire régulièrement des objets en métal variés ?

On voudrait s'intéresser maintenant à la façon dont ce processus se traduit dans l'étude d'activités professionnelles socialisées, en essayant de faire le lien entre ces dimensions globales et celles qui gouvernent la production régulière d'objets ou de services. Le modèle de l'activité sert pour nous de contexte à l'interprétation des informations ; les situations constituent des cadres de l'interprétation de l'information. L'activité constitue un modèle bien plus générique. La modélisation de l'activité articule les dimensions sémantiques et les constructions matérielles (notamment les outils) de l'activité. La modélisation que l'on propose pourrait constituer le fondement d'une ontologie de l'activité.

En reprenant les travaux fondateurs de L.S. Vygotsky[215] , on s'efforcera dans un premier temps de caractériser certaines différences d'approche liées justement à la finalité de l'activité en question. Dans un second temps, on essaiera de modéliser ce processus, en montrant que ces distinctions peuvent être référées à la définition de la tâche, proposée par Marr.
L.S. Vygotsky distinguait un bas niveau, composé des opérations de vision, manipulation et de réflexes, un niveau intermédiaire, caractérisé par le champ de vision, l'environnement matériel et la tâche, et enfin un haut niveau, comprenant la mémoire de travail, la mémoire du contexte, et l'anticipation du déroulement de l'activité.

Dans ce cadre, la dynamique de l'activité se traduit par de mécanismes d'internalisation et d'externalisation. Un fait externe est internalisé par abstraction dans le cadre de l'élaboration d'un plan, qui ensuite sera externalisé par sa traduction dans des manipulations. On aboutit ainsi à un système fonctionnel hiérarchique.
Dans un tel cadre, l'outil constitue l'étape ultime de la traduction d'une pensée dans un objet externe.

Cognition sociale et activité. Cadre de travail.
Les différents ensembles de recherches qui se réclament de l'anthropologie cognitive ont des objets de grandeur différente : cela peut aller d'une compétence collective, caractéristique d'une profession, comme les ferronniers d'art (C. & J. Keller, op. cit.) à des segments d'activité précis, comme les procédures de descente pour un atterrissage chez G. Hutchins (op. cit.).

Ces choix dépendent de la théorie de la cognition présupposée. En effet, il faut distinguer une théorie de l'activité reposant sur un partage des connaissances et la négociation des compétences et des domaines d'expertise, de la cognition située et distribuée, qui s'attache à caractériser les unités et processus cognitifs dans certaines relations entre les capacités opératives du monde et l'esprit. Les deux perspectives peuvent se rejoindre, comme dans les travaux de C. & J. Keller. La problématique de l'activité s'inscrit plus particulièrement dans

le cadre de la cognition sociale mais limite considérablement l'objet de cette étude à des phénomènes localisés structurés autour d'objets et de processus matériels.

Par ailleurs, le problème de l'activité peut être caractérisé soit dans le cadre de règles sociales régissant les relations entre différents acteurs impliqués dans la construction de la réalité du travail, ou au contraire, dans le cadre du traitement d'objets du monde par un système cognitif (largement externalisé dans la plupart des cas). Nous n'avons pas l'opportunité de développer cette dimension sociale de l'activité. Cette remarque permet de marquer une limite de notre travail. En n'intégrant pas la dimension sociologique, on ne prend en charge ni les déterminants et contraintes collectifs, ni l'autonomie de l'acteur dans le cadre social[216]. Cette limitation correspond également à la caractérisation de l'usage que l'on retient, qui ne prend pas en compte justement la dimension sociale.

Définition de l'activité.

L'activité constitue une segmentation comprise entre d'une part les pratiques, qui constituent des ensembles de connaissances et d'actions ayant une certaine fonction dans une société, et d'autre part des échanges structurés, à savoir des unités de coordination et de coopération situées à l'intérieur d'une activité, et qui visent à instaurer une communauté de connaissances et de définition d'états de faits à l'intérieur de la menée d'une activité, de façon à pouvoir atteindre collectivement l'objectif.

Considérée ainsi, l'activité constitue un niveau intermédiaire entre des finalités partagées et des accommodements locaux et temporaires, en fonction de circonstances systématiquement différentes. L'apprentissage consiste justement à intégrer les solutions trouvées par rapport à un certain type de circonstance à l'intérieur même de l'activité.

Considérée à un niveau intermédiaire, entre des rôles et des finalités sociales considérées comme un cadre de justification et des séquences d'action dirigées par des circonstances, l'activité consiste à caractériser à la fois le processus générique de production d'un objet et les différentes façons de le produire, en fonction des circonstances. L'activité caractérise une panoplie de stratégies et de méthodes pour réaliser l'activité prévue (C. & J. Keller).
Cette panoplie est caractérisée à la fois par des dimensions internes et externes de la cognition :

- les outils, les réalisations d'action, les transformations de l'objet et la configuration de l'espace constituent ces dimensions externes. L'usage des outils s'inscrit dans l'extension de leurs fonctionnalités. Il résulte de cette extension une pluralité d'usages des outils, pluralité inscrite dans la panoplie.
- Les choix, les stratégies et les représentations prévisionnelles de l'objet constituent les dimensions internes de la cognition. Il s'agit du contrôle interne de la menée de l'activité, intégrant à la fois les contraintes extérieures et l'autonomie de l'acteur.

Dans le cadre de l'activité, la tâche est ce que fait le système et ce pourquoi il le fait. Il s'agit de la compréhension et de la maîtrise du dispositif global. Celui-ci explique l'activité dans son fonctionnement et les motivations de l'agencement. Il s'agit ici d'une compétence humaine.

L'analyse de l'activité permet alors de faire converger des processus instrumentaux et des processus abstraits au travers de l'observation de transformations d'états. Que l'on soit dans le

cadre psychologique ou anthropologique, les analyses de l'activité reposent sur des unités observables et discrètes que sont les objets, les opérations et les changements d'état. Par conséquent, on dispose d'une communauté d'objet d'investigation scientifique. Elle inclut également de nombreux travaux en ergonomie cognitive. Elle se distingue alors des perspectives sociologiques, qui elles ne requièrent pas l'ancrage dans ces différents types d'objets construits.

Néanmoins, l'analyse de l'activité, telle qu'elle vient d'être décrite, présuppose une unité de lieu qui certes correspond parfaitement à l'atelier du ferronnier d'art, mais beaucoup moins à des activités professionnelles où les outils sont conçus indépendamment d'une tâche ou d'un acteur, ou d'activités reposant essentiellement sur des marqueurs symboliques. Alors, l'importance des dimensions linguistiques et de l'interprétation ne pourra être ignorée. Nos propositions s'inscrivent dans ce cadre.
En effet, il est difficile, considérant des pharmaciens hospitaliers dans la réalisation d'une adaptation de posologie, de décrire un dispositif d'outils, d'espace de travail et de repères suffisamment élaboré pour amener des propositions intéressantes. Par contre, le lien entre ces outils et les objets traités (le malade dans son lit, les molécules et les bactéries), comme les relations entre ces objets réels et les populations statistiques du logiciel utilisé est apparu comme une question centrale pour notre propos.
Cette dimension de l'activité nous est affectivement apparue comme la plus intéressante et jusqu'à aujourd'hui la moins explorée. Elle mobilisait des outils de description et d'analyse spécifiques à la dimension symbolique et plus particulièrement l'interprétation sémantique.
Ce constat nous a amené à reformuler très largement la question de l'activité et à proposer un modèle reposant sur le traitement des objets symboliques.

Le phénomène qui nous a semblé alors le plus intéressant était le caractère à la fois distribué de l'activité et des raisonnements, mais aussi et surtout l'importance de l'information et le fait que le praticien se caractérisait d'abord comme un interprète d'une information.
Ces données empiriques nous ont amené à réorienter très considérablement notre travail vers justement ces questions d'activité fondées sur une structure symbolique.
On a donc élaboré un modèle qui représente des contraintes à l'interprétation de faits du monde à l'aide d'un dispositif fondé sur la distribution et l'externalisation des connaissances.

Activité et information.
L'analyse d'une activité vise, quel que soit le cadre dans lequel on se situe, à élaborer le modèle cognitif du processus permettant la réalisation d'un certain objet. Dans ce cadre, analyser l'information consiste à étudier comment des informations (étendues au sens de représentations de faits inédits) peuvent être interprétées dans un cadre contraint par des finalités. Les finalités du service de pharmacie sont doubles : elles concernent d'une part le service fournit aux prescripteurs, et d'autre part l'acquisition de connaissances relatives à ce type de soin. L'acquisition pour le service réside d'abord dans l'acquisition d'un savoir supplémentaire, ou d'un renforcement de positions thérapeutiques.

Par ailleurs, l'activité construit un système matériel d'information[219], au sens où les canaux sont élaborés de façon à permettre l'acheminement de l'information, son interprétation et la diffusion du résultat de cette interprétation dans un dispositif d'analyse externalisé (le logiciel probabiliste) et enfin dans un autre système de communication. Mais parallèlement, le contenu informationnel transmis n'appartient pas à l'activité en question : il s'agit de données

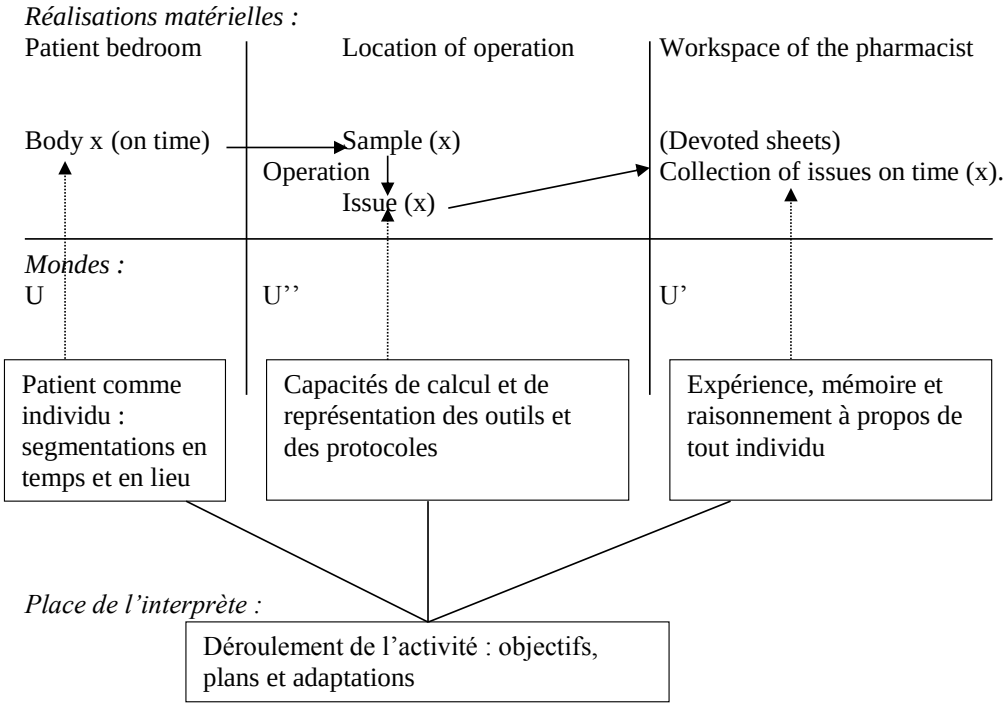
[219] On entend système dans sa définition générique (un ensemble d'outils solidaires conçus pour une répondre à une fonction de façon régulière et optimale.

relatives au patient, à la bactérie et à la molécule, dont les comportements sont autonomes par rapport à l'activité. Ainsi on explique que l'activité s'adapte aux circonstances transmises.
En ce sens, l'étude de l'information, donc des circonstances, ne peut être confondue avec la modélisation de l'activité. L'analyse de l'interprétation de l'information permet de comprendre de quelle façon les circonstances sont représentées et insérées dans l'analyse de l'activité.

Dans les analyses d'activités, celles qui font le plus intervenir les systèmes d'information sont celles dont la problématique est la convergence entre les acteurs par la coopération[217]. Les contraintes « externes » ont assez peu été étudiées, entre autre parce que justement, étant externes, elles sortaient du cadre de l'analyse. Par contre, en se focalisant sur le fait que ces contraintes étaient essentiellement informationnelles, donc manifestées sous forme symbolique, il devenait possible de proposer un modèle.

Caractérisation des mondes d'interprétation dans le cadre de l'activité.
Nous avons précédemment émis l'hypothèse de trois mondes distincts pour caractériser le cadre dans lequel l'information prend sens.
Le monde U est purement extensionnel : il est composé d'objets du monde.
Le monde U' est intensionnel au sens où il est composé de représentations du monde U. Ces représentations ne sont pas seulement mentales ; elles sont également matérielles, mais sont toujours relatives au monde U. (On peut, si l'on suit J. Barwise & J. Allwein (op. cit.), considérer les représentations analogiques comme des structures intensionnelles).
Le monde U'' est la projection du produit de l'intension sur les univers extensionnels. Il s'agit d'un produit pragmatique au sens où il s'agit d'élaborations intellectuelles projetées dans des outils matériels (comme les outils d'analyse et de mesure).
La dimension matérielle et le caractère imprédictible des phénomènes de ces mondes font qu'il est impossible de les caractériser comme des domaines. Si le monde intensionnel pourrait constituer un domaine de connaissances, il n'en reste pas moins que dans un cadre d'analyse, distinct de la modélisation formelle de ce domaine, une telle appellation n'est pas nécessaire. Par conséquent, les mondes désignent des régions de référence et constituent des concepts intermédiaires entre la description et la modélisation formelle.

Caractérisation informelle des trois mondes.
Nous reprenons maintenant la schématisation de l'activité proposée précédemment, en 4.3.6., pour ne caractériser maintenant que les relations entre les mondes et les parties du monde matériel concernées par l'activité.



Schématisation des trois mondes de l'interprétation.

: représente la référence.
: représente une opération.
: représente la relation opérée entre les trois univers.

Figure 10. Schématisation des trois mondes dans le cadre de l'activité.

Usage d'une définition située et distribuée de la cognition pour caractériser l'interprétation à l'intérieur d'une activité.

Au préalable, l'interprétation ne concerne pas seulement les unités du langage mais leur contexte. Autrement dit, l'interprétation constitue une activité inscrite dans le cours d'action. Ainsi, les connaissances acquises et les informations précédemment traitées sont intégrées dans le déroulement de l'activité.

Une telle modélisation s'intègre en partie à l'intérieur d'activités distinctes de la pharmacie hospitalière, comme par exemple la pratique scientifique. Ainsi, pour l'étude d'un fait particulier du monde, on construit des données scientifiques observables et des outils pour les appréhender. Ces deux univers requièrent un troisième, qu'ils enrichissent également, et qui est un cadre théorique et méthodologique.

Les connaissances relatives à la construction des données, à la segmentation des phénomènes du monde, sont distinctes de celles qui caractérisent les développements théoriques tout autant que celles qui définissent les connaissances relatives au monde (plus précisément les résultats d'une étude). Cette modélisation pourrait être particulièrement pertinente pour caractériser et

structurer les dispositifs d'e-science. En effet, la dualité données primaires et publication est un peu sommaire, relativement à la complexité des objets et des pratiques.

Cette approche permet ainsi de spécifier des domaines de connaissances distincts, liés chacun à des mondes constitutifs d'une activité, et qui permettent d'expliquer la découverte, l'établissement de faits, et l'explication. Autrement dit, l'action.

5.3.2.2.Hypothèse de la cognition étendue et modélisation.

Le développement qui suit concerne le cadre le cadre théorique sur lequel se fonde la théorie de la bi-dimensionnalité. Ce cadre théorique permet également de faire le lien entre les conclusions des descriptions et modélisations d'anthropologie cognitives et les propositions de sémantique formelle.
L'émergence de ce cadre théorique peut être envisagée en deux parties : la première concerne la question des états mentaux et la contestation de l'individualisme méthodologique (que nous avons déjà partiellement abordée précédemment), la seconde concerne en propre la théorie de l'esprit étendue qui est en quelque sorte un produit des remises en question précédentes. Nous quittons le cadre de l'anthropologie cognitive pour celui d'un autre constituant des sciences cognitives, la philosophie de l'esprit.

Publicité et communauté des états mentaux.
La première contestation de l'individualisme méthodologique provient de la question des états mentaux. Si ces états sont individuels et internes, comment sont-ils communiqués, quelles conditions doivent être réunies pour que l'état soit commun.
On distingue des états communs à chacun ou « publics » (« j'ai soif ») et des états personnels communiqués dans un but : « je crois que Pierre va venir ». Or, ces états n'existent que par leur expression dans le langage, ce qui constitue un paradoxe.
Par ailleurs, si « je » ou « Pierre » constituent des états mentaux ayant une existence empirique, par contre les valeurs associées (comme « pense » ou « arriver » ne le sont plus. Or, pour fonder ces états mentaux, une perspective causaliste est développée : un état mental est toujours associé à une cause qui le produit.

Si les états mentaux ne sont plus individuels mais publics du fait de la langue[218], et si la pensée devient un mécanisme reliant l'individu à un environnement, par conséquent ce qui forme la collectivité d'une expérience est la communauté des contextes et celle du langage.
En ce sens aussi, une telle définition des fondements de la cognition accorde une place essentielle d'une part à la culture (collectivité des expériences du monde) et d'autre part au langage comme forme permettant de transmettre les expériences individuelles. Or, rien n'indique que seules les unités du langage naturel soient susceptibles d'assurer cette transmission ; d'autres formes, comme les outils et les supports de médiation, pourront effectivement remplir une telle fonction.

A partir du moment où on intègre des objets dans les facultés cognitives, se pose le problème immédiat de la caractérisation du sujet, et par conséquent de l'incorporation de la cognition. On verra que a socialisation des états mentaux s'inscrit à l'intérieur des propositions de D. Chalmers (op. cit.). Le fondement de son argumentation consistera à caractériser distinctement les états mentaux issus de la perception individuelle des conventions sociales.

Une conséquence de cette proposition est la caractérisation des états mentaux en type et occurrents, mais également leur dimension propositionnelle. Si effectivement la

caractérisation non-individuelle des états mentaux est avérée, néanmoins, ces états peuvent rester des expériences individuelles, et l'argument alors est physiologique. Ce sont bien des expériences individuelles vécues dans le cerveau.
Or, justement, ce qu'un psychologue comme T. Burge[219] essaiera de faire, c'est de contrer cette idée, en mettant en avant la dimension relationnelle de l'expérience. L'idée serait que les états mentaux seraient toujours des états internes de l'individu, mais ne seraient pas déterminés par les limites physiques de celui-ci.

Environnement et pensée individuelle.
Le paradoxe posé par T. Burge notamment, consiste à montrer que toute pensée individuelle serait dépendante de l'expérience propre du monde[220] (p. 304). L'individualisation d'une pensée dépend d'une relation à un certain environnement particulier.

Cette position permet de montrer que l'individualisation de la pensée dépend de la construction de l'environnement. Or, cet environnement est multiple ; doit-on considérer les seuls paramètres de la localisation, de l'interaction ou au contraire intégrer les environnements et les outils ?
Méthodologiquement, le problème s'est centré sur d'une part les outils et d'autre part l'interface entre le monde et le sujet. Sont alors convoquées les méthodes de constructions de ces représentations et en premier lieu la vision. Ainsi, la psychologie individuelle ne peut seulement, au mieux, parler que d'une partie de la cognition, du fait de la non-limitation des processus cognitifs à des aspects mentaux.

Dès lors, ce qui apparaît central dans l'étude de la cognition, c'est le lien entre le monde et les états mentaux. En quelque sorte, il s'agit d'une question d'implémentation.

Par ailleurs, à partir du moment où l'on considère des états mentaux individuels, le processus de causalité n'est plus focalisé sur les effets (qu'est-ce que tel état mental type entraîne comme autre état), mais sur les causalités antécédentes. Cette proposition est intéressante par rapport à l'interprétation dans le cadre d'une activité : l'interprétation de la signification d'un résultat d'analyse requiert l'interprétation des opérations (donc de l'activité) qui a permis d'obtenir ce résultat. Le résultat à propos d'un patient est interprété dans l'ensemble de la chaîne causale antérieure, postulant en fait l'ensemble des opérations ayant produit ce résultat.
.
La causalité envisagée par T. Burge diffère de celle présentée par J. Fodor, entre autre par la distinction entre causalité physique (le mouvement causé par l'activation d'une certaine zone du cerveau) et causalité psychologique.
Or, cette causalité est déterminée par ce qui est considéré comme type, et donc dépend de la façon dont l'expérience individuelle construit l'expérience. Autrement dit, il s'agit d'une causalité de reconstitution explicative, et non de conséquence de types abstraits antérieurs ; les types ne sont pas des causalités initiatrices d'un comportement ou d'un état mental, mais sont marqués par une antécédence sur une situation occurrence avec laquelle il serait en relation d'adéquation. Cette proposition, qui relève en apparence de l'analyse des états mentaux, a quelques conséquences sur la façon dont les sciences cognitives traitent de ces questions.
En tout état de cause, les modèles formés par l'expérience scientifique physiologique ne sont pas considérés comme pertinents, d'un point de vue psychologique, dans la caractérisation de la causalité des états mentaux. Les causalités sont normales dans un environnement particulier. (P. 317) Elles ne sont pas caractérisées par une quelconque ontologie mais bien par la formation de patrons d'expérience. Les régularités de cette expérience peuvent être

celles de l'activité et de ses opérations réalisées à l'aide d'outils : les patrons d'expériences deviennent alors particulièrement réguliers.
(E. Rosh[221] a exploité cette idée, notamment pour la théorie des prototypes, en insistant sur la dimension collective des expériences).

En définitive, en considérant la relation entre l'individu et son environnement comme primordiale, on construit une caractérisation des états et des liens entre ces états fondée sur la spécificité de chaque environnement. La contestation de la dimension individuelle de la cognition entraîne la remise en cause d'un postulat fondamental du fonctionnalisme : la primauté de l'intention (et donc de mécanismes purement mentaux et individuels dans la formation des états mentaux). Néanmoins, T. Burge ne remet pas en cause la causalité, mais seulement sont fonctionnement : les conditions et la situation sont alors considérées comme des causalités physiques explicatives.
Cette construction théorique pourra être fondamentale pour positionner le modèle de l'activité par rapport à l'interprétation des structures d'information. L'idée selon laquelle certains états mentaux ne sont possibles qu'en fonction de certaines situations et relativement à des événements antérieurs et réguliers constitue un argument fondamental pour Articuler le monde naturel (comme celui du patient) et celui de l'activité et de ses opérations régulières.

Incorporation.
A partir du moment où l'on postule que l'activité cognitive n'est pas réduite au cerveau, se pose la question du rôle du corps et de la matérialité, entre autre si l'on distingue l'esprit et la cognition. La nature des traitements des informations en provenance du monde est réévaluée par le fait que les processus corporels peuvent être caractérisés comme de premiers traitements cognitifs.

Par ailleurs, l'incorporation joue un rôle essentiel dès lors que l'on considère la structuration de l'espace matériel en fonction des accès physiques aux objets. Cette caractérisation de l'espace de manipulation (espace accessible directement au sujet par ses mouvements), amène l'idée que la cognition externe se fonde sur l'accessibilité des objets dans le cadre des activités quotidiennes. Cette perspective permet de caractériser le lien entre les dimensions internes et externes de la cognition par le geste.
Comme le montre le travail de C & J. Keller, si les outils inscrits dans un espace approprié s'intègrent dans la caractérisation de l'activité cognitive, c'est pour une modélisation intégrant connaissance et pratique dans le cadre de la fabrication d'objets.
On peut également citer dans ce cadre les travaux de J. Lave et surtout de L. Suchman[222].

Même si ces questions sont quelque peu éloignées de nos problèmes d'activité, la caractérisation du traitement de l'information comme seules inférences sur des symboles est remise en cause. C'est donc l'architecture générale du « langage de la pensée » qui est contestée. Plus généralement, ce sont les questions de l'abstraction des processus cognitifs qui sont posées ; l'incorporation requiert un faible niveau d'abstraction des objets qui par ailleurs sont ceux qui sont directement perçus et linguistiquement désignés par des indexicaux (ou déictiques). Ainsi, une autre implication de l'incorporation de la cognition est la caractérisation d'objets et de raisonnement de faible niveau d'abstraction.

Ce lien, formulé notamment parD. Ballard & alii.[223], permet de lier les questions de corps dans la cognition avec celles des langues et d'information, mais également celles d'activité et d'usage d'artifices de médiation. L'expérimentation menée vise à saisir de quelle façon l'orientation du corps et des mouvements corporels structurent le raisonnement. L'hypothèse

centrale est qu'un changement de position par rapport à un environnement modifie la façon dont on appréhende cet environnement et donc la façon dont ses éléments sont utilisés dans un raisonnement. Autrement dit, la façon dont on internalise ou on externalise un élément de l'environnement dépend du développement progressif d'un raisonnement. En résumé, l'expérimentation caractérise comment on utilise des déictiques dans l'émergence d'un raisonnement, à savoir la façon dont les déictiques accompagnent la prise de distance par rapport à un signal et le développement d'un raisonnement par rapport à ce signal.

Externalité, indexicalité et représentations.
La première difficulté qui caractérise l'entreprise de la cognition située et distribuée, est la caractérisation des objets externes exerçant une partie de l'activité de raisonnement. Par ailleurs, l'indexicalité constitue un argument contre les propositions d'une cognition non symbolique : les déictiques seraient les outils permettant de connecter une pensée symbolique au monde concret. Or, justement, la cognition située et distribuée va considérer l'indexicalité comme un outil permettant de construire des expériences.

L'idée proposée par A. Clark & J. Chalmers et développée par J. Sutton[224] (p. 506) consiste à considérer que les outils, les technologies et les dispositifs constituent des relais de l'esprit individuel, et donc sont en continuité d'un individu. Il s'agit d'objets définis par une capacité fonctionnelle (qui peut être de traitement de l'information) et une propriété représentationnelle.

La fonctionnalité des objets et leurs propriétés représentationnelles constituent les deux arguments fondamentaux d'une appartenance des propositions de cognition située et distribuée aux sciences cognitives. (Rappelons qu'un postulat fondamental des sciences cognitives réside dans leur réfutation du béhaviorisme et sa restriction des objets d'analyse aux observables et donc le refus d'une modélisation des contenus internes de la pensée).
Ainsi, la définition d'un niveau représentationnel de la pensée constitue une condition pour une théorie cognitive complète, comme le rappelle Z. Pylyshyn[225], même s'il ne s'agit ni du seul, ni du principal niveau de la théorie.

La plus grande difficulté, et ce qui va distinguer différentes théories de la cognition, c'est la façon dont les contenus représentationnels sont liés au monde. L'hypothèse de la cognition située et distribuée consiste à considérer que ces représentations sont externalisées, marquées au sein de supports de médiation. Les raisonnements de l'acteur s'appuient alors sur ces représentations externes.

Le problème initial est celui de l'insuffisance des propriétés descriptives des objets pour caractériser la façon dont une pensée est liée au monde, et donc la nécessité d'une dimension indexicale afin de lier la pensée aux objets et donc rendre possible l'action. De cette insuffisance, qui est à l'origine de la philosophie du langage (dès B. Russell et le problème de l'expression « le roi de France est chauve »), on verra se développer à la fois la problématique de la référence directe (à la suite notamment de S. Kripke[226] et D. Kaplan[227]), mais également la théorie des situations (J. Barwise & J. Perry, op.cit.), et enfin les propositions de cognition située et distribuée. Si les indexicaux sont au départ considérés comme des unités permettant d'ancrer un concept dans un contexte, et donc de lui attribuer une valeur de vérité, la caractérisation de ce contexte devient elle-même une problématique importante : comment peut-il être caractérisé, comment par ailleurs peut-il être abstrait ? Ces questions amènent progressivement à reconsidérer la construction mentale qui contient, organise et produit les

concepts. C'est ainsi que progressivement la nature purement interne des concepts pourra être remise en question.

Cette façon de considérer la conceptualisation, et par conséquent les représentations symboliques et leurs enjeux dans la cognition, a quelques conséquences sur l'appréhension à la fois du langage et des représentations. En effet, les questions d'opacité référentielle (définies notamment par W. Quine[228]) sont notablement reformulées par l'explicitation, voire la modélisation, du contexte de réalisation du langage. La dimension interprétative permet cette articulation en considérant le contexte comme les causalités externes de la réalisation des structures informationnelles. Si par ailleurs, on considère ces causalités comme des représentations externalisées de connaissances, elles peuvent être considérées comme des objets (donc des entités discrètes). Elles peuvent donc être intégrées dans une modélisation.

Caractérisation des états mentaux dans le cadre de la cognition située et distribuée. Matérialisme et phénoménologie.

Si effectivement, il est relativement facile de montrer que certains raisonnements s'appuient sur des données externes, par contre, il est plus difficile de considérer des états mentaux qui seraient fondés sur une dimension externe. Or ces états mentaux sont essentiels pour aborder la dimension symbolique de la cognition (qui sans être la seule, en constitue une dimension essentielle).

La cognition située et distribuée ne fournirait un cadre philosophique crédible que si elle donnerait une définition cohérente de ces états mentaux. Autrement dit, l'externalité ne consisterait pas seulement en quelques supports environnementaux pour des calculs.

On distingue les mécanismes de traitement de l'information, qui relèvent de la cognition, des états mentaux (dont font partie les expériences et les croyances), qui eux, relèvent de l'esprit (R. Wilson & A. Clark, *How to situate cognition: Letting nature take its course*, op.cit.). L'activité cognitive étant étendue à des objets et des outils externes, elle ne peut être seulement considérée comme une activité interne du sujet, à la différence de l'expérience et des croyances. Si l'activité cognitive est distribuée et située, l'esprit se caractérise par la conscience et l'intention à propos de cette activité cognitive. En considérant que l'intention est une partie de la conscience et que les états mentaux sont directement à propos de cette activité cognitive « étendue », on en conclut le fait que les états mentaux sont autant socialisés que l'activité cognitive.

Cela dit, la contestation de l'individualité des états mentaux ne constitue pas un rejet du concept, mais la nécessité d'une reformulation. Ceux-ci possèdent à la fois une dimension phénoménale (marquée par l'association de propriétés aux objets, et donc par l'expérience) et une dimension intentionnelle, associée à la capacité à représenter des buts, des plans et des objectifs)[229]. La dimension phénoménale, composée de concepts et de propriétés phénoménales alimentés par les capteurs sensori-moteurs notamment, construit l'expérience et permet donc l'introspection.

La cognition située et distribuée pose de façon nouvelle la question de l'individu ; en effet, au-delà de l'incorporation, la cognition située et distribuée, et son extension philosophique qui est « l'esprit étendu » procède à une redéfinition complète du fonctionnement de l'esprit humain parce qu'elle postule le primat de la dimension socialisée des phénomènes cognitifs.

Par ailleurs, toutes ces compétences externalisées font des objets et des outils du monde des représentations dans le cadre des raisonnements. Cet usage du monde n'annule pas le fait que par ailleurs, évidemment, le monde existe.

Par ailleurs, la remise en cause des états mentaux, et surtout de leur dimension individuelle, a des conséquences importantes sur la définition d'une proposition, et plus généralement sur la sémantique associée aux états mentaux et leur typage. Les conséquences de cette remise en question seront essentielles pour caractériser ensuite l'approche appropriée du langage et plus particulièrement de la signification[230]. Considérant l'externalisation et l'incorporation de la cognition, le langage signifie dans cette dualité : à la fois dans le cadre externalisé de la cognition et internalisé de l'esprit.

En définitive, dans les théories de la cognition située et distribuée, l'esprit possède une dimension ontologique. C'est ce que la philosophie essaie de traiter au travers de la problématique des zombies et de l'impossibilité de les concevoir[231].

Une caractérisation de la cognition fondée sur le matérialisme.
Comme permet de l'indiquer le dernier fondement, c'est sur les questions de la personne et de son rapport au monde via le langage, que se développe la proposition théorique. Elle trouve son origine dans les travaux de J. Perry, eux-mêmes à l'origine de la théorie des situations. Ils évolueront vers les propositions de D. Chalmers.

Le matérialisme serait l'idée selon laquelle il ne pourrait y avoir de conscience sans un rapport à des objets du monde, et que c'est ce rapport qui construit la conscience. Or cette conscience matérialiste repose sur soit sur un rapport direct aux objets, soit sur des phénomènes, qui constituent des constructions mêmes de ce monde.

L'indexicalité est ici un concept crucial puisqu'il permet de caractériser matériellement la conscience (celle d'être ici, dans un certain temps, environné de tel ou tel objet). Mais parallèlement, cette conscience est dépendante des objets et des outils de cet environnement.

L'indexicalité constitue un point essentiel aussi pour aborder la question des connaissances dans la mesure où elle permet de caractériser comment les connaissances s'intègrent dans le rapport au monde. Ainsi, en vertu d'une telle définition de la conscience, les connaissances seront d'abord caractérisées par une relation constante entre le sujet et l'objet perçu. Il s'agit ici d'une connaissance phénoménologique. Ainsi, une expression a un sens parce qu'elle a d'abord une extension[232]. Mais cette extension n'implique pas que l'expression ait un seul sens : toute extension peut avoir des descriptions différentes, entre autre parce que les représentations associées à ces extensions (dans le cadre de l'activité cognitive) sont différentes. A la base le sens est épistémique, à savoir qu'il caractérise une certaine croyance sur ce qu'est l'objet, il est immédiatement adjoint sur cette croyance un certain nombre de scénarios différents, liés à des usages et donc des intentions différents à propos de ces objets.

L'externalité et l'internalisation. La primauté de l'expérience.
On identifie maintenant quelques lignes directrices de la pensée de D. Chalmers qui fondent et positionnement la théorie de la cognition située et distribuée dans le cadre de la philosophie analytique. Comme on peut s'en douter, il s'agit plus particulièrement d'insister à la fois sur l'expérience individuelle et sur la construction sociale de représentations.

Par l'indexicalité, on décrit les passages d'entités depuis le monde externe vers celui des représentations internes. A la suite de Ballard & alii. (op. cit.), l'indexicalité est suivie par un

processus de mémorisation qui permettent à l'acteur dans le cadre d'une tâche, de ne plus avoir à se référer aux instructions formulées dans l'environnement.

On reprend ici le réalisme phénoménologique de D. Chalmers parce qu'il permet de caractériser le contenu de l'état mental comme étant une forme d'expérience (et non un type, comme dans la conception de J. Perry). Un état mental est une mémoire à la fois d'un certain type de phénomène et de la façon dont on peut réagir par rapport à lui. Cette conception est liée à une double définition de l'intension, comme étant d'une part épistémique et d'autre part « subjonctive ». La part épistémique caractérise le fait que l'on puisse associer une pensée ou un concept à un espace de réalités plausibles. La part subjonctive ou contrefactuelle renvoie au fait que cette pensée ou ce concept renvoie à un espace de réalités contraires ou contrefactuelles.
Nous reprenons l'exemple de D. Chalmers[233]. Pour le concept "eau", l'intension épistémique choisit H_2O dans notre monde et XYZ dans une terre jumelle. Cela reflète le fait que si j'accepte le monde actuel pour moi est comme l'univers de la terre jumelle, (le liquide des océans est et a toujours été XYZ), je dois accepter que XYZ est un liquide buvable. Par contraste, l'intension subjonctive du concept « eau »choisit H_2O à la fois sur terre et sur la terre jumelle. Cela reflète le fait que l'eau étant H_2O dans le monde actuel, la terre jumelle contrefactuelle est bien décrite comme un univers dans lequel l'eau est bien H_2O, et dans lequel XYZ est simplement une chose liquide. Dans l'intension épistémique, « eau » choisit une substance avec certaines caractéristiques superficielles dans n'importe quel monde possible, et dans une intension subjonctive « eau » choisit H_20 dans tous les univers.
Ce qui est jeu ici, c'est la définition même du concept : il n'est pas seulement constitué par son contenu actuel, mais par l'ensemble des contrefactuels de ce concept, ce qui advient quand on pose ce concept.

Néanmoins, on ne peut vraiment comprendre la portée de la proposition précédente que si l'on a caractérisé ce qu'est un concept. D. Chalmers[234] propose une distinction épistémique entre phénomènes physiques et concepts phénoménaux. Les concepts phénoménaux ne sont pas systématiquement déduits des phénomènes physiques. Les concepts phénoménaux sont déduits de l'expérience des phénomènes physiques, et y adjoignent les dimensions de l'incorporation et de la mémorisation. Ainsi, un concept phénoménal est dual et associe la référence à l'objet de l'expérience et l'expérience elle-même. Le concept phénoménal ne peut être selon D. Chalmers réduit à un démonstratif (au sens de J. Perry) mais contient des qualités descriptives indépendantes. Cette dualité individuelle est corrélée à la caractérisation communautaire du concept, sa dimension publique et collective. Cette dernière distinction permet d'associer la connaissance phénoménale à une connaissance communément partagée, donc pouvant être externalisée. Elle se caractérise par une connaissance phénoménale qualitative, pouvant être dissociée de la connaissance phénoménale indexicale. Ainsi, on peut caractériser la connaissance phénoménale de telle ou telle propriété (comme par exemple de la couleur « rouge »).

Cette primauté de l'expérience articulée à la publicité des états mentaux comme contrainte de sociabilité amène D. Chalmers (*imagination, individuality and intension*, op. cit.) à redéfinir les questions d'intension. Il distingue alors l'intension primaire (fonction depuis des mondes centrés autour d'individus dans le temps vers des valeurs de vérité), de l'intension secondaire (fonction depuis des mondes non centrés vers des valeurs de vérité). On peut dire que S est concevable de façon primaire (ou de façon épistémique) lorsqu'il est concevable que c'est actuellement le cas. Lorsque S est secondairement concevable (ou subjonctivement – contrefactuellement concevable), c'est qu'il peut avoir été le cas.

Cette caractérisation permet de rendre compte de l'apprentissage et de la révision.

Ainsi, on postule une distinction épistémique entre les différentes formes de connaissances, en distinguant celle qui relève d'une cognition centrée sur l'individu de celle qui serait commune. Les deux formes sont couplées, ce qui permet d'articuler les connaissances communes à celles qui sont centrées sur l'individu. Cette articulation peut trouver une application sur les textes, qui constitueraient une médiation entre différentes formes de connaissances. Cela permet de rendre compte de ce qui est plausible comme étant une forme de connaissance parce que c'est concevable.

Ce développement permet de rendre compte de façon relativement élaborée de la façon dont on peut caractériser le monde de l'interprétation et des connaissances par rapport au monde externalisé, celui des outils et des représentations. Il permet de considérer l'interprétation non d'un seul point de vue référentiel, mais dans le cadre d'une connaissance construite à partir de fondements matériels. L'intérêt du travail de D. Chalmers consiste à donner une consistance au travail de l'esprit dans le cadre d'une cognition située et distribuée.

La première traduction de cette matérialité est la caractérisation des supports des représentations symboliques, donc, en lien avec la perception et le maniement des objets contenant des représentations. C'est ainsi que les supports de représentation, qui constituent des médiations entre les productions symboliques et les activités d'interprétation constituent des objets d'étude centraux.

Le principe des paires.
Les propositions de D. Chalmers possèdent une dimension plus immédiatement utilisable et qui est déduite des élaborations précédentes. Le principe des paires consiste à caractériser chaque système cognitif comme un couplage entre l'égocentrisme humain et une entité externe. Le système est défini par ce couplage, où autant le composant interne que le composant externe jouent un rôle causal actif. Ils co-gouvernent le processus cognitif[220].

Le couplage constitue un phénomène systématique, et qui ne renvoie pas, comme dans le cas de T. Burge et H. Putman, à une causalité historique, à savoir la différence des expériences individuelles. Dans l'hypothèse de D. Chalmers, les distinctions de comportement sont liées aux particularités de la relation entre l'individu et le monde externe au moment où l'action est réalisée. L'externalité a alors un impact direct sur la cognition interne, sans qu'il y ait à rechercher une causalité historique.

L'exemple pris par D. Chalmers (*The Extended Mind*, op. cit., p. 9) est celui de l'alignement des lettres sur un porte-lettres au jeu de scrabble : tout arrangement et modification d'ordre apparaît non comme une partie d'action, mais de la pensée. En effet, un état d'agencement caractérise un raisonnement en cours, à savoir une mémoire d'un mot possible, ou d'un agencement qui permettrait de trouver un mot. A ce moment-là, il joue un rôle actif dans la recherche d'un mot, ou dans la mémorisation du coup à jouer.

L'élargissement de la cognition repose sur la définition des objets membres de paires en relation aux capacités du cerveau. Il s'agit ici essentiellement d'outils de calculs, de médias et de mémoires. Nous avons utilisé ce principe des paires, sans l'avoir justifié théoriquement, lorsque nous avons présenté les opérations en 4.3.6.

[220] Les paires ne doivent pas être confondues avec le principe informationnel de duplication, tel que formulé par F. Dretske.

Propriétés actives des membres des paires.
On peut objecter qu'un tel système ne permet pas de caractériser des ressources, ou plus exactement le réemploi de certaines opérations, voire même l'apprentissage. En effet, un externalisme comme celui que l'on présente semble associer la cognition à un modèle de reconnaissance. Or, les entités externes ne sont pas passives. Au contraire même, elles se caractérisent par une évolution, étant comme le patient ou la molécule dans notre cas, des êtres vivants.
Les paires peuvent être envisagées comme des modèles de l'interprétation des unités lexicales. Elles permettent de rendre compte de la sous-détermination des unités linguistiques au regard de l'évolution des objets du monde auxquels elles réfèrent. La solidarité entre le référent et l'entité linguistique rend possible l'adaptation des contenus de celle-ci aux évolutions de l'objet. La notion de caractère reflète les règles gouvernant la façon dont le contenu d'une expression peut varier d'un contexte à n autre. Le contenu caractérise le profil de cet objet, et plus précisément les propriétés qui lui sont affectées de façon constante.

Par ailleurs, comme on l'a vu, l'intension est structurée de façon duale, entre les contenus épistémiques et les contenus subjonctifs. Ainsi, il est possible d'envisager une fonction depuis ces premiers vers les seconds, et donc un flux d'information.

Par ailleurs, les paires peuvent être représentées non seulement comme des classifications de chaque objet du monde considéré par un token dans un type qui caractérise sa signification cognitive, mais également par une fonction depuis le type épistémique vers le type subjonctif ou contrefactuel.
Pour en rester dans le cadre proposé par D. Chalmers[235], l'intension est organisée en plusieurs niveaux, entre intension primaire (qui regroupe les intensions centrées sur l'interprète et les intensions secondes, qui caractérisent des significations détachées de l'ancrage dans l'expérience d'un individu. Les canaux peuvent satisfaire cette dimension.
Plus précisément, un scénario, qui structure l'intension primaire du concept ou de la proposition, se définit par un individu et un temps dans ce monde. L'intension seconde est composée de mondes possibles dans lesquels la proposition peut se vérifier.
On peut appliquer cette proposition à notre cadre de travail. C'est par rapport à une intension seconde qu'il est possible d'interpréter une expression portée par un individu, et l'interpréter en tant que proposition formulée par cet individu. On peut généraliser cette hypothèse et l'appliquer à propos des résultats fournis par les outils de calcul. On peut ainsi caractériser le fait que les résultats des opérations peuvent être considérés à la fois dans le cadre de l'univers des outils, de ce qu'ils disent de l'individu considérant les possibilités expressives des outils (leur monde propre et les théories qui ont permis de les produire), et enfin de la position d'acteur interprétant les résultats (relativement aux mondes des outils et de l'individu).

D. Chalmers utilise la notion de monde possible, qui comme on l'a vu, n'était pas nécessaire pour caractériser l'interprétation des expressions. En positionnant des situations à l'intérieur de chacune des constructions, il devient possible de limiter les mondes possibles d'interprétation et de construire un domaine dans lequel l'interprétation de ces expressions est circonscrite.

Concept de portabilité et généralisation de l'hypothèse.
Les ressources cognitives ne sont pas de nature contingente et doivent pouvoir être « portatives » ce qui signifie utilisables dans des contextes différents. Les outils de calcul externalisés peuvent être suivant les disponibilités, soit une feuille une calculatrice, un

portable, les doigts, etc. En ce sens, les ressources seront utilisables dans des circonstances différentes, à l'aide d'objets différents. La dimension portative concerne donc en premier lieu les ressources externalisées, mais également les constituants intensionnels des paires. L'apprentissage consiste alors à établir des relations à des objets externes distincts de façon à produire des routines.

Le langage fait partie des objets de la cognition externalisée, en ce qu'il couple les processus mentaux et le monde extérieur. Ce monde extérieur peut être effectivement une écriture, dans le cas d'une pensée qui s'élabore dans le cours d'une écriture, donc de l'externalisation d'un état, mis en relation à un état suivant.
Par couplage, il faut entendre que le processus se réalise à la fois partiellement dans l'esprit et dans l'outil externe. Si l'on dispose en interne de certaines relations entre des mots et différents objets possibles, c'est la relation établie à un contexte (qui peut être une écriture en cours) qui dirige la sélection des unités et de leur contenu.
Si l'on suit cette hypothèse, la production langagière mobilise à la fois la production antérieure dans l'ici et maintenant, mais également l'ici et maintenant du monde et de l'acteur. Dans une allocution, tout ce que je viens de dire est externalisé et se trouve donc en relation, dans une paire, par rapport à ce que je suis en train de construire.

Ainsi, si le langage fonde le couple pensée-expression, la langue écrite constitue une externalisation de plus haut niveau, puisqu'elle va servir de médiateur pour l'utilisation d'autres outils et donc d'autres paires. Ainsi, on peut envisager des structures externalisées plus complexes, entre autre parce qu'elles mobilisent pour une même unité comme peut l'être le langage, des ressources différentes : les modèles externalisés de textes (longueurs et contenus pour un article), les ressources d'informations déjà écrites (dépêches d'agence, sources, écrits d'autres journaux), axe choisi au début du texte et contraignant l'ensemble de l'écriture, et enfin vocabulaire (ou classes de mots) décliné.

La cognition externalisée ne rend pas seulement compte de calculs localisés. Elle nécessite le principe de portabilité, à savoir le fait que certains calculs, représentations et mémoires, peuvent être inscrits des outils différents et pouvoir être exploités dans le cadre de relations différentes. Ces relations et ressources inscrites dans des relations peuvent être mobilisables dans des contextes différents pour former des systèmes cognitifs localisés : « external coupling is part of the truly basic package of cognitive resources that we bring to bear on the world ».

La première ressource portable est justement la relation, et l'une des plus importante est le langage : les unités du langage permettent à la fois une référence (qui est la relation ressource) et une certaine dépendance par rapport au monde externe dans lequel s'exerce la parole (à savoir la façon dont s'exerce la première relation). Ce rapport sera caractérisé par un mode d'existence particulier des objets. En effet, « eau » et « H_20 » caractérisent un même objet, mais dans un mode d'existence différent, autrement dit d'un mode d'apparence du monde extérieur. La notion d'univers devient alors fondamentale dans la construction des rapports au monde, en considérant d'abord qu'il s'agit d'une relation entre le monde ici et maintenant et une certaine possibilité d'être. Cette formulation s'inspire des univers possibles de S. Kripke, mais caractérisés par une dimension relationnelle : un univers est mis en relation avec celui dont il est question maintenant et ici.
Ainsi, les univers se caractérisent comme des paires ordonnées depuis des univers possibles vers des extensions, et dès lors qu'un univers serait choisi, les autres seraient contrefactuels.

Sur les croyances et les états mentaux.
Les processus externalisés reposent sur certaines croyances, comme par exemple le statut qui est donné à certaines ressources de constituer un outil fiable, adéquat pour calculer ou représenter un certain fait.
Ce qui caractérise l'esprit, en dehors de l'ici et maintenant, c'est le fait de créditer une ressource, un monde possible, pour supporter une relation au monde extérieur. L'ensemble des univers possibles est caractérisé par ce qui est envisageable pour décrire telle ou telle part du monde dans lequel on est. La limite est alors constituée par ce qui est ou pas envisageable.

Le crédit constitue une croyance, à savoir la causalité d'un choix relationnel. Or ce crédit peut être associé à des objets externes, comme par exemple des outils de mesure (et notamment le rapport entre une échelle et une unité). Ainsi, si l'extension est un patient, et l'univers choisi une batterie de mesures physiologiques, la croyance réside dans la relation établie entre cet univers choisi et les autres possibles comme représentation adéquate de ce patient. Par conséquent, la ressource est également constituée de ce que ne représente pas cette représentation, mais qui est envisageable en fonction de ce qui est représenté.

Par-là, « l'extension de l'esprit » ne concerne pas seulement quelques processus de calcul, mais bien le fonctionnement de l'esprit.
Cela dit, un rôle essentiel est donné au langage, comme membre d'une part essentielle des couples comme des états mentaux externalisés. Simplement, la dualité donne une importance élevée à la dimension intensionnelle et à sa structuration autour de la distinction entre signification, connaissances apriori et possibilités. La connaissance apriori est caractérisée par la reconnaissance par le sujet d'un phénomène perçu dans le cadre de connaissances déjà traitées. Cette situation est celle de la cognition non épistémique présentée par F. Dretske. La signification correspond à l'intension de l'expression dans les différents mondes possibles où cette expression a un sens. Elle reflète le profil modal de l'expression, les choses ou les propriétés actualisées. Enfin, les possibilités correspondent aux contrefactuels et aux connaissances subjonctives. Ce système est relationnel : la connaissance a priori définit la connaissance contrefactuelle, de même la signification actualise et enrichit la connaissance a priori et enfin limite les possibilités de contrefactuels.
La dimension épistémique caractérise ce qui peut être connu, la signification ce qui est possible et nécessaire et enfin, les contrefactuels caractérisent la nature des choses du monde.

Enfin, les composants externes des paires jouent ainsi un rôle dans l'ancrage du raisonnement à l'intérieur du monde, considérant ainsi un ici et un maintenant fondateur. Cette indexicalité est encore renforcée par le postulat d'un égocentrisme fondateur de la dimension interne de la cognition. Il faut entendre par égocentrisme le point de vue individuel systématique, qui ne peut exister que par rapport à son couple, lequel est une entité externe. Par conséquent, dans cette proposition, l'impact du monde sur l'individu est direct.

Remarquons que les propositions de D. Chalmers sont antimatérialistes au sens où elles réifient le rôle de la conscience dans la cadre de la sémantique des expressions. Néanmoins, ce propos doit être intégré dans le cadre général de la cognition située et distribuée, considérée par D. Chalmers, et non dans la seule sémantique des expressions du langage naturel.
Nous développons maintenant ce point, qui met en perspective la proposition de D. Chalmers.

Conséquences de l'approche du langage : statut des représentations.
La cognition située et distribuée s'appuie d'abord sur des travaux remettant en cause le fonctionnalisme symbolique dominant, notamment sur la perception et certaines propositions neurologiques, de biologie cognitive et d'anthropologie cognitive (A. Clark & D. Chalmers, *The extended mind*, op. cit., p.10). Les problématiques d'anthropologie cognitive, entre autre parce qu'elles s'appuient sur des perspectives évolutionnistes, font émerger d'autres principes d'organisation et de fonctionnement de la cognition.

Néanmoins, de telles propositions posent un certain nombre de problèmes autant au fonctionnalisme linguistique (cf. J. François, p. 6[236]), pour lequel l'adaptativité linguistique constitue un argument circulaire, tout autant qu'au formalisme, pour lequel la langue n'aurait pas de lien direct avec son contexte d'usage. La plupart des perspectives que l'on a présentées jusqu'à présent omettent la spécificité du langage en tant que système de symboles articulés et structurés.
On peut donc considérer que le langage constitue un problème pour les théories de la cognition située et distribuée. Egalement, ce problème est accentué par la volonté qu'ont les théories de la cognition située et distribuée de substituer aux modèles de la cognition fondés sur le symbolisme, des modèles relationnels dans lesquels les discours s'inscrivent à l'intérieur des processus cognitifs (voir par exemple P. Agre[237], pp. 28-33). Néanmoins, il n'est question guère que de discours, et non de langage.
Dans la mesure où nous n'avons pas à prendre en compte la dimension linguistique dans son ensemble, du fait notamment d'un vocabulaire réduit et spécialisé, tout autant que d'une syntaxe normée.

Par ailleurs, ces propositions, fondées sur les relations entre un sujet et son environnement, ne permettent pas de penser le langage autrement que par des formes qui circulent au sein des processus décrits. Ainsi, on peut considérer que l'on n'a pas analysé des entités linguistiques mais des formes de représentations symboliques.
Les représentations sont des constructions externalisées et socialisées, ce qui permet de caractériser des textes, et plus précisément des modèles de textes (dans lesquels s'exerce l'énonciation individuelle) comme des mémoires et leur actualisation par l'intégration de toute information nouvelle à l'intérieur de chaque feuille.
C'est donc l'ensemble de la structure épistémique (de ce qui apparaît crédible), qui est redéfini par cette caractérisation représentationnelle et mémorielle des textes. La conséquence de cette externalisation des mémoires est l'accroissement de la précision (et l'amoindrissement des risques d'interprétation erronée) des propositions thérapeutiques.

Ces représentations sont distinctes des symboles, en vertu de la distinction -d'abord méthodologique- constituée entre l'activité cognitive interprétative et la compétence linguistique. Au départ, cette distinction est considérée comme une critique de la cognition symbolique, pour laquelle toute activité de raisonnement manipule des symboles (voir W. Clancey[238])]. Dans un second temps, cette distinction ouvre la possibilité de caractériser plusieurs niveaux d'abstraction dans la cognition et de placer au centre de celle-ci la question de la perception. Cette intégration permet d'aborder de façon nouvelle des objets comme les textes, en ne considérant plus seulement leur dimension symbolique, mais visuelle et matérielle.

L'externalisme a comme conséquence une sous spécification des contenus des mots et une prégnance du contexte dans la détermination de l'interprétation. On entend par là le contexte

linguistique, mais également, en suivant la cognition située et distribuée, le contexte des outils de médiation. Or de ce point de vue, les objets et outils représentationnels ne réduisent pas leur interprétation discursive à celle des expressions symboliques, mais à l'ensemble des constituants d'une représentation, comme les marques planaires ou plus généralement, les objets considérés culturellement comme des symboles ou des indices voire des icônes.

La cognition située et distribuée ne développe pas une théorie ni un modèle de l'information. Seule la théorie des situations et des flux s'intéressent, au travers du concept d'infon, à cette question. Elle s'est essentiellement intéressée soit au raisonnement dans le monde, soit à la façon dont l'esprit pouvait, dans un tel cadre, être pensé et comment son rôle défini.

D. Chalmers s'appuie pour cela sur les concepts d'action épistémique et d'action pragmatique. L'action épistémique altère le monde de façon à aider et augmenter les processus cognitifs tels que la reconnaissance et la recherche. (Les actions épistémiques altèrent le monde parce qu'un certain changement physique est désirable). L'action épistémique requiert un crédit épistémique, à savoir qu'une partie du monde soit une partie des processus cognitifs.

Nous ne pouvons prétendre avoir établi un modèle de n'importe quelle activité : nous avons établi quelques parties de ce modèle, sachant que l'activité que l'on a analysé repose uniquement sur de l'information et consiste à en produire. En ce sens, l'activité que nous avons analysée est à la fois exemplaire de l'activité scientifique et de toutes les activités entièrement dématérialisées, reposant donc essentiellement sur de l'information symbolique. Par conséquent, l'adaptation de posologie peut être considérée comme exemplaire des activités s'exerçant aujourd'hui essentiellement à l'aide du web.

D'une approche sémantique des flux vers une approche par les connaissances.
Comme nous l'avons déjà évoqué dans notre présentation antérieure, les pharmaciens travaillent avec un modèle de population, qui est une base de données contenant des populations de patients et répertoriant les comportements cinétiques de ces patients, en fonction de leurs propriétés physiologiques. Cette mémoire externalisée et structurée n'a pour le moment guère été exploitée dans notre modélisation.
Nous avons essentiellement focalisé notre propos sur les inférences réalisées dans le cours de l'activité sans considérer la façon dont les mémoires externalisées sous forme de bases de données pouvaient influer non sur l'interprétation des expressions mais sur les propositions de doses et les inflexions de stratégies thérapeutiques.

Si l'on intègre la dimension des mémoires externalisées, les flux pourront caractériser d'autres inférences que celles qui sont précisément liées à des interprétations d'expressions. En ce sens, on établit la distinction entre une approche sémantique des flux et une approche par les connaissances à partir du moment où l'on introduit cette base de données et les calculs qu'elle permet (à savoir des probabilités de comportement de patient relativement à ceux de la population envisagée). La dimension de l'action requiert, dans l'adaptation, ce passage par la mémoire externalisée.
Notons tout de même que l'adaptation de posologie à l'aide d'une base de données et de calculs probabilistes ne constitue pas la seule méthode possible pour l'adaptation. Des calculs plus sommaires, ne nécessitant que la mémoire des opérations, rendent cette activité possible.

Notre modélisation peut apparaître en retrait par rapport aux possibilités offertes par le cadre théorique de la cognition située et distribuée. Nous aurions pu poursuivre l'élaboration de ce

modèle. Néanmoins, les limites de notre caractérisation de l'information et de sa circulation rendent une telle modélisation encore peu pertinente dans l'immédiat. Les développements de l'activité, et notamment de l'activité scientifique sur le web rendent cette perspective crédible à moyen terme.

5.3.3. Solidarité des processus matériels et symboliques : principe de bi-dimensionnalité.

Les résultats de l'analyse précédente montrent que l'on peut considérer un parallélisme entre les processus matériels et symboliques d'une part, et des activités cognitives interprétatives d'autre part. C'est cette hypothèse que l'on voudrait maintenant étayer, en envisageant non plus seulement l'hypothèse cognitive mais la modélisation des phénomènes de cognition étendue. Il s'agit pour nous d'ouvrir un certain nombre de pistes, considérant que le développement des activités sur le web pourrait s'accélérer.

La bi-dimensionnalité de l'information constitue une hypothèse demandant d'être sérieusement étayée, notamment en ce qui concerne son lien à la connaissance. (On entend par bi-dimensionnalité le fait que la signification est parallèlement intensionnelle et extensionnelle, et non pas complémentairement. Un argument pour ce parallélisme est l'existence de contrefactuels). En effet, l'opération présupposée est purement une inférence en interprétation : à la suite des paires de D. Chalmers, on envisage la bi-dimensionalité comme un type de représentation qui permettrait d'associer la dimension de l'activité à la circulation d'information.
C'est la raison pour laquelle il nous apparaît fondamental d'inscrire notre définition de l'information en référence aux théories de la cognition qui postulent une distribution de l'activité cognitive entre des sujets, des objets et des dispositifs.

Pour donner une idée de la pertinence de la bi-dimensionnalité, il est essentiel de considérer des phénomènes tout-à-fait empiriques dont on n'arrive pas vraiment à rendre compte jusqu'à maintenant. Il s'agit de l'ensemble des interruptions, des retours en arrière, des révisions, que l'on peut observer dans le cadre des adaptations de posologie. En d'autres termes, il ne saurait y avoir symbiose entre les trois univers, mais bien quelques difficultés à valider une représentation dans un univers à l'intérieur d'un autre.

Représentation de la bi-dimensionnalité dans le cadre de la théorie des situations.
La bi-dimensionnalité constitue un concept d'IA qui a été intégré par J. Seligman & L. Moss (Jerry Seligman and Larry Moss. « Situation theory". In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*., op. cit.) dans le cadre de la théorie des situations. J. Seligman & L. Moss en donnent une représentation formelle, en continuité des principes de la théorie des situations, mais le l'associent pas à une conceptualisation.

Dimensions relationnelles des logiques dynamiques et caractérisation des transitions.
Une logique relationnelle ne constitue pas un modèle différent de la logique classique (P. Blackburn, *Representation, reasoning, and relational structures: a hybrid logic manifesto.* op.cit, p. 341), mais une façon de représenter des liens sans avoir à caractériser leur spécificité. Par exemple, une relation d'héritage quelconque peut être représentée par une relation, donc sans avoir à spécifier quelle est la direction du lien entre les entités. Ce premier argument sert à définir quels objets, quelles structures d'objets et enfin quelles opérations constituent les unités de l'objet d'étude. Parallèlement, les objets sélectionnés et leurs opérateurs ne sont pas

expliqués dans le cadre de ces logiques, ce qui permet de les exploiter selon des théories et des hypothèses variées. (Néanmoins, un tel propos présuppose que l'on considère toute explication comme étant de type fonctionnel donc causale ; nous y reviendrons).
Les logiques relationnelles ne sont pas simplement pour nous un mode de représentation, mais une façon de considérer les objets sur lesquels on travaille : ils sont décrits par les relations qu'ils entretiennent et font évoluer. « Dans cette perspective, la représentation n'apparaît plus comme un état mental doté d'un contenu linguistique qui tiendrait lieu d'un objet prétendument originaire, mais comme une *activité relationnelle* où le sujet et l'objet d'une visée intentionnelle co-adviennent. Simondon (1989a) appelait déjà *transduction* ce type de relation, où les deux pôles n'existent que dans leur relation réciproque, où la relation elle-même constitue les deux termes sans que l'un précède l'autre ou vice versa. A titre d'exemples, en peut citer les odeurs ou les couleurs, qui n'existent en tant que telles que dans leur relation avec un sujet constitué d'une manière particulière ; sans organismes dotés d'un système nerveux central et de dispositifs de couplage sensori-moteur particuliers, il n'y aurait que des molécules chimiques ou des ondes électromagnétiques » Havelange, V., Lenay, C., & Stewart, J. *Les représentations: mémoire externe et objets techniques*, op. cit., pp.3-4).

Perspectives articulant activité, langage et information.
Nous introduisons maintenant quelques perspectives qui permettraient de modéliser l'aboutissement de notre analyse de l'activité dans le contexte d'une prise en charge plus globale par le web de l'activité humaine.
Nous avons considéré des modèles comme les flux et les structures d'information comme étant régis par des principes qui sont ceux de l'information. L'information est inscrite dans une activité dont elle constitue le fondement de la dynamique. L'information inscrite dans une activité constitue un fondement pour une approche de la cognition (et plus précisément des raisonnements) en exercice, dans des situations pratiques. Dans ce cadre, le langage est considéré comme une ressource au sens de Wittgenstein. Cette perspective permet d'éviter de considérer la totalité du système linguistique comme une unité à étudier. Elle permet surtout de considérer les unités linguistiques dans le cadre de contextes contraignants qui peuvent être ceux d'une activité.
Les derniers travaux de R. Cooper donnent une idée de ce que pourrait être une approche articulant langage et activité dans un cadre de sémantique interprétative. La restriction par rapport à cette perspective est qu'elle ne prend pas en compte la dimension distribuée de la cognition.
Si R. Cooper argumente son propos sur son expérience à propos de la modélisation du dialogue et notamment des jeux de question-réponse, la visée de son propos est plus générale (R. Cooper, *Type Theory And Semantics In Flux*, op. cit., p. 273) : « The idea there is that natural languages are to be seen as toolboxes (resources) that can be used to create limited languages for use in particular language games in the sense of late Wittgenstein. These limited special purpose languages may be formal in the sense that Montague had in mind. We will argue, however, that there is a lot of linguistic interest in trying to discover not only how natural languages provide these formal languages but also how agents using the language apply and develop these resources which are constantly in a state of ux as we use the language. We will argue that our particular kind of type theory is appropriate for such an analysis whereas the kind of semantics of the classical model theoretic approach represented by Montague does not provide us with enough structure to capture the notions of variation in meaning that appear to be necessary. »
Les ressources ont comme intérêt de pouvoir à la fois être considérées comme des structures cognitives et linguistiques et de pouvoir être typées. Cette dernière propriété est essentielle pour l'intégration des outils dans la modélisation. Le concept d'enregistrement (« record »)

permet de représenter des cadres dans lesquels les dimensions linguistiques sont intégrées à l'intérieur de graphes représentant des opérations. L'intérêt du propos de R. Cooper est qu'il permet d'intégrer dans un cadre unique à la fois des dimensions cognitives et linguistiques. Plus particulièrement, il permet de lier la représentation des situations par des types et une caractérisation beaucoup plus précise de la signification lexicale que ne le fait généralement la sémantique formelle. Ainsi, le lien opéré avec FRAMENET peut avoir un rôle important pour la spécification d'une relation entre les représentations lexicales et les modèles d'opérations dans les activités.
Bien entendu, cette base de travail n'intègre pas explicitement la dimension bi-dimensionnelle ni les flux. Néanmoins, le fondement de cette modélisation dans la théorie des situations rend l'intégration de ces dimensions possible.

Une autre perspective est relative au raisonnement lui-même et à sa caractérisation dans un contexte d'activité réelle. La théorie des situations et les flux (on ne parle pas ici de la sémantique des situations, qui constitue une perspective de linguistique formelle développée à partir de la théorie des situations) ont été construits dans l'objectif de fonder sur des entités psychologiquement et référentiellement réalistes un modèle logique. K. Devlin développe ce programme de façon constante. Néanmoins, il reste dans un cadre essentiellement mathématique et ne propose pas de méthodologie permettant d'articuler les perspectives de Sciences Humaines et Sociales et de logique de façon suffisamment précise pour que l'on puisse déduire des premières des conclusions qui pourraient être utilisées dans un cadre de formalisation[239].
Nous avons proposé une méthodologie par laquelle la modélisation formelle caractérisait certaines régularités de comportement des phénomènes observables dans le cadre des activités.

Travaux relatifs à l'application de modèles d'activité dans le cadre du renseignement de procédures.
Dans le domaine médical et pharmaceutique, en premier lieu sous l'impulsion de la transparence des procédures d'expérimentation et des protocoles d'étude, un travail important comment à être mené pour caractériser ces activités.
La ressource ethnographique est généralement utilisée pour une modélisation utilisant UML (Unified Modelling Language)[240]. Ces travaux, s'ils permettent une représentation exacte des protocoles, ne prennent pas en compte la dimension cognitive et se limitent à la représentation d'une procédure.
De tels modèles ne prennent pas en compte le rôle de l'information dans la réalisation du cours d'action, ce qui assimile la représentation à celle d'un workflow. Un standard comme celui-ci permet de caractériser un processus pour sa reproductibilité, mais n'intègre pas justement l'information, qui, a contrario, différencie chaque réalisation de ce protocole.

En posant un présupposé cognitif, pour la caractérisation des outils, on ne fait pas qu'argumenter le fondement cognitif de la structure d'information. On argumente la dimension culturelle des opérations réalisées par les outils documentaires, notamment les classifications. On associe à des opérations structurantes d'un domaine de connaissances un fondement universel, caractérisé par des primitives, et pouvant être exprimé par un langage considéré comme universel, celui de la logique et de la sémantique formelle.

Les classifications et les thésaurus peuvent être considérés comme des réalisations particulières, dans le contexte des bibliothèques et des centres de documentation, de compétences nettement plus générales, correspondant à des facultés humaines fondamentales.

Cette relation à l'anthropologie cognitive est motivée par un postulat de l'ensemble de notre travail, à savoir le fait que le raisonnement est fondamentalement culturel. Ce positionnement théorique a des conséquences importantes dès lors que les outils d'information peuvent être utilisés dans le cadre d'activités. En effet, jusqu'il y a peu, l'information nécessitait un outillage technique tel (pour interrogation, la navigation, la sélection) que la recherche ne pouvait être menée dans le cours d'une activité professionnelle. Or, à partir du moment où des outils techniques permettent cet usage et que les outils d'information s'adaptent à cette nouvelle situation, on peut reconsidérer le rapport qu'il y aurait entre l'information et les cours d'action professionnels.

Il existe une autre raison qui permette justifier cette problématique de l'activité : la progressive intégration de l'information dans l'activité quotidienne. Comme le mouvement s'est déjà engagé en médecine et pharmacie, au sein des hôpitaux, les terminaux de type portable, tablettes et portables réduisent la distance entre la pratique informationnelle et le déroulement de l'activité. Cette relation entre le cours d'action d'une activité professionnelle et les pratiques professionnelles est émergente ; elle est encore peu étudiée parce qu'elle demande à la fois une modélisation très précise de l'activité en cours et une spécification des modalités et des stratégies de la recherche d'information, étant donnée la spécificité du support et de ses fonctionnalités.

La cognition située et distribuée constitue un cadre dans lequel ont pu se retrouver des perspectives philosophiques (en philosophie de l'esprit tout d'abord), de sciences sociales et plus particulièrement d'anthropologie cognitive, et enfin les acteurs de l'IA. Cette corrélation constitue un phénomène qui ne s'est pas reproduit dans le cadre de la structuration du web de données. En effet, les langages de représentation ne sont guère confrontés à des questions d'automatisation de raisonnement humain…

Gap entre théories de l'information et théories de la cognition située et distribuée.
Les propositions de D. Chalmers sont encore très peu utilisées dans un cadre de modélisation, notamment dans le web de données, à la différence par exemple de la théorie des situations. Les raisonnements de D. Chalmers ne sont pas accompagnés d'une modélisation, ce qui rend difficile un usage immédiat dans un contexte de conception.
Par ailleurs, la question de l'information n'est pas centrale dans les problématiques philosophiques de la cognition située et distribuée. En effet, les auteurs sont beaucoup plus préoccupés de la structuration de l'esprit dans ce cadre plutôt que de la proposition de modèles permettant d'utiliser cette représentation de l'esprit et de la cognition. Par ailleurs, les théories de l'information qui se développent partiellement à partir des travaux de J. Barwise et J. Perry, ne se préoccupent pas de questions de cognition.
Or, les constats sont les mêmes : on ne peut éliminer le rôle que jouent les outils, les artifices de représentation et les dépôts de mémoire externalisés dans l'activité cognitive humaine. La

grande différence entre ces deux perspectives est que les théories de la cognition ne se préoccupent guère du raisonnement alors que les théories de l'information vont se centrer sur cette question (sans nécessairement se référer à un cadre cognitif explicite (qu'il soit psychologique ou anthropologique).
Notre travail tend à amorcer un rapprochement entre ces perspectives, considérant que les flux et les structures d'information constituent des expressions et des processus que l'on peut analyser dans un cadre interprétatif.

Conclusion.
Nous avons dans cette partie caractérisé la structure d'information que l'on a pu établir à partir des flux et à la suite, défini précisément la sémantique qui lui était associée. Cette sémantique, si elle pouvait être représentée d'un point de vue de sémantique formelle, était fondamentalement associée à une représentation cognitive particulière des univers de référence. La modélisation de ces mondes a permis une articulation forte entre les questions d'information et celles de cognition située et distribuée.

Le concept de neutralité génétique de l'information, que nous avons identifié comme un trait marquant de la définition, est validé par notre modèle, de même que la dimension communautaire (ou sociale) de l'information.
On explique ce principe de neutralité par le contrôle de la symbolisation au travers des opérations.

On vient d'identifier un cadre théorique qui permettrait de mettre en valeur et d'assurer un fondement pour le déploiement de la caractérisation de l'interprétation dans le contexte d'une activité. Ce projet pose deux questions :

- Relativement à l'information, nous entrerions dans une problématique totalement différente. Nous nous limitons donc à montrer les implications de notre caractérisation des flux et des structures d'information dans le domaine des études de la cognition.
- Relativement à la finalité de notre travail, notre objectif n'est pas de concevoir des substituts à l'activité humaine, mais seulement de représenter sur le web des relations entre des données structurées hétérogènes. En ce sens, la pertinence d'une théorie de la cognition reste encore sujette à interrogations.

Néanmoins, c'est sur le lien entre ces perspectives et les questions des bibliothèques numériques que nous aimerions conclure, de façon à montrer ne perspective de recherche particulièrement productive et en continuité de l'ensemble de notre projet.
Les bibliothèques numériques et l'e-science ne se posent guère les questions d'intégration de l'utilisateur dans les outils qu'ils mettent en place, considérant essentiellement jusqu'à présent l'exploitation des potentialités offertes par les langages de représentation et les formats qui sont aujourd'hui à disposition. Or, aujourd'hui, de nouveaux matériels permettant d'utiliser l'information documentaire deviennent d'un usage commun. Ce sont les mobiles, tablettes et autres portables. Ces matériels rendent possible une recherche et une consultation d'information directement en situation de travail. Alors que la recherche d'information a toujours été considérée comme une activité autonome, mobilisant la totalité de l'attention de l'utilisateur, on a un changement profond dès lors que l'information peut être recherchée dans le cours d'une activité professionnelle. Cette question ne constitue pas une utopie, mais correspond, notamment en médecine, un élément essentiel du développement de l'offre informationnelle.

L'usage des smartphones, des ordinateurs portables et des tablettes devient de plus en plus important dans le cours d'une activité professionnelle, ce qui implique une modification de la façon dont on peut utiliser l'information en situation de travail, notamment dans un cadre de prescription. Par ailleurs, les bases de données, notamment THERIAQUE, proposent des applications « mobile » qui permettent cet usage. Comment faire en sorte que ces applications soient directement articulées à des activités de choix décisionnels dans une pratique quotidienne [241]?

L'information dans les unités de soin constitue une problématique qui donne lieu à de plus en plus d'expérimentations et donc d'outils et de retours d'usage[242]. Ainsi, on dispose aujourd'hui d'un corpus de connaissances relativement important sur le sujet et sur lequel nous pourrons nous appuyer pour proposer une analyse de Sciences Humaines et Sociales.

Des études ont montré que les praticiens modifient leurs décisions en considérant l'information scientifique à disposition[243].

La question de l'usage des dispositifs d'information pourrait devenir ainsi solidaire de celle de la menée des activités professionnelles. Elle va dans le sens de l'intégration des dispositifs d'information à l'intérieur des environnements de travail, et permet d'envisager la question des usages de l'information à l'intérieur de l'analyse des raisonnements menés dans les activités professionnelles.
Les usages de l'information s'intégreraient dès lors dans l'étude des activités professionnelles, et plus particulièrement dans le cadre des raisonnements insérés dans les cours d'action. Il résulterait de cette articulation une définition des usages par rapport à un cours d'action.

Du point de vue des bibliothèques numériques, il s'agirait de faire émerger un usage original par rapport à ceux des bibliothèques traditionnelles. En effet, comme on y reviendra, les structures de données des bibliothèques sont surtout considérées comme une facilitation de l'accès aux données et une plus grande fluidité de la navigation. Or, en considérant les questions de l'activité, on peut envisager l'information comme un dispositif totalement intégré dans le cadre d'une activité professionnelle. L'usager serait considéré comme un professionnel, dans le cadre de la menée de son activité, et pour lequel l'information constituerait un outil très particulier de travail. Or, justement, cette activité doit être considérée comme un cadre contraignant au sens où il spécifie très exactement l'information qui sera recherchée. La caractérisation de cette activité, des raisonnements et des configurations d'objets comme d'outils qu'elle mobilise peut ainsi être considérée comme un contexte d'usage particulièrement spécifié.

PARTIE. 6. Elaboration et mise en place d'un projet concernant les bibliothèques numériques et dans le cadre du web de données.

Une finalité de notre travail consiste à élaborer un projet permettant de produire un outil pour les bibliothèques numériques dans le cadre du web de données. Nous allons reprendre l'ensemble de ce qui vient d'être présenté et en montrer l'intérêt dans le cadre d'un projet web. Nous revenons donc maintenant à ce contexte dont nous avons parlé au début du document où nous avons effectivement décrit un contexte institutionnel, technologique et organisationnel qui permettait le développement des bibliothèques numériques, de la mise en commun des ressources et leur enrichissement mutuel. Nous allons y inscrire un projet dont nous allons donner une formulation relativement générale, sachant que les versions les plus précises ont été présentées dans le cadre des réponses aux appels d'offre ANR[221].
Les travaux dont nous venons de parler dans les parties intermédiaires, soit la plus grande partie du travail, se sont déroulés juste avant la montée en puissance du web de données. Ils ont pu commencer, à l'instar des flux, à s'appliquer de plus ou moins accentuée, dans le cadre du web de données. Nous allons donc pouvoir caractériser à quoi ces connaissances, qui sont le produit d'observations ou d'analyses de données linguistiques, peuvent contribuer à formuler des projets de recherche.

Le projet que l'on présente maintenant constitue un exemple d'application du modèle général dont on a présenté les contours. Il ne constitue qu'une application relativement restreinte du modèle général. Cette partie concerne un projet en cours, qui plus est collectif. Ce projet ayant par ailleurs évolué récemment du fait d'un changement de contexte d'application, un certain nombre de pistes sur lesquelles nous avions travaillé pourraient être abandonnées. Nous présentons donc maintenant plus un travail de positionnement qu'une réalisation effective, qui elle est en devenir.

La principale évolution dont il est question ici concerne la structuration du web et les services fondés sur cette structuration. Comme nous allons pouvoir le constater, les langages du web de données constituent la meilleure application possible pour la logique des prédicats. Ils constituent donc un étonnant champ d'application pour les outils de la sémantique formelle. Ces outils existent et ont été élaborés avant le web de données. Par exemple, on retrouve dans une publication sur l'apprentissage d'ontologie à partir de texte dans le cadre du web de données la référence aux DRT de H. Kamp, qui sont au départ des outils construit pour la levée d'ambigüités relatives aux anaphores dans le cadre du traitement automatique des langues[244].Ainsi, ce qui change vraiment avec le développement du web de données, c'est un environnement, des méthodes et des objectifs d'application pour ces modèles. Une large part de cette partie sera dévolue à la caractérisation précise des enjeux et des positionnements dans le cadre du web de données, une autre part étant plus particulièrement dédiée à l'application des outils que l'on vient de présenter dans ce cadre.

[221] Le travail effectué dans le cadre du montage de ce projet est collectif. L'ontologie, le « pattern » pour l'extraction et les structurations lexicales sont en cours d'élaboration. Avant publication, nous ne voulons pas nous approprier ce travail collaboratif et partagé. Nous resterons donc relativement discret sur ces travaux en cours. Seules les publications relatives à ce projet et les dossiers de réponse aux appels d'offre pourront attester concrètement du travail mené par le consortium.

Nous considérerons cette approche sous trois angles différents :

- 1. L'émergence et la construction de projets.
- 2. La question de l'usager.
- 3. Les questions relatives à l'analyse et au traitement des unités symboliques.

La première partie de notre travail est consacrée à la formulation générale et le contexte du projet. Nous préciserons la façon dont il s'insère dans le cadre de projets similaires. La seconde caractérise la façon dont notre projet s'articule et produit des usages au sein de communautés d'usagers qui sont en train de se construire à l'intérieur du web. Nous nous intéresserons plus particulièrement à la reconfiguration des communautés scientifiques autour des pratiques émergeant au travers de l'e-science. Enfin, nous développerons plus précisément les enjeux de notre projet dans le cadre plus fondamental des structures conceptuelles et du langage, et notamment dans la relation qui se construit au travers de la description des documents et de son évolution.

6.1. Emergence de projet.

Conformément à ce que nous venons d'énoncer, nous aimerions insister sur le caractère générique du projet, donc sur sa portée dans le cadre des problématiques liées à la circulation et à l'enrichissement des contenus informationnels.

L'idée de départ de ce projet est que des données hétérogènes mais liées d'une façon ou d'une autre (utilisation de l'un par l'autre, importance de l'un dans la genèse de l'autre, etc.) peuvent fournir des informations sur leur contenu respectifs. Ainsi, par exemple, ce que je peux décrire d'un outil me permet de postuler quelques compétences de son utilisateur. (A titre d'exemple, le soc nous apprend sur les savoirs, les tâches et l'activité du laboureur).

Le postulat, dans le cadre des bibliothèques numériques et du projet de l'e-science, est que la mise en relation des publications avec les données primaires qu'elles utilisent (comme les corpus dans le cadre de la linguistique) permet de montrer des propriétés de contenu de ces publications. Ainsi, le choix de données primaires pour réaliser une recherche est une source d'informations sur les résultats de cet apport.
Dans le cadre d'une recherche d'information, la description documentaire peut avoir d'autres applications que la seule identification du document qu'elle décrit ; elle peut concerner tous ceux qui sont en relation de contenu avec lui. Jusqu'à présent, l'exploitation des mises en relations repose sur les liens par citation et co-citation, au travers d'outils comme CITESEER, ou sur des liens entre personnes, via FOAF. Le projet VIVO[245] s'inscrit également dans cette perspective. Simplement, qu'il s'agisse des liens entre personnes, ou des liens de citations, ou encore des liens de thématiques, on exploite dans ces outils des liens de complémentarité ou d'usage (comme les citations).
Néanmoins, tous ces outils ne proposent que des liens, jamais des enrichissements mutuels. Or, dans la cadre de l'e-science, les liens entre ressources sont marqués par des inférences au sein de structures de connaissances (qui elles-mêmes peuvent assembler des structures de connaissances déjà constituées).

Le projet prend donc racine dans le type de phénomène dont le modèle des flux veut rendre compte : comment représenter le lien entre des structures de données hétérogènes ?
Ainsi, par exemple, entre des signaux visuels comme le code de la route et des comportements physiques de conduite, il n'y a pas de relation homogène, comme il peut y en avoir entre le code et sa signification. Néanmoins, il y a bien une relation puisque l'on obéit de

façon réflexe aux injonctions du code.
Rendre compte de ce fait est essentiel à partir du moment où des objets de plus en plus hétérogènes intègrent le web. Simplement, ces objets hétérogènes sont des constituants des activités humaines (et plus particulièrement de recherche), et c'est par rapport à cette activité que cette relation entre ressources hétérogènes prend sens.

L'autre contrainte fixée par les flux est que ces deux objets hétérogènes doivent être insérés à l'intérieur d'une structure de données. C'est le cas à la fois du code de la route et des comportements physiques de conduite. Les objets sont classés, ce qui permet d'envisager la régularité du phénomène, mais également de circonscrire la portée du modèle.

Nous allons donc maintenant expliquer de quelle façon on a pu construire un projet à partir de cette base théorique et du constat d'une multiplication des structures de données hétérogènes sur le web.

6.1.1. Un projet centré sur une application pour les flux et rôle des autres modèles.

Les flux ne constituent pas un modèle permettant à lui seul d'élaborer un outil quelconque. Il requiert des méthodes d'autres disciplines et d'autres objets (comme par exemple les ontologies) pour construire un projet : les flux permettent de poser un problème et donc certaines façons d'y répondre.

Ainsi, le fait que les flux traitent toute relation entre structures de données hétérogènes requiert la caractérisation de ces structures de données. En effet, les structures qui servent de fondement à la mise en relation se doivent d'être explicites et régulières. Ces structures ne sont pas prédéfinies par les flux : si nous associons les flux à une ontologie et à des structures lexicales, c'est du fait de contraintes liées aux données elles-mêmes, en l'occurrence à des métadonnées.

Les structures de données (ontologies, mais aussi lexiques sémantiques) ne caractérisent pas d'inférences entre données hétérogènes. Elles sont fondées sur la représentation des termes ou des concepts d'un domaine et de leurs relations. Les ontologies, les lexiques sémantiques mais aussi les usages (envisagés au sens des usages sociaux) se caractérisent par une cohérence conceptuelle propre définie par l'unité du domaine. C'est en ce sens que l'on parle d'ontologie de domaine, de lexique de domaine, etc.

Or, le problème est que cette unité de domaine ne permet pas de rendre compte de l'hétérogénéité des données et de leur lien. Elle ne permet pas de rendre compte du fait qu'une activité, professionnelle notamment, est fondée sur des outils fonctionnellement hétérogènes mais qui par leur assemblage et leurs relations, construisent l'unité de l'activité (Voir C. & J. Keller, *Cognition and Tool Use*, op. cit.).

Comme nous l'avons déjà évoqué, l'application essentielle des flux s'est déroulée dans le cadre des alignements d'ontologie. Les ontologies constituant les premières structures de concepts numériques, elles constituaient le meilleur objet d'application pour la théorie. Marco Schorlemmer[246], en 2010, élargit la portée des flux dans le cadre du web de données. Il reprend les arguments relatifs à la nécessité d'une interopérabilité sémantique. Celle-ci ne peut être aisément obtenue lorsque l'on a affaire à des systèmes hétérogènes. Les méthodes

classiques d'alignement d'ontologies[247] n'intègrent pas la possibilité d'interaction ni la situation dans laquelle la relation entre les deux structures intervient. Or, justement, les flux intègrent un niveau dit des tokens qui caractérise la situation occurrente et individuelle. Les flux permettent de fonder une relation inférentielle entre deux systèmes hétérogènes sur la base d'une situation occurrente. Les flux s'appliquent donc à des situations précises, comme par exemple le lien entre une donnée primaire de la science et les publications qui les utilisent. Dans ces situations, les flux représentent les unités d'information circulant entre deux objets appartenant à cette situation et permettant d'expliquer un changement d'état à l'intérieur de cette situation.

Comme nous l'avons déjà indiqué, les flux travaillent à partir de données de niveaux d'abstraction différents et surtout leur intérêt réside dans la prise en compte de la variabilité des situations occurrentes.
La variabilité des contextes est ce qui caractérise à la fois les publications et les ressources qu'elles utilisent : elles traitent toutes de thèmes particuliers, envisagés au travers de problématiques différentes, etc. Ces variables sont justement ce qui nous intéresse de représenter. En effet, tant que l'on reste dans le cadre de données bibliographiques, les représentations sont homogènes. Par contre, à partir du moment où on entre à l'intérieur des contenus, y compris de façon restreinte, on se situe au niveau des contextes et des variabilités individuelles. Tout l'intérêt des flux consiste par un système de classification à associer des entités de ce niveau des réalisations occurrentes (ou tokens) à des entités d'un niveau d'abstraction plus élevé (ou type). Ainsi, il est aisé de caractériser les entités occurrentes en les classant.

Dans le cadre de l'e-science, on a deux structures de données ; d'une part les descriptions des publications et les publications, d'autre part les ressources primaires sur lesquelles s'appuient les recherches dont il est question dans les publications, et qui sont elles-mêmes décrites.
Les relations entre les deux reposent sur des données structurées que sont les métadonnées de l'un et de l'autre. Par ailleurs, chacune de ces entités (les publications et les ressources), portent un discours sur leur propre usage pour les données primaires, et sur l'usage qu'il a été fait de ces données primaires dans le cadre de la recherche relatée par la publication.
On a donc considéré que les métadonnées étaient des représentations au niveau des types, et les descriptions d'usages des entités de niveau token.

Ainsi, si les données primaires nous apprennent à propos des publications, les publications nous renseignent à propos des ressources. Ce double mouvement, qui constitue une paire contra-variante de fonctions, fonctionne depuis les contenus des publications vers l'ontologie des ressources primaires et inversement, depuis les données primaires vers les publications.

Les questions liées à l'extraction de séquences d'information, à l'établissement d'une ontologie (modulaire) des ressources primaires et à la construction d'un lexique électronique sont liés à cette construction des données destinée à faire fonctionner les flux. Ces questions sont déterminées et contraintes par les besoins présentés par les flux.

La diversité des apports disciplinaires et leur complémentarité tient à la façon dont on a travaillé. En effet, nous avons tout d'abord construit un modèle, permettant de faciliter la compréhension et la représentation d'un phénomène, pour ensuite envisager une segmentation des problèmes, des approches et des outils, de façon à ce que chacune puisse travailler corrélativement et de façon complémentaire. Cette perspective a permis de construire progressivement l'environnement de ce flux : ontologies et métadonnées tout d'abord, puis

usages, lexiques et extraction d'information. Ainsi, un flux ne constitue pas une ontologie, ni un thésaurus, mais à l'intérieur de ces structurations, il contribue à la reformulation des problèmes de ces outils de par la façon dont il les met en relation. Par exemple, les flux prenant en charge certaines inférences, l'ontologie proposera des structures de concepts permettant ces inférences.

6.1.2. Hypothèse fondamentale du projet et domaine.

Notre projet s'inscrit dans le cadre de la description des documents et explore plus particulièrement la question des contenus.

L'idée que l'on défend ici, c'est le fait que les flux permettent de faire circuler des contenus entre des données structurées hétérogènes : à partir du moment où un lien est caractérisé entre des objets distincts, décrits par des structures de connaissances hétérogènes, les structures peuvent être enrichies l'une par l'autre à partir du moment où le lien (de causalité, de contrainte) entre les deux objets a pu être caractérisé. Il est effectivement possible, grâce aux langages du web et aux structures de données disponibles, de mettre en relation deux structures de données de façon à enrichir une description dans notre cas, mais également d'améliorer la qualité d'une activité (comme la recherche d'information par exemple). (Les ontologies bibliographiques comme BIBO entre autres, servent à structurer des descriptions bibliographiques de façon à faciliter les mises en relations internes des données. Mais ces descriptions ne sont pas enrichies les unes par les autres parce qu'elles ne sont pas inscrites dans un processus caractérisé comme peut l'être une recherche scientifique). Or, ce que l'on apporte, en utilisant les raisonnements mis en œuvre dans le cadre des flux, c'est le transfert de descriptions depuis une structure vers une autre : ce transfert caractérise ce que l'une apporte à l'autre dans le cadre de l'activité de production et donc représente certains contenus de ce dernier. Cette proposition constitue une avancée scientifique importante dans le cadre de l'information numérique, et nous aimerions donc la présenter dans son ensemble.

On entend par contenu l'opération et le résultat du choix d'une entité symbolique parmi l'ensemble des entités possibles à partir d'un prédicat existant. Cette définition reprend celle que nous avions présentée dans la partie 3, en caractérisant l'information. Dans notre cadre, un contenu est entendu comme le résultat d'un flux, à savoir d'une opération sur des ensembles d'entités symboliques substituables. On doit donc distinguer cette définition, qui concerne les opérations des flux, de la référence des attributs et valeurs dans les publications qui sont décrites. A l'aide de ces métadonnées, on ne décrit pas le « contenu » de la publication, mais certains traits contextuels qui spécifient le sens du discours. Cette relation entre le contenu des métadonnées et le discours dont elles spécifient la signification est conforme à la définition de la sous-spécification du sens des unités linguistiques et des discours et de leur spécification contextuelle.

Appréhender globalement l'e-science constitue un défi trop vaste pour un seul projet. Nous avons choisi le cadre des ressources linguistiques. Ces ressources ne sont pas liées à une seule discipline, qui serait la linguistique, mais à l'ensemble des acteurs académiques et industriels les utilisant. Notre domaine est celui des sciences Humaines et Sociales, en lien avec les sciences de l'Ingénieur, essentiellement l'informatique, à travers le Traitement Automatique des Langues.
Ce domaine explique que l'on soit plus attentif aux données primaires et à leur traitement que, par exemple, aux conditions de reproductibilité dans les Sciences du vivant et notamment la médecine (cf. par exemple l'équipe « Information Magangement Group » de l'Université de

Manchester[222]).
Dans le domaine qui nous a jusqu'à présent concerné, celui des ressources linguistiques, la représentation des données primaires se limite à des métadonnées associées aux corpus et outils. On exclut les annotations parce qu'il s'agit de descriptions intégrées dans les données primaires elles-mêmes (qui par conséquent portent mal leur nom). Le problème se pose relativement différemment pour les données primaires reproductibles (notamment en biologie) ou aux essais cliniques. Dans ce cadre, une modélisation est nécessaire pour représenter de façon normée le protocole mis en œuvre. Un effort de normalisation internationale est d'ailleurs à l'œuvre actuellement[248].

Un autre intérêt de ces ressources est qu'elles offrent un panel relativement complet des outils d'analyse et de traitement en ligne. En effet, mis à part le recueil des données (comme des conversations par exemple), l'ensemble des procédures de traitement et d'analyse peut se faire à l'aide d'outils disponibles sur le web. On dispose ainsi d'une panoplie diversifiée de l'ensemble des outils de traitement, données structurées voire même de modélisation nécessaires à une analyse des données linguistiques. L'effort de structuration terminologique et lexicale constitue une dimension également très importante de la construction du domaine.
Enfin, d'un point de vue plus économique, les ressources linguistiques entrent dans des domaines d'application très importants, notamment dans le cadre du plurilinguisme, de la recherche d'information et de l'indexation, et par conséquent font l'objet d'une attention soutenue.

Globalement donc, cette idée d'enrichir la description d'objets par celle de ceux qui entrent dans leur élaboration trouve une application particulièrement appropriée dans le cadre des données linguistiques. Elle s'inscrit par ailleurs dans le développement de l'e-science vu comme l'intégration de l'activité scientifique (émergence d'hypothèses, méthodologies, collection de données, traitement, analyse et diffusion de résultats) dans des procédures en ligne.

En définitive, si notre projet consiste à enrichir certaines métadonnées par d'autres, auxquelles la première ressource est en relation, on peut considérer cet aspect comme une première étape dans un raisonnement plus vaste. On peut poursuivre en considérant que les relations entre publications et ressources utilisées par ces dernières ne sont pas simplement fonctionnelles. Les publications (et avant tout les recherches qu'elles relatent), constituent des usages des ressources, et donc, en extrayant les contenus des publications relatifs aux ressources, on a accès à la façon dont certaines ressources ont été utilisées.
Ainsi, notre projet peut avoir un autre aspect : en développant considérablement l'extraction d'information (comme nous le verrons au travers de structures d'information), il permet de rendre compte de la façon dont une communauté relate son usage d'une ressource.
La perspective de l'extraction a une autre conséquence : elle oblige à considérer les contenus statiques (comme les discours) et à ne plus opérer seulement à partir de descriptions. On peut donc envisager une mémoire et une analyse de ces relations. Cela se fait sous forme de web service et d'une base de données servant à enregistrer les résultats.

[222] http://img.cs.man.ac.uk/

6.1.3. Caractérisation du projet par rapport aux bibliothèques numériques et aux évolutions des métadonnées.

Les deux types d'objets hétérogènes sont les publications et les ressources utilisées par ces publications. Ces deux types d'objets sont décrits par des outils eux-mêmes hétérogènes, à savoir des jeux de métadonnées distincts. De façon emblématique, le Dublin Core pour les publications, et les métadonnées IMDI, CMDI, BAMDES[249] et OLAC pour les ressources.
Nous aimerions positionner notre projet par rapport à ceux qui ont cours dans le cadre des bibliothèques numériques et de l'édition de métadonnées. On illustre ainsi deux mouvements distincts, qui sont d'une part l'accentuation des mises en relation de ressources, et d'autre part l'intégration des activités scientifiques dans les descriptions de documents. Ces deux mouvements doivent être associés à un troisième, qui caractérise plus globalement la structuration du web utilisant ce que l'on peut savoir de l'utilisateur.
Ce positionnement permettra aussi de clarifier notre travail par rapport aux usages et en premier lieu les liens à l'utilisateur final. En présentant la politique générale des projets des bibliothèques numériques, on comprendra l'usage de toutes les mises en relation de métadonnées et par là même de notre projet.

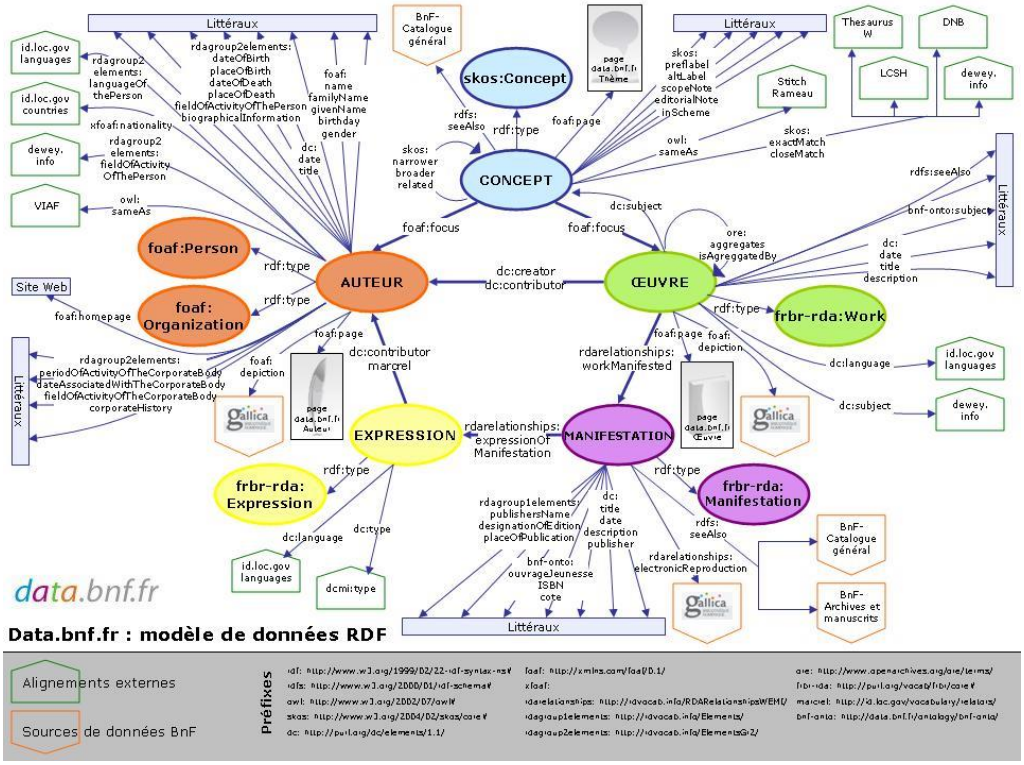
Les jeux de métadonnées des ressources linguistiques portent de l'information sur le document, mais considèrent toujours une distinction entre le travail descriptif du document et la description de contenus intégrés dans ce même document :'annotation et l'édition de documents structurés étaient jusqu'à présent distincts de la description documentaire de ces ressources. Une ressource est ainsi le fruit d'un travail de construction des données primaires.

La prise en compte des contenus (non pas au sens sémantique et informationnel tel qu'on l'a défini plus haut, mais au sens formel des données composant le document) va demander l'utilisation d'outils, de méthodes, et de modèles qui sont extérieurs à l'univers scientifique des bibliothèques numériques : l'extraction d'information, par exemple, n'en fait pas partie. Dès lors, notre projet vise à proposer dans le cadre des outils des bibliothèques numériques de nouveaux objets, notamment en ce qui concerne l'utilisation des contenus.

La réussite de notre projet passe systématiquement par des opérations sur les descriptions de documents. C'est pour cela que les métadonnées jouent un rôle crucial dans notre projet : elles constituent des outils dont on se sert pour mettre en relation les objets et sont des outils sur lesquels on agit pour proposer d'autres usages.
Notre objectif consiste d'abord à introduire des éléments descriptifs des contenus (par leur contexte) dans les jeux de métadonnées. On peut toujours rétorquer que certains éléments des jeux contiennent déjà des indications sur le contenu. On peut prendre comme exemple les attributs <subject>, <description> et éventuellement <title> du Dublin Core. Néanmoins, ces attributs ne sont pas normalisés, ce qui entraîne une imprécision, dommageable en termes de recherche d'information. Les efforts pour associer des vocabulaires contrôlés et des thésaurus à ces attributs ne résolvent le problème que du point de vue de la normalisation des descriptions, et non de celui de la spécification du sens des discours.
Notre projet vise donc à utiliser les relations qui structurent de plus en plus l'offre documentaire de façon à les accompagner par l'explicitation des conséquences de cette relation. Il s'agit d'accompagner les relations de contenu informationnel.

6.1.3.1. Mises en relations opérées par les bibliothèques numériques.

Notre projet vise à améliorer la pertinence des métadonnées pour les bibliothèques numériques. Les métadonnées servant à proposer des descriptions pour la recherche d'information, il s'ensuit que notre objectif est articulé aux questions de recherche d'information. Nous considérons que les outils de recherche d'information existant évoluent dans le sens que l'on a évoqué plus haut, à savoir une prise en charge de plus en plus adéquate des métadonnées. La première possibilité qui s'offre alors pour améliorer les produits de la recherche d'information est inscrite dans l'adaptation des outils descriptifs de façon à permettre une meilleure représentation du contexte d'un document. Or, les métadonnées sont héritières des notices bibliographiques et sont fondamentalement statiques. Les relations opérées par des ontologies de l'activité scientifique comme VIVO ou encore les relations mises en œuvre grâce à BIBO ne constituent que des accentuations de mises en relations entre des données différentes. Le projet de la BNF de proposer un modèle de données relationnel s'inscrit dans cette perspective[223] (). En voici le schéma complet[224] () :

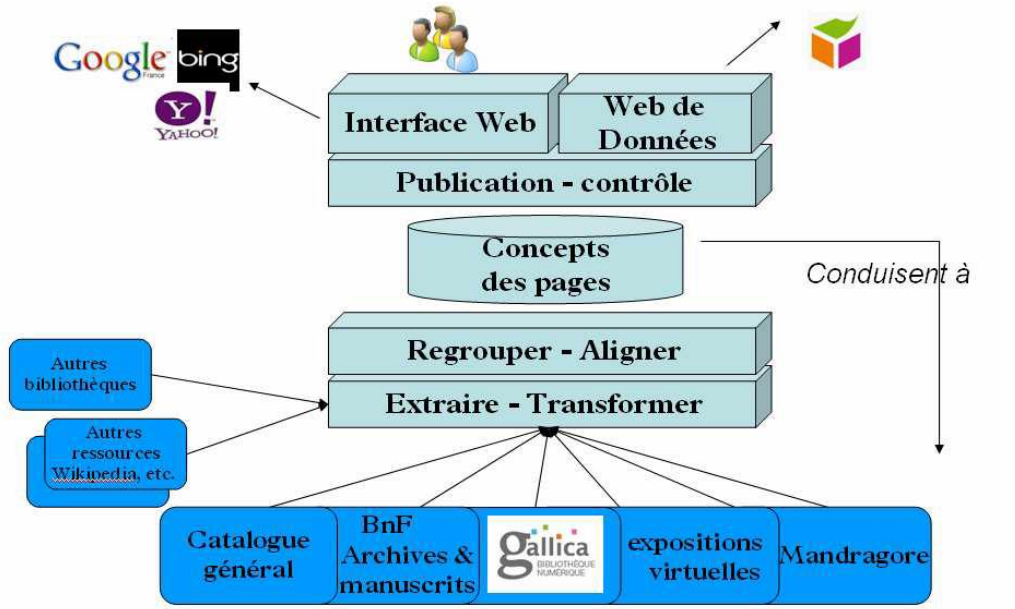


Le modèle proposé par le BNF, qui ne constitue pas un modèle d'e-science, vise à accentuer la mise en relation entre les documents de type œuvres et l'ensemble des structures de données descriptives de cette œuvre. Ainsi, à l'aide notamment de protocoles comme FRBR, il est possible de proposer une interopérabilité accentuée entre les différentes structures[225].

[223] http://data.bnf.fr/semanticweb
[224] http://data.bnf.fr/images/graphe_complet.jpg
[225] Nous pouvons également citer le modèle de données de la British National Library, qui est relativement similaire à celui de la BNF : http://bnb.data.bl.uk/docs

La proposition de l'e-science constitue une problématique différente par rapport à celle-ci parce qu'elle propose des métadonnées en lien à la menée d'activités de recherche et qui permettent d'accéder à du matériau de travail et non à de seules connaissances. Or, les questions suivantes : « Qu'est-ce que l'on recherche dans le cadre d'une interrogation ? Qu'est-ce que l'on cherche au travers des documents que l'on vise à extraire ? » ne donnent pas lieu à de mêmes réponses que l'on soit dans le cadre d'une activité de recherche ou dans un cadre de recherche d'information qui par ailleurs peut être utile dans le cadre d'une recherche.

La perspective illustrée ici par l'exemple de la BNF[226] ouvre des perspectives importantes en matière de navigation. En effet, l'interopérabilité syntaxique rend possible une navigation accrue entre des documents reliés. Ces échanges de données permettent une visibilité accrue des collections et des documents sur le web, comme l'illustre le schéma suivant http://data.bnf.fr/docs/databnf-presentation.pdf :



En somme, le but de cette structuration se situe dans l'usage commun du web, donc des moteurs de recherche généralistes. Les moteurs de recherche usuels du web sont considérés comme une donnée, à savoir que l'on construit l'ensemble des outils complexes dans l'objectif d'une meilleure visibilité pour ces moteurs.

A priori, ces perspectives peuvent sembler éloignées de nos préoccupations. Or, du point de vue de l'usager, le recours aux moteurs de recherche les plus faciles d'accès et d'usage

[226] Le projet de la British National Library est relativement similaire à celui de la BNF : http://www.bl.uk/bibliographic/datafree.html

constitue une pratique non seulement courante mais d'autant plus développée que ces moteurs eux-mêmes proposent des services reposant sur la mise en relation de données (voir par exemple l'évolution du SCHOLAR-GOOGLE[227], qui propose des profils chercheurs). Néanmoins, comme le montre le schéma précédent, l'utilisation « web de données » n'est pas ignorée pour autant : elle est considérée comme complémentaire et encore insuffisamment généralisée dans son utilisation pour pouvoir être privilégiée.

Comme on l'a dit dans la première partie de ce travail, il existe une complémentarité d'intérêt entre les moteurs de recherche et les organisations proposant des descriptions de documents. Notre projet possède donc ce versant : il vise à produire des métadonnées enrichissant les descriptions et permet donc d'améliorer la pertinence des recherches d'information sur le web. Or, il ne s'agit pas là du seul usage de notre travail. En lien plus précisément à l'e-science, notre projet propose une interface de navigation.

Retenons pour le moment que l'on situe le projet dans le cadre de cette mise en relation des données, qui constitue le cœur du travail actuel des bibliothèques numériques. Néanmoins, ce cadre n'est pas celui de l'e-science. A la différence de la BNF, les infrastructures de recherche sont intéressées au plus haut point par la perspective d'une intégration de l'activité de recherche dans leur offre. Néanmoins, il n'existe pas de connexion effective entre les projets d'e-science, qui devraient être hébergés dans des infrastructures spécifiques, et les bibliothèques numériques. Cela s'explique entre autre par le fait que les projets d'e-science se développent plus rapidement dans le cadre des sciences du vivant, et notamment la médecine, alors que les Sciences Humaines et Sociales privilégient la mise à disposition sur le web de leurs ressources. Les méthodes de travail des Sciences Humaines et Sociales (et dans le cas des ressources linguistiques la situation est encore plus complexe du fait de l'informatique) sont très diverses et ne permettent pas une standardisation comme celle qui opère dans le cadre des Sciences du vivant. Nous avons donc choisi une voie intermédiaire, qui permet de s'inscrire dans le cadre du développement des SHS sur le web tout en proposant, au travers du web service, des développements d'e-science.

Les bibliothèques numériques généralistes utilisent des jeux de métadonnées, et en premier lieu le Dublin Core. Par contre, à partir du moment où elles sont relatives à un domaine, elles peuvent utiliser une ontologie comme GOLD[228] pour la Linguist List[229]. DBLP reprend les descripteurs bibliographiques sous forme d'ontologie et propose des liens vers les moteurs de recherche (SCHOLAR GOOGLE, CITESEER). Mais aucune ne contient d'outils d'e-science. (GOLD est une ontologie des concepts de la discipline). La perspective reste celle de la structuration de l'information et de la mise en relation de cette information à d'autres ressources.

On modifie de façon assez importante la perspective en proposant de décrire des propriétés de contenu des publications, propriétés relatives non pas au sujet de ces publications (qui peuvent être correctement décrits par les mots-clés, le titre notamment) mais à la nature de la recherche menée sur ce sujet. Une telle caractérisation permet d'envisager des catégorisations différentes des ressources en opérant des liens entre des données hétérogènes, comme par exemple entre un corpus en linguistique textuelle et une publication en extraction d'information ou en text mining.

[227] http://scholar.google.fr/schhp?hl=fr&as_sdt=0,5
[228] http://linguistics-ontology.org/
[229] http://linguistlist.org/

Ces usages permettent une mise en valeur du fond de la bibliothèque au sens où il permet une approche complémentaire des celles des métadonnées usuelles dans un premier temps, et dans un second une révision de la façon dont les collections elles-mêmes peuvent être envisagées. En effet, on pourrait proposer des catégorisations par usage de ressources et non plus seulement par discipline ou thématiques.

Cette construction a une autre conséquence sur le positionnement de notre travail, non plus en termes scientifiques mais d'hébergement, ce qui constitue un problème stratégique essentiel. Si les formats utilisés (qui sont validés par des consortiums tels le W3C ou l'ISO notamment) sont universels, permettant ainsi une très grande liberté d'action, il n'en reste pas moins que dès lors que le web service constitue un outil comportant une mémoire (corpus d'apprentissage, résultats d'opérations), il requiert un emplacement de stockage. Or, la définition de cet emplacement est une question fondamentale pour la généralisation de l'outil. Ainsi, nous proposons que l'hébergement passe par des grands équipements, tels le TGE-ADONIS en France ou les infrastructures européennes CLARIN et DARIAH. Or, dans ce cadre, le lien avec les bibliothèques n'est pas assuré, sachant que les utilisateurs communiquent essentiellement avec ces dernières. Les infrastructures sont essentiellement des répertoires d'outils, des entrepôts et des ressources.
Dès lors, stratégiquement, le rôle des bibliothèques numériques reste fondamental parce qu'elles permettent d'assurer à la fois un lien à l'usager, aux publications et aux moteurs de recherche.

Il faut envisager pour expliciter cette question les différentes dimensions du projet : la distinction entre les usages internes des bibliothèques numériques et donc dans le cadre de la recherche d'information d'une part, et l'intégration dans le cadre de l'e-science d'autre part n'est guère pertinente dans la mesure où justement notre projet articule les publications et les ressources primaires.
L'idée fondamentale de la construction est que le service d'e-science est complémentaire de l'usage du service dans le cadre des bibliothèques numériques. Ainsi, il peut fournir une réponse aux questions que se posent les bibliothèques par rapport à l'e-science en proposant un outil hybride.

Cette réalité est loin d'être nouvelle si l'on considère les projets universalistes de Paul Otlet. La différence est que les problématiques documentaires sont totalement intégrées à l'intérieur de celles du web, à commencer par le lien entre métadonnées et formats de représentation, comme XML ou RDF. Simplement, cette intégration permet de prendre en compte des objets, comme les données primaires de la science, qui jusqu'à présent sont extérieurs au cadre des bibliothèques numériques. En deçà même de l'intégration d'un outil d'e-science dans le cadre des bibliothèques numériques, notre outil pourrait permettre d'inscrire les données primaires dans l'offre des bibliothèques numériques, ce qui constitue une question fondamentale pour les bibliothèques elles-mêmes. (Voir G. Puech, fondateur de PERSEE, communication personnelle).

6.1.3.2. Les jeux de métadonnées : usages et évolutions.

Notre propos considère les documents à partir de leur description : il s'intègre dans le domaine de la description documentaire, relativement notamment aux questions de métadonnées. Il trouve donc sa pertinence dans le cadre de la structuration du web. Sont donc concernés des langages de description et surtout leur connexion dans le cadre de services web.

Si les métadonnées constituent au départ une adaptation des notices bibliographiques au contexte de l'information électronique, le rôle qu'elles jouent aujourd'hui évolue nettement. Confrontées aux questions d'e-science, mais aussi des structurations des communautés d'usagers, ou encore de questions liées à l'invisibilité de données numériques, les métadonnées évoluent considérablement. On distingue trois thématiques fondamentales dans l'évolution des métadonnées :

- le recours de plus en plus fréquent aux langages contrôlés extérieurs,
- les profils, et les questions des communautés d'utilisateurs,
- les micro-données et l'insertion de descripteurs dans les documents eux-mêmes.

Ces trois enjeux essentiels contiennent en eux-mêmes certains paradoxes : d'une part l'articulation entre la normalisation des descriptions et la diversification des profils, d'autre part entre la consolidation des principes de métadonnées et le développement de micro-data, qui constituent des formes de taggage internes à la ressource.

Par rapport aux **langages contrôlés**, le DC propose toute une série de liens à l'ensemble des organisations de connaissances existantes. On n'entre pas encore dans le cadre des profils, mais simplement dans la normalisation des valeurs associées aux éléments. Un problème des métadonnées a toujours été l'écart entre la signification d'une valeur, vue par celui qui les renseigne, et la signification de cette même valeur pour celui qui interroge : le rôle des langages contrôlés est de résoudre ce problème. Les langages contrôlés permettent ainsi une meilleure découverte des ressources et surtout l'élimination de bruit dans l'affichage des résultats de recherche d'information. Nous verrons à propos de la DEWEY notamment que les systèmes de classification vont adopter leur propre stratégie concernant la représentation de leur structuration en RDF, ce qui ne manque pas de complexifier la situation. Cette intégration des langages contrôlés est par ailleurs un premier résultat de l'interopérabilité et constitue une régulation dans la production de métadonnées comme de leur interprétation.

Relativement aux profils, on peut reprendre la définition qu'en propose le Dublin Core, et qui sert de ligne directrice pour caractériser comment on entend un profil dans le cadre des métadonnées : « The term *profile* is widely used to refer to a document that describes how standards or specifications are deployed to support the requirements of a particular application, function, community, or context. In the metadata community, the term *application profile*has been applied to describe the tailoring of standards for specific applications."[230]
Il s'agit donc d'adapter un outil universel à des préoccupations communautaires sans perdre l'interopérabilité des descriptions.
Techniquement, une telle articulation repose sur le principe des URI, donc l'écriture RDF du Dublin Core. On relie ainsi un concept du DC avec un autre, appartenant à un standard voisin. Ainsi, pour notre cas, un profil est constitué par l'association d'un jeu comme l'IMDI ou l'OLAC, qui sont spécialisés, au DC. La métaphore du Lego est utilisée pour décrire cette construction.
L'application d'un profil se spécifie dans le cadre de la grammaire abstraite du DC. C'est au niveau des propriétés que le lien est établi avec les modèles du domaine spécifique à une communauté.
En ce sens, les profils requièrent un usage de plus en plus important du concept de relation.
Un profil comprend trois composants :

[230] http://dublincore.org/documents/singapore-framework/

- la caractérisation fonctionnelle qui définit ce que fait le profil et ce à quoi il sert, à savoir ses usages.
- La caractérisation du modèle, qui concerne les entités de base et leurs relations.
- L'ensemble des descripteurs du profil, qui caractérise l'ensemble des métadonnées qui constituent des instances valables du profil d'application.

Ce cadre reste très général et les descripteurs du profil ne sont pas encore définis. De plus, si l'interopérabilité syntaxique constitue un fait facile à définir, de nombreux problèmes se posent comme par exemple l'actualisation des versions de KOS : un thésaurus, une terminologie sont des objets vivants, et donc le profil doit pouvoir prendre en compte ce paramètre. Il en va de même des traductions des différents KOS.

L'enjeu de cette structuration communautaire est essentiel pour traiter des images, qui du fait de leur nombre, de leur volume et de leurs usages, nécessitent un traitement communautaire. Pour les ressources scientifiques, un tel traitement apparaît également opportun parce qu'il permet une meilleure découverte de ressources et une meilleure préservation[250].
L'intérêt de cette publication est qu'elle donne une image complète des différents jeux de métadonnées spécialisés qui constituent autant de profils. L'inventaire n'est pas complet, car il manque entre autre les métadonnées IMDI, OLAC et LDC.

Les profils reposent ainsi sur des structures de données pouvant être relativement complexes, surtout qu'aucune limitation n'est associée aux types de structure associée aux profils. Ainsi, les propositions de M. Zumer & alii[251] esquissent la caractérisation d'un modèle de domaine utilisable pour caractériser des relations entre structures de connaissances diverses, ou KOS. Ce modèle permet de caractériser les ressources insérées dans le profil. Il s'agit d'un modèle permettant de décrire les ressources utilisées à l'aide des descripteurs suivants : caractérisation de la ressource comme un tout (« work »), caractérisation des différences « expressions » (notamment langages) de l'outil et enfin les différents formats de cet outil (« manifestation »).
Ces réalisations montrent que la question des profils est encore loin d'être close dans la mesure où le processus de leur caractérisation n'est pas encore établie. Cette question relativement technique se double de celle de l'usage même de ces profils : les métadonnées ne sont pas encore insérées dans des dispositifs d'e-science encore émergeants.

Les micro-données ont vu leur rôle accentué par la publication de HTML5[231]. Il s'agit de balises répertoriées (notamment dans SCHEMA.ORG), et qui permettent de tagger des structures dans des textes de façon à ce qu'elles puissent être reconnues et extraites par les moteurs de recherche. L'enjeu pour la visibilité des articles de presse sur le web, par exemple, est fondamental. Pour nous, il s'agit de premières caractérisations normalisées de contenu qui pourraient avoir un enjeu fondamental dans le cadre de relations entre les contextes caractérisés par les métadonnées et les contenus des publications.
Les micro-données sont fondées sur les tags structurés insérés dans les pages représentées en utilisant HTML5.

Ainsi, le lien entre la construction d'outils pour le web et les bibliothèques est relativement complexe. Si l'élaboration de métadonnées constitue un travail qui a pu être réalisé indépendamment des organisations du web, ce n'est plus le cas aujourd'hui : les LINKED DATA ont ainsi permis l'intégration syntaxique des métadonnées dans le cadre du web de données, mais comme nous venons de le voir, cette mise en commun est à la fois

[231] http://www.w3.org/TR/html5/

techniquement et intellectuelle complexe. Notamment par le biais de l'incubateur W3C Library Linked Data, qui comprenait la BNF comme membre à part entière, les questions de bibliothèques se sont insérées de façon plus précise à l'intérieur des problématiques développées par le W3C.
Néanmoins, pour bien comprendre ce problème, il est opportun prendre l'exemple de notre propre domaine d'application, parce qu'il montre assez bien comment peuvent s'articuler des métadonnées généralistes et spécialisées.
A la différence des métadonnées DC, les métadonnées IMDI ou encore IPTC sont élaborées en fonction de besoins professionnels et sont ensuite rédigées en utilisant les différents formats de représentation des données du web.

Comme nous l'avons évoqué plus haut, nous avons centré le projet sur les ressources linguistiques. Elles possèdent leurs propres jeux de métadonnées, dont une des particularités est quand même d'avoir une histoire relativement longue, marquée par une diversité importante au moins en termes d'institutions.
Une présentation en a été proposée par P. Wittenburg[252]. La première distinction opérée avec le Dublin Core a été établie par l'OLAC, à savoir des archives ayant des caractéristiques différentes des publications : le fait qu'il s'agisse de données relatives à une langue précise, située géographiquement, recueillies à un certain moment et par une certaine méthode. Or ces données ont été construites pour un usage précis, et cette question des usages est fondamentale pour d'autres éditeurs de ressources, comme le LDC, qui construit et annote des corpus afin de les vendre à l'industrie. Le projet de l'IMDI, situé dans le Max Planck Institute for PsychoLinguistic a été de représenter ces ressources en spécifiant l'ensemble des conditions de leur élaboration et de leur usage. Les métadonnées IMDI sont publiées à partir de 2003.

Depuis, les métadonnées IMDI ont été intégrées dans les propositions présentées par CLARIN. Les propositions METASHARE, qui sont intégrées dans un dispositif centré autour des technologies de la langue, sont globalement similaires à celles du dispositif CLARIN. Si l'on compare ces deux jeux, on constate une ressemblance fondamentale dans ce qu'ils décrivent des données. Si les deux jeux semblent effectivement faire doublon, et qu'ils sont par ailleurs financés dans le cadre de programmes européens, leur différence essentielle est liée au cadre de leur diffusion : les infrastructures, et surtout CLARIN pour CMDI, les plateformes pour METASHARE.

Aujourd'hui, les métadonnées IMDI constituent un profil à l'intérieur d'ISO-cat. Elles sont ainsi intégrées à l'intérieur de l'offre de l'infrastructure, et sont identifiées par un profil appelé « metadata ». (Les profils d'ISO-cat sont différents de ceux du Dublin Core).

En définitive, les développements sont marqués par des différences très importantes de choix stratégiques :

- Une perspective centrée sur la lisibilité des données sur les moteurs de recherche et l'intégration des données dans leur offre (y compris le développement de SCHOLAR-GOOGLE)
- Une perspective centrée sur la maîtrise de la structuration des relations entre données structurées. Il s'agit là de la perspective du Dublin Core et des acteurs des bibliothèques numériques. La position du Dublin Core se distingue des bibliothèques numériques par son indépendance vis-à-vis de n'importe quelle sorte de données numériques. en ce sens, il peut au mieux synthétiser les perspectives en cours de développement.

- Enfin, une troisième stratégie est le développement des infrastructures permettant le stockage et la disponibilité des outils et des données.

L'interopérabilité et la disponibilité des standards n'impliquent pas une uniformité des stratégies des acteurs. En effet, au-delà des standards, les différents acteurs ont construits des objets et des outils situés dans des pôles différents de la chaîne de l'information et donc adoptent des points de vue différents sur des enjeux communs.

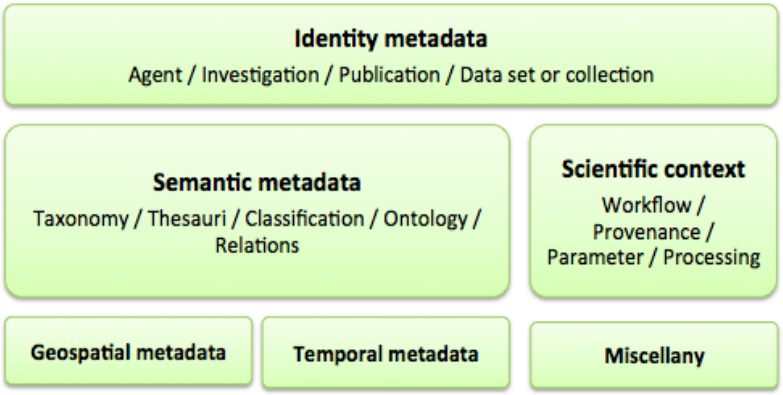
6. 1.3.3. Evolutions des formulations des métadonnées en lien aux perspectives de l'e-science.

Comme nous l'avons vu, les créateurs de jeux de métadonnées ont progressivement adopté les nouveaux formats d'expression proposés par le web. Ils ont accepté une syntaxe et ont adapté leur formulation aux contraintes. L'e-science constitue un de ces enjeux et les réponses apportées sont dépendantes des positions dans le domaine. Nous commencerons par caractériser la façon dont les jeux de métadonnées appréhendent la question, entre que les métadonnées mettent en relation les descriptions de documents et les modèles d'activité qui sont au centre des dispositifs d'e-science.

La question est formulée par J. Qin & alii[253] : comment peut-on articuler les questions d'e-science et les métadonnées ? A cela, on peut ajouter une autre question, celle des communautés qui sont impliquées dans le concept même d'e-science, et donc le lien aux profils.
L'idée généralement développée pour répondre à ces questions consiste à construire des niveaux de structuration, considérant que les besoins pour l'e-science sont relatifs à des dimensions d'usage des ressources : les descriptions des données doivent permettre de rendre ces dernières vérifiables, utilisables (que l'on puisse les explorer), analysables et interopérables. (P.4) Ces propriétés favorisent le développement d'outils dans le cadre des sciences expérimentales

Les métadonnées sont considérées traditionnellement comme des couples attribut/valeur et ce n'est qu'assez récemment que l'effort de structuration a été engagé de façon à caractériser les descriptions transmises comme des représentations conceptuelles des objets décrits. Ces réflexions du Dublin Core, motivées par l'adoption d'une représentation utilisant RDF, ont abouti à la modélisation des métadonnées. Cette modélisation permet ainsi, également, la relation avec les autres structures conceptuelles de représentation de l'information.

Une telle modélisation permet de caractériser plus précisément les objectifs d'une proposition de métadonnées pour les données scientifiques. Elles sont synthétisées de la façon suivante par Greenberg :



Un certain nombre de précisions doit être apporté à cette architecture. On entend par métadonnées sémantiques l'ensemble des vocabulaires structurés permettant d'identifier précisément le sujet de la publication. Cette définition du sujet permet de grouper les publications de ce même sujet par un système de relation. Cela suppose l'utilisation d'une structure lexicale ou conceptuelle reconnue dans une communauté. Par exemple, l'utilisation de l'Unified Medical Language System (UMLS) et du Library of Congress Subject Headings (LCSH) permettent de construire un lien par sujet, reposant sur l'usage commun du RDF. On retrouve ainsi les questions de navigation et de catégorisation des ressources.
Le contexte scientifique regroupe l'ensemble des descripteurs permettant l'usage des données primaires de la science.

Cette architecture permet de répondre au moins en partie aux critères définitoires des objets de l'e-science : les données doivent être vérifiées, utilisables (traitées, analysées) dans le cadre d'une recherche et donc interopérables. J. Greenberg considère leur description au travers d'un module spécifique ; néanmoins, jusqu'à présent, aucune proposition concrète de jeu n'est proposée dans le cadre du Dublin Core. On peut envisager des structurations différentes en fonction des domaines scientifiques. Par ailleurs, les métadonnées spécialisées comme celles que l'on a explorées ne sont pas à proprement parler adaptées à l'e-science, puisqu'elles ne sont relatives qu'à la description d'objets numériques. Néanmoins, elles peuvent s'y inscrire du fait de la possibilité qu'elles ont d'être structurées dans une modélisation de l'activité scientifique. C'est ce à quoi nous travaillons.

Pour Greenberg, il n'y a pas à créer de nouveaux jeux, mais simplement à structurer des descriptions existantes. Les principes utilisés sont ceux du moindre coût (qui caractérise le fait que les descriptions, notamment administratives, peuvent être récupérées), des infrastructures (la récupération des métadonnées spécialisées intégrées dans les bases de données contenant les ressources) et de la portabilité (à savoir le fait que le lien entre les descriptions contenues dans les infrastructures et celles développées dans les métadonnées généralistes se fait par le biais d'ontologies (« ontological methods, »)).
Une telle position présuppose que l'on ne se préoccupe pas de modèle de l'activité. Or justement, ces modèles sont nombreux mais hétérogènes. De plus, les métadonnées spécialisées, comme celles relatives aux ressources linguistiques, représentent implicitement au moins le processus d'élaboration de la ressource.
Nous exploiterons cette propriété implicite de ces métadonnées descriptives de ressources de façon à élaborer une ontologie de l'activité. Enfin, de quelle activité parlons-nous ?

Conclusion.
En termes de positionnement du projet, on observe une très grande distinction entre les stratégies des acteurs en fonction d'usages différents du web. Pour tous, il s'agit de répondre aux questions posées non seulement par la multiplication des données, mais par la diversification de ces données, de leurs usages, et la formation de communautés en lien à ces usages. Seul le Dublin Core fait le lien entre les publications, les ressources et leurs différentes structures conceptuelles de description et les profils. Même si les réalisations sont loin d'être abouties, cette position est la plus intéressante.

Pour les bibliothèques et les métadonnées spécialisées, les standards mettent en relation des documents hétérogènes et permettent une navigation entre les différentes ressources. Par contre, pour les jeux généralistes, l'enjeu consiste d'abord à assurer l'homogénéité des représentations (et donc l'adoption de leurs schémas) tout en prenant en compte cette dimension communautaire : la position du Dublin Core est centrale parce que ce jeu universel se dote progressivement d'une capacité à s'adapter à des offres communautaires et spécialisées.

Notre projet se situe dans cette perspective puisqu'il vise à proposer un jeu de métadonnées à usage communautaire tout en s'assurant de son intégration dans les jeux existant, et notamment le Dublin Core.
La grande différence entre notre approche et celle du DC réside dans l'utilisation d'un raisonnement. On entend par là le fait que les métadonnées sont produites à l'aide d'inférences automatisées.

6.1.5. Le domaine des lieux de réception, de gestion et de conservation de l'information. Les mises en relation de structurations.

Aborder la question de notre projet uniquement sous l'angle des bibliothèques numériques et des métadonnées peut sembler relativement restrictif dans la mesure où les évolutions marquées par les grilles laissent entrevoir de toutes autres possibilités, notamment relativement à l'accessibilité des contenus pour un traitement descriptif.
Nous aimerions maintenant traiter un peu plus précisément la façon dont le concept même de bibliothèque numérique est en cours de reformulation du fait des perspectives d'e-science, mais plus globalement de la mise en relation des données et de leur visibilité sur le web. Nous avons introduit cette question en présentant les propositions de la BNF. Simplement, nous n'avons pas développé les conséquences que peut avoir cette mutation sur les bibliothèques numériques et sur la façon dont la notion de collection change de dimension du fait de la visibilité sur le web.
Nous revenons sur l'offre structurée d'information utilisant les grilles afin de présenter la notion d'environnement de recherche. Puis, en considérant les mutations des outils liés aux bibliothèques (notamment les systèmes de classification), nous pourrons envisager la façon dont les dépôts de connaissance peuvent être considérés dans un cadre plus global. Cette analyse permet de comprendre l'enjeu de notre projet dans le cadre de la reformulation du rôle des bibliothèques.

6.1.5.1. Usages des grilles.

Même si l'on considère la réelle diversité des outils de recherche et donc l'accès à l'information, on reste dans une perspective où l'accès à la notice précède l'accès au contenu.

C'est cet ordre-là qui pourrait évoluer, et dont nous essayons de donner les contours.

On utilise les métadonnées pour permettre d'accéder à un aperçu du document, sachant que le document demande d'autres procédures (en l'occurrence le téléchargement) pour pouvoir être lu et plus globalement utilisé. Cette question de procédure est liée en grande partie aux limites des canaux de circulation de l'information, même si par ailleurs les capacités de transfert se sont très largement améliorées.

La mise en place des grilles s'inscrit dans le cadre d'une nouvelle phase de la mise en réseau des documents et se caractérise par le fait que les quantités d'information transmises augmentent considérablement. A ce titre, il serait possible de travailler directement sur les contenus. Nous estimons que les projets de grandes infrastructures (telles que celles qu'élabore l'UE) constituent des enjeux fondamentaux pour l'élaboration de valeurs ajoutées au document numérique. A titre d'exemple, une rupture pourrait bien être le fait que l'on définisse l'unité d'information par le quantum, et non plus le bit. Le quantum constitue une unité de cryptage de l'information beaucoup plus précise, et donc permettant de faire circuler un nombre beaucoup plus important d'informations par unité.
Ces orientations sont illustrées notamment au sein des projets CHIST-ERA de l'Union Européenne. Ceux-ci sont considérés comme des projets à long terme. Sachant qu'ils concernent entre autre les capacités de circulation et d'échange de l'information, ils sont situés très en amont de la description de l'information. Les travaux en cours concernent surtout la construction des différents constituants des infrastructures techniques (ou grilles), à distinguer des infrastructures reposant sur un serveur et liant différentes bases de données. Si ces infrastructures constituent une perspective à moyen, l'enjeu réside maintenant dans la caractérisation des outils pour l'échange, les connexions et donc les utilisations de ces données. C'est dans ce cadre que nous inscrivons notre projet : quels services il sera possible d'intégrer dans ces grilles, considérant que les capacités techniques de circulation de données permettent la transmission d'informations complexes à propos des documents. Néanmoins, on ne peut véritablement élaborer un projet intégrant ces mutations que si au départ on a pris en compte la façon dont les principaux acteurs concernés réagissent par rapport à elle. Les bibliothèques numériques constituent les premiers bénéficiaires de ces mutations et il est donc essentiel de comprendre comment elles se positionnent.

Les premiers acteurs à avoir réagi sont ceux qui sont directement associés au projet technique, à savoir les institutions de la recherche. Les possibilités techniques de traitement offertes par les grilles (GRID) montent en puissance. Les espaces de travail scientifiques commencent à voir le jour, comme par exemple l'environnement de recherche open-source SciDoc[232]. L'ensemble de ces projets est énuméré dans le Virtual Research Environment Collaborative Landscape Study de janvier 2010 (auteurs : A. Carusi et T Reimer).[233] eSciDoc présente l'avantage d'être une plate-forme pluridisciplinaire et modulaire. D-SPIN[234] est l'infrastructure d'e-science dévolue à la linguistique et incluse dans CLARIN. Ces plateformes (comme également en France le TGE-Adonis au travers d'ISIDORE[235]) commencent à fournir des services, notamment des répertoires. L'idée de mettre en relation

232 https://www.escidoc.org/

233 www.jisc.ac.uk/media/documents/publications/vrelandscapereport.pdf

234 http://weblicht.sfs.uni-tuebingen.de/index.shtml

235 www.rechercheisidore.fr

répertoires et collections de façon à accroitre la qualité de l'information émerge à travers des liens, via les possibilités expressives des RDF (Resource Description Framework), entre des publications et leurs références par exemple. Des outils existent et sont utilisés, notamment l'encodage TEI, un jeu de métadonnées, etc.

On vient de présenter rapidement un contexte largement à venir et dont les contours, notamment techniques, ne sont pas encore totalement fixes. Les services actuels des plateformes comme leurs capacités de traitement, sont susceptibles d'évoluer relativement rapidement d'une dimension de stockage vers des capacités de traitement et d'utilisation libre, reposant sur une véritable structuration des données.
Les promoteurs d'eSciDoc ne s'y sont pas trompés ; après avoir mis en place une infrastructure de stockage des données, établie avant 2010, leurs travaux se sont orientés plus spécifiquement vers l'élaboration d'une ontologie du travail scientifique fondée sur les métadonnées des documents scientifiques. Cette ontologie préfigure les outils d'e-science permettant de raisonner à partir de données primaires, d'outils et données préparées. Mais pour cela le cadre technologique n'est pas encore finalisé.
Enfin, il n'est pas non plus sûr que la configuration même des services d'e-science soit totalement établie. Des projets nationaux (de type TGE-Adonis ou EsciDoc) seront-ils privilégiés, ou au contraire des initiatives plus sectorielles, reliant différentes universités et sur des thématiques spécifiques, comme c'est le cas par exemple aux Etats-Unis (CINet de Virginia Tech[236]).

En définitive, le cadre de travail de l'e-science est extrêmement évolutif, entre autre parce qu'il mêle un nombre très important de technologies différentes, ce qui introduit un niveau de complexité tout autre que celui des métadonnées et des ontologies que nous utilisons aujourd'hui.
On peut néanmoins dans ce cadre caractériser un certain nombre de principes et d'orientation de notre projet qui lui permettra de perdurer quelle que soit l'évolution de l'e-science. Ainsi, la question de la distribution et de l'hétérogénéité des données, l'articulation entre les contenus et les descriptions, des web services évolutifs et une ontologie facilement actualisée.

On remarque que ces infrastructures se développent sans lien direct aux bibliothèques « traditionnelles », qui ne sont pas inscrites dans le cadre de CLARIN notamment. La distinction entre les bibliothèques numériques et les institutions chargées de gérer et développer les bibliothèques « traditionnelles » pourrait s'amoindrir dès lors que les mêmes outils pourraient être utilisés et que les bibliothèques numériques pourraient avoir un besoin plus important de structurer leur masse de données, entre autre du fait de l'élargissement de leur public, lui-même dû à un meilleur accès à l'information (voir notre propos sur les métadonnées). Justement, c'est ce que nous allons voir maintenant.

6.1.5.2. Usages des classifications, thésaurus et taxonomies.
Les outils servant à classer, donc à organiser les bibliothèques, se voient transformés aujourd'hui dans leur mise en œuvre technique (adoption de formats de représentation RDF et SKOS) et par conséquent, du fait des échanges de données (Linked Data) dans leurs usages. L'utilisation de la classification DEWEY comme langage contrôlé pour le renseignement des métadonnées du Dublin Core constitue un des usages innovants de ces classifications.
Pour illustrer cette question, on peut reprendre l'exemple de la DEWEY. Jusqu'à présent, elle était considérée comme un système classificatoire et permettait d'organiser les connaissances

[236] http://www.ci.uchicago.edu/escience2012/pdf/CINet-A_CyberInfrastructure_for_Network_Science.pdf

dans les bibliothèques. Le fait qu'elle soit décrite en utilisant les langages du web et donc d'une part qu'elle s'inscrive dans la sémantique de ces langages et d'autre part, du fait de la disponibilité et de l'interopérabilité des structures, qu'elle puisse être associée à des descriptions d'objets, constitue en fait une transformation radicale à la fois du sens et des usages de la DEWEY. Justement, si la signification, donc pour nous l'interprétation, de cette mutation pouvait être explicite, on aurait effectivement vraiment avancé.

La mutation technique des outils documentaires concerne d'abord leur disponibilité pour d'autres usages que la structuration des bibliothèques, et ensuite une autre définition des termes et de leurs relations est proposée par l'usage même de RDF. Un outil sensé au départ organiser les connaissances académiques pour les bibliothèques devient un vocabulaire contrôlé destiné entre autres à renseigner des métadonnées[237]. La classification décimale devient ainsi un service web. Il en résulte une perte d'autonomie de la Dewey par rapport à ses usages initiaux. Néanmoins, les URI de la DEWEY permettent d'associer à une ressource non informationnelle (un objet du monde, abstrait ou concret), un ensemble de ressources génériques et leurs représentations. Ces dernières sont des documents web décrivant l'objet. Néanmoins cette description inclut le fait que la classe associée à l'objet est elle-même inscrite dans une hiérarchie. Ainsi, on représente un aperçu (« snapshot ») de la collection à laquelle appartient la classe de l'objet. Cette collection constitue alors le type de l'objet. Ainsi, la description proposée pour un objet est double puisqu'elle concerne à la fois un composant descriptif et un composant classificatoire obligatoire. Ce dernier est géré par la relation « about ». L'expression fondamentale de la Dewey en RDF est la suivante :
http://dewey.info/{object-collection}/{object}/{snapshot-collection}/{snapshot}/about

Le cas de la Dewey est exemplaire parce qu'il s'agit de la plus ancienne classification. C'est dans le cadre des thésaurus que la transformation d'usage est la plus facile à développer, entre autre du fait du standard SKOS. Néanmoins, comme on l'a vu à propos des profils, SKOS constitue un outil pensé à partir de la structuration des thésaurus et ne possède donc pas la flexibilité nécessaire pour concevoir des relations entre des structures fortement hétérogènes. C'est pour cela que les liens FRBR sont préférés dans certains cas, comme pour l'édition de profils Dublin Core.
On peut également noter que la CDU (Classification Décimale Universelle) est également publiée en XML/RDF et utilise SKOS[238] . A la différence de la DEWEY qui cherche à conserver ses principes hiérarchiques fondamentaux, la CDU traduit son propre système de classification dans le cadre de SKOS mais sans véritablement s'attacher à préservation de la structure hiérarchique dans l'utilisation du vocabulaire SKOS. La spécificité d'une classification par rapport à une autre organisation de connaissances n'est pas prise en compte par l'UDC[254] (à la différence de la DEWEY qui utilise le système des « snapshots »). Pour RAMEAU, le problème est légèrement différent puisqu'il ne s'agit pas d'une classification à proprement parlé, mais d'un répertoire d'autorité-matière. Seuls les vedettes-matières sont représentées.

6.1.5.3. Transformation des outils de recherche d'information.
Nous aimerions maintenant caractériser plus précisément le fait que les infrastructures modifient considérablement l'accès à l'information, proposant un modèle intermédiaire entre la recherche d'information sur un portail et les moteurs de recherche sur le web.

Les stratégies des bibliothèques, notamment de la BNF mais aussi des Humanités

[237] http://oclc.org/developer/documentation/dewey-web-services/using-api
[238] http://udcdata.info/

Numériques, consistent à donner accès à des données primaires de la science qui sont aussi des éléments du patrimoine culturel. En ce sens, une politique d'ouverture des données comporte une dimension grand public, donc une visibilité par les moteurs de recherche généralistes[255].

Ces bibliothèques numériques ne sont pas seulement des portails existant, mais constituent les aspects les plus visibles d'une politique extrêmement prégnante. L'accessibilité des connaissances et des sources de ces connaissances ne constitue pas seulement un problème scientifique mais bien de diffusion des résultats vers l'industrie.
Qui héberge ces informations, ces connaissances ? Le W3C est une organisation assurant la structuration de la circulation de l'information, mais pas l'hébergement des documents. Les bibliothèques ne constituent pas les seules institutions aptes à offrir ce service et les bibliothèques numériques ne fournissent pas d'espace de stockage et de publication de données de la recherche. De plus, les enjeux ne sont pas seulement académiques : la recherche et développement, l'innovation et la compétitivité sont très largement impliqués dans ce processus. Ces infrastructures qui permettront de mutualiser des ressources très importantes et de réaliser des économies d'échelle impressionnantes constituent des lignes de force des politiques européennes et du NSF.
Cette préoccupation s'inscrit aujourd'hui dans des infrastructures permettant d'organiser et de partager des collections extrêmement importantes de document et de les structurer. Dès lors, la recherche d'information se développe parallèlement à l'intérieur des bibliothèques et des infrastructures, utilisant alors des outils paramétrés par le domaine. Par exemple, les outils de recherche de la bibliothèque numérique des ACM ou de DBLP sont effectivement paramétrés pour des recherches plus spécialisées, mais en même temps, constituent des outils de recherche proches de ceux utilisés dans les bases de données traditionnelles. Les outils de recherche fondés sur les technologies de type web sémantique et permettant d'interroger de multiples ressources ne sont guère, à l'exception d'EUROPEANA et de ISIDORE, que des projets. La mise en relation des données de la science et des publications n'est pas encore un fait. Justement, notre projet se situe à cette interface. Il permet de donner consistance aux relations entre données primaires et publications.

6.1.5.3. Sur la description des données.
Tout ce que l'on vient de dire donne une image relativement euphorique des évolutions technologiques à l'œuvre dans la mise en réseau des ressources. Néanmoins, il existe une très grande distance entre ces capacités techniques, voire l'investissement des institutions, et les descriptions des documents eux-mêmes, par leurs auteurs ou toute autre personne chargée de les mettre en ligne.
C'est ainsi qu'est né le projet Data-At-Risk Initiative (DARI)[256] . En partant du constat qu'une bonne partie des données scientifiques étaient insuffisamment décrites, Jane Greenberg et son équipe ont établi un panorama de ces données concernées. Comme le remarquent Zinn, Wittenburg et Ringersma[257], les données de la science qui servent de base pour l'e-science, ont longtemps été considérées comme des archives. Elles n'ont donc pas nécessairement été considérées comme des ressources vivantes, destinées à être réutilisées fréquemment. Elles n'ont donc pas été renseignées comme telles. Ce projet s'inscrit également dans l'idée que le web doit permettre d'obtenir des résultats exhaustifs relativement à une requête, et donc les mêmes fonctionnalités qu'une bibliothèque.

Justement, cette question du renseignement des ressources se pose dans notre cas de façon cruciale, et l'on ne peut que faire le même constat. Par ailleurs, certaines expériences comme BAMDES ont disparu du web. On a pu faire le constat que seules les ressources répertoriées

sur les serveurs comme OLAC, IMDI et dans une moindre importance BAMDES et le LDC, pouvaient être considérées comme visibles, à condition bien sûr que le renseignement ait été relativement complet.

La perspective d'OLAC, issue de l'OAI, reste dans le cadre de la disponibilité et du répertoire de ces archives. Grace à son outil de moissonnage (PMH) et dans une moindre mesure d'agrégation (ORE), l'OAI rend possible la visibilité des archives sur le web mais ne transforme pas leur statut. La publication très récente (février 2013) d'une nouvelle version montre un changement de la perspective : il s'agit d'un outil permettant de saisir les révisions, actualisations d'une ressource. La « ResourceSync Framework Specification »[239] permet d'actualiser des descriptions depuis un serveur source vers un serveur de destination. Ce protocole permet à la destination d'avoir des descriptions toujours à jour des ressources qu'elle propose.
En dehors des aspects techniques, cette spécification révèle une transformation de la perception des ressources : elles ne sont plus considérées seulement comme des données, mais comme des objets vivants, évolutifs.
Un autre phénomène important émerge dans cette spécification : les actualisations de ressources sont essentiellement associées à des parties de document et donc concernent très directement les contenus des ressources. Il n'est certes pas question de signification, mais de segmentation dans la ressource. Cette segmentation est temporelle (**<lastmod>**`2013-01-02T13:00:00Z`**</lastmod>**) mais peut contenir des éléments de description. Ces derniers concernent la nature et les formats de la transformation.

Les différents projets et propositions dont nous venons de parler montrent un écart important entre les possibilités descriptives et l'état des ressources à disposition. En d'autres termes, la disponibilité des ressources n'est pas égale et il existe un fossé entre les ressources dûment décrites et actualisées d'une part, et celles qui n'ont qu'une description sommaire, limitée aux seuls aspects administratifs et techniques.

Notre modèle, s'il se sert des métadonnées associées à la ressource, peut également être intéressant parce que grâce à l'extraction, il permet d'obtenir des descripteurs (issus de l'usage) pour des documents.

On doit tout de même préciser que les métadonnées ne sont pas nécessairement bibliographiques : les IPTC pour le journalisme (notamment photographique), ou le MPEG 7 pour le multimédia, constituent des ensembles de descriptions professionnelles de documents pour des objectifs qui ne sont pas bibliographiques. Elles caractérisent néanmoins des documents.

Une ontologie bibliographique se caractérise par le fait que les connaissances liées ont toutes trait à la caractérisation de documents. Les IPTC ne caractérisent pas le document mais certains paramètres de son contenu, afin qu'il soit pertinent dans l'échange des nouvelles.

Enfin la structuration des métadonnées dans les langages de type XML/RDF permet une interopérabilité forte, comme celle qui par exemple est développée par la société ADOBE, dans la plateforme XMP.

[239] http://www.openarchives.org/rs/0.5/resourcesync

6.1.5.4. Sur les bibliothèques et les connaissances.

L'ensemble des propositions précédentes et les questions soulevées par l'e-science ne peuvent manquer de transformer le rôle des bibliothèques. Comme nous venons de le voir, les propriétés des bibliothèques (organisation des connaissances, exhaustivité et neutralité) tendent à être partagées par d'autres entités, notamment les infrastructures, mais également, dans l'objectif du Dublin Core, le web dans son entier.

Nous envisageons les mutations des bibliothèques autour de trois points essentiels :

- La structuration des connaissances
- La capacité grâce au MARC 21 de constituer un réseau universel d'échange de données et de relier les structures de données grâce aux protocoles FRBR.
- Enfin, du fait de l'hébergement des documents tout autant que la capacité des grilles de faire circuler des quantités d'information plus importantes, on peut considérer que les bibliothèques sont les plus à même pour structurer l'entrée dans les contenus.

Nous entendons par bibliothèque en premier lieu les institutions de type bibliothèques universitaires. Elles se distinguent par le fait qu'elles pratiquent l'échange électronique de données (sous forme de notices bibliographiques) depuis l'apparition du MARC en 1961, qu'elles sont donc constituées en réseau et que par ailleurs leurs collections sont structurées à l'aide de classifications standardisées. Cette maîtrise des réseaux associée au dépôt permet d'envisager une dimension nouvelle des bibliothèques, dans l'hypothèse d'une entrée dans les contenus. En effet, les dispositifs d'e-science, tels qu'ils sont considérés aujourd'hui, ne prennent pas en compte les publications ; en ce sens, notre projet est également novateur.

La structuration des connaissances.

Il existe un certain nombre de modèles des bibliothèques, évoluant au fur et à mesure de la mutation du document (depuis les papyrus et tablettes de l'Antiquité) et répondant à l'exigence de l'optimisation et de la maximalisation de l'utilisation sociale de la connaissance enregistrée (Herold, op. cit.). Or justement, les perspectives de l'e-science transforment considérablement la demande sociale concernant la connaissance déposée. A terme, elles devraient permettre d'inscrire les connaissances déposées à l'intérieur des processus de travail. En ce sens, le lien entre un travail de recherche et les connaissances déposées est renforcé et de plus en plus approchant le temps réel.

De façon toujours très générale, pour Herold, les bibliothèques peuvent être considérées au travers de trois niveaux de représentation : les données, l'information et les structures d'information, et enfin les connaissances et les structures de connaissances. Ces définitions relativement générales reposent systématiquement sur trois objets qu'il est important de distinguer : les documents, l'information sur le document et les réseaux. Les bibliothèques peuvent être définies globalement, d'après Herold comme un ensemble d'unités agrégées connectant des unités distinctes de connaissances au travers d'objets d'information que sont les textes et les documents. La médiation de l'ensemble serait assurée par des processus d'augmentation de la connaissance convergeant vers l'utilisateur informé. Ces processus sont ceux de la recherche d'information (qui dans ce cadre peut être considérée comme un apprentissage). L'utilisateur reprend et s'approprie ce système des échanges d'information, qui constitue une dimension essentielle des bibliothèques. Les flux s'inscrivent dans cette dynamique en proposant d'accentuer encore les mises en relations de document par la spécification de relations qui ne sont pas inscrites dans les descriptions et classifications usuelles. Du point de vue des bibliothèques, on propose une explicitation des relations. Si les flux permettent de mettre en relation des données hétérogènes, ces mises en relation, au travers du web service, produisent des connaissances.

Enfin, tout échange présuppose une donnée enregistrée, donc un composant temporel. Cela est particulièrement valable pour toute entreprise de recherche d'information : la composante temporelle est toujours associée aux documents enregistrés, et toute connaissance enregistrée peut alors devenir une connaissance transmise. En ce sens, les connaissances produites sont dynamiques et évolutives. Elles ne sont pas sans rapport avec le lien opéré entre l'e-science et le workflow.

Tout notre travail se situe entre les deux derniers niveaux, information et connaissances. Il insiste donc sur la façon dont les contraintes des réseaux peuvent effectivement construire des structures à la fois logiques et linguistiques (au travers de la structure d'information comme représentation du résultat d'un flux). Ces structures indiquent quelle information est transmise et interprétée, et donc les conditions pour qu'une information soit transmise à propos d'un document. Cela concerne donc la structure d'information, définie comme le type de représentation permettant de transmettre un contenu depuis un espace vers un autre. Plus concrètement, la structure d'information permet de décrire un document au travers de la description d'autres, avec lesquels elle est en relation. En ce sens, la structure d'information permet de lever d'enrichir l'information représentée à propos d'un document parce que la structure d'information inclut la connaissance du contexte du document. Plus précisément, la structure permet de caractériser le contexte de description d'un document, et donc explicite la signification : la structure permet de lier plusieurs descriptions de ce document, l'une étant la précision d'une autre.

Dans ce cadre encore relativement théorique et général, on peut aisément entrevoir le rôle des flux par rapport aux classifications ; la complémentarité réside d'une part dans la dynamique inférentielle, fondée sur la causalité et marquée par l'hétérogénéité des entités mises en relation, proposée par les flux, et d'autre part les relations structurelles, fondées elles sur la proximité maximale des entités, proposée par les classifications et autres langages documentaires. Dans le projet, l'articulation se fait entre une ontologie et les flux.

Le point de vue relativement général adopté se justifie par le fait que l'on ne caractérise pas un type particulier de bibliothèque, mais le concept même. En effet, le rôle de bibliothèque peut être joué par des organisations différentes, sachant qu'aujourd'hui, entre les bibliothèques elles-mêmes, les bibliothèques numériques et les infrastructures, de nombreux acteurs se partagent ce rôle. De plus, ils ne peuvent l'exercer indépendamment des consortiums et autres institutions transversales comme le Dublin Core.

Les bibliothèques numériques ont montré que les dépôts de documents et de publications pouvaient être organisés et gérés par des organisations directement liées aux acteurs et institutions de la recherche. Les dépôts de ressources, outils et données de l'e-science, sont également élaborés par des institutions appartenant au même domaine. En ce sens, le rôle de bibliothèque est joué aujourd'hui par des acteurs qui ne sont pas à l'origine des bibliothécaires mais des producteurs de connaissance. Cette montée en puissance ne devrait que croître avec le développement de l'e-science.

Cette maîtrise de la diffusion et de la valorisation des résultats de la recherche (depuis les bibliothèques numériques jusqu'aux perspective d'e-science) par les instituions la gouvernant nous a permis d'opérer certains choix en matière de partenariat. Si en France la situation est complexe du fait de la multiplicité des acteurs (les différentes bibliothèques numériques, le CIMES, le TGE-ADONIS), l'exemple allemand, où D-SPIN centralise à la fois les dépôts de publication et les données de la science (ce qui n'exclut pas des initiatives plus disciplinaires

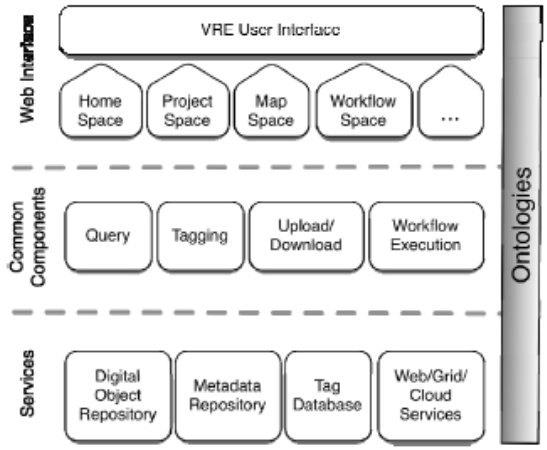
comme DBLP), montre que les questions d'e-science et de bibliothèques numériques sont considérées comme des éléments stratégiques de la politique de recherche. L'avantage d'une telle perspective est bien le fait que les publications et les ressources sont stockées dans de mêmes structures ou des structures apparentées. A cette question, on doit ajouter la complexité disciplinaire, à savoir les différences de pratiques scientifiques entre les disciplines.

Ce constat nous a permis d'orienter nos partenariats vers les acteurs de ces infrastructures et des bibliothèques numériques, le CCSD et le TGE-ADONIS. Ce sont eux qui au niveau français sont les plus à même de développer les perspectives d'e-science et d'effectuer une mise en relation avec les infrastructures européennes.

La maîtrise des réseaux d'échange.
Qu'il s'agisse des projets DARIAH ou CLARIN en Europe, l'objectif consiste à fédérer des ressources isolées et de permettre le déploiement d'outils utilisant simultanément ces ressources hétérogènes. L'utilisation d'une telle masse informationnelle conjuguée aux capacités des réseaux entraîne une redéfinition des outils de description des documents et de la façon de rendre public leur contenu.

La maîtrise des échanges par le biais des infrastructures constitue un enjeu avant tout pour la structuration de la recherche et sa valorisation. En ce sens, les institutions en charge de cette recherche sont avant tout nationales et proposent des assemblages relativement différents.
Il peut apparaître contradictoire de développer d'un côté des projets nationaux plus ou moins coordonnés avec les politiques européennes et d'un autre de promouvoir l'interopérabilité et l'échange généralisé des données. La publicité des données et des productions est une façon de construire une visibilité et donc de prendre position à l'échelle internationale. Il s'ensuit que la présence sur le web constitue un impératif pour la mise en réseau du travail scientifique. En effet, et c'est là notamment un objectif du projet britannique, la fédération des produits et des outils permet d'accroître à la fois la qualité et la quantité des résultats et de diminuer les coûts. Ce point de vue est par exemple représenté par les promoteurs de l'e-science en GB : « As scientific research becomes increasingly interdisciplinary in nature, the need for technologies that support collaboration and provide access to heterogeneous data and computational resources becomes ever more critical »(p.1)[258].

Ainsi, les réseaux se développent à plusieurs niveaux. Si l'accès public concerne les résultats et les données, ils sont considérés comme des préalables pour l'élaboration de projets pour des acteurs plus restreints. Ce sont les environnements virtuels de recherche. Ce qui est important de voir, c'est que ces deux niveaux sont complémentaires et que la dimension de diffusion s'accompagne de la collaboration. Ainsi, les politiques d'open data, dans lesquelles s'insèrent les questions d'e-science, modifient les dimensions visibles de la science en rendant publics les données et les outils. Les laboratoires se trouvent dépossédés de leur matériau tout autant que cette dépossession garantit la scientificité des publications. La conséquence est donc une obligation à travailler sur de nouvelles constructions de données, et donc à innover.
On insistera sur la continuité entre ces différents niveaux, à savoir le fait qu'entre l'espace collaboratif communautaire et l'espace de la diffusion publique des résultats, on a affaire à une continuité, illustrée par exemple par le VRE « ourSpace » (op.cit., p.2) :



Ce schéma montre exactement les relations qui peuvent être établies dans le cadre d'un dispositif, entre le chercheur et son interface d'une part, et le web en général d'autre part. Il montre également comment les différents outils et objets se positionnent entre l'espace de l'utilisateur et le web, avec en intermédiaire le système collaboratif.

Ce sont donc de mêmes infrastructures qui assurent le travail de recherche et la mise à disposition sur le web. Si effectivement le travail demeure inscrit dans des espaces nationaux, les échanges de données sont par contre universels.

Ces espaces, dans leur mise en œuvre, varient considérablement en premier lieu au niveau du développement national des infrastructures. Ces variations concernent essentiellement les espaces de travail.

Notre projet s'inscrit dans ce développement : en effet, les relations entre données hétérogènes que l'on propose ne doivent pas demeurer dans le cadre invisible des métadonnées, mais se traduire dans les espaces de travail. Cette perspective amène notre travail à se situer dans le double cadre des spécifications de relations entre données et dans celui des interfaces aux utilisateurs (y compris dans leur capacité à engendrer la production de ces métadonnées). Les mises en relation que l'on construit peuvent être pertinentes à la fois dans le cadre d'une navigation que dans la structuration des ressources et des publications.

La question des contenus.
En effet, à l'heure où les métadonnées sont inscrites dans le document (et non externes), il devient plus facile de penser une analyse des contenus pertinente pour la description. Cet état n'a pas à être considéré relativement au document seul, mais au niveau des collections ou même de l'offre globale.

Il est également prévisible que les microdonnées d'une part, le développement de la TEI d'autre part, vont amener à réviser la distinction entre la description documentaire et les contenus. En effet, ces deux outils différents quant à leur usage ont en commun de décrire des contenus. On peut par ailleurs estimer que les modèles d'extraction d'information deviennent de plus en plus précis. Mais surtout, ce sont les capacités des grilles qui constituent l'avancée la plus importante : l'augmentation des capacités de traitement permettra alors d'analyser les contenus, sans que ceux-ci aient été préalablement annotés.

6.1.6. Réseaux et description des documents ; des métadonnées à la navigation.

Les usages du dispositif sont liées à la recherche d'information, mais également à la navigation. La navigation centre la question des choix sur l'usager, et donc transforme les perspectives d'automatisation de raisonnements d'information sur des documents. Si l'on reprend le modèle défini précédemment de structure d'e-science, l'espace de travail est en continuité de celui de traitement de l'information. Ainsi, les questions relatives aux relations entre publications et ressources doivent être considérées au travers de plusieurs niveaux d'usages. Enfin, ces différents niveaux permettront de comprendre la dynamique du système, à savoir comment fonctionne l'acquisition des relations.

Préalablement, il faut pour cela caractériser ce que l'on entend par navigation. A la suite des travaux sur l'apprentissage, on peut caractériser la navigation au travers d'un contexte : "Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves"[259].
On distingue donc deux niveaux dans l'organisation du travail. Le premier niveau traite de la relation entre données, et en premier lieu entre métadonnées, le second permet de maîtriser la trajectoire de l'utilisateur.

On n'envisage pas la navigation à partir d'une tâche précise (comme par exemple récupérer un cours) mais bien une activité complexe impliquant l'interaction entre le contexte d'une ressource utilisée et l'attention de l'utilisateur. Dans l'activité scientifique, on distingue fondamentalement l'acquisition de bibliographie de celle de ressources et d'outils pour la menée d'un projet scientifique. On peut exprimer cette distinction de façon différente : on aurait d'un côté la construction d'une compétence, et d'un autre la formulation, le développement et la résolution de problèmes liés à la menée d'une activité de recherche. La navigation est bien plus pertinente dans le second cas que dans le premier.

Pour bien comprendre cette question, il faut revenir sur les questions de l'e-science. Si l'on dépasse le simple cadre de la mise en ligne de données et d'outils, l'e-science transforme considérablement les rapports que l'on peut avoir avec la recherche d'information, le travail en ligne et le travail collaboratif. En effet, l'e-science crée une mise en commun des données et du travail sur celles-ci qui n'a guère d'équivalent.
Elle est donc en lien à la question des communautés, et plus particulièrement des communautés d'utilisateurs de ressources. Il est non seulement important de connaître qui utilise telles ou telles données, mais également quel lien est créé par l'utilisation de ces données.
Qu'est-ce que l'on produit alors comme usage en mettant en relation des données hétérogènes et en créant du sens entre elles ?
Zinn, Wittenburg, et Ringersma (op. cit.) présentent un environnement de recherche permettant à l'utilisateur de gérer (annoter, commenter) ses ressources d'e-science. Le cadre présenté est limité à la gestion des ressources et ne contient pas d'indication de navigation, et surtout de raisonnement : il s'agit surtout de construire un espace de gestion de ses données (y compris en donnant la possibilité d'une dimension collaborative). Cette perspective est articulée à l'élaboration d'un workflow pour les activités de recherche.

La question de l'activité revient dès lors qu'il est question de navigation dans un cadre d'e-science. L'importance des travaux de G. Hutchins peut s'avérer essentielle parce qu'elle concilie raisonnement, distribution et information dans le cadre d'une activité finalisée. Elle intègre également la représentation de contenus de documents dans le cadre de la navigation, notamment celle qui concerne les relations que tout objet informationnel entretient avec d'autres ressources dans le cadre d'une activité.
On envisagera cette question autour de trois niveaux de pertinence :

- Le niveau des documents où l'activité est associée à des relations entre entités symboliques (généralement normées).
- Le niveau des outils, où il est fonction de relations fonctionnelles,
- Au niveau de l'activité, où il est question d'usage.

Cette question de la navigation permet également de préciser notre apport dans le cadre de l'e-science. En effet, on ne propose pas un outil d'e-science relativement complet, au sens de David de Roure et des perspectives de workflow, mais un service documentaire adapté à une perspective d'e-science.

Il existe une dimension essentielle à l'e-science : c'est sa connexion aux techniques du workflow. La perspective permet d'articuler les données primaires de la science avec les techniques de travail collaboratif. L'intérêt du travail de Carole Goble est qu'il associe la caractérisation des données primaires au modèle du travail collaboratif, mais également qu'il associe cette caractérisation au Dublin Core, et donc à une description universelle des objets. La seule perspective qui ne soit pas envisagée encore concerne les publications.

La perspective workflow proposée par Carole Goble est fondée sur l'idée que si les Linked Data proposent des outils pour naviguer entre des données différentes, ils ne proposent pas d'outils permettant de caractériser comment les acteurs travaillent ensemble. La distinction entre infrastructure proposant des services structurés pour une finalité explicite et plateforme proposant différents services sans spécification de l'usage se trouve ici totalement illustrée. D. De Roure et C. Goble[260] proposent de mettre en ligne les communautés scientifiques, ce qui consiste à dupliquer le fonctionnement d'une communauté de chercheurs sur le web.

La perspective d'une telle structuration de la recherche amène à situer la navigation non pas dans une perspective individuelle mais dans le fonctionnement collectif des équipes de recherche. En d'autres termes dans la perspective d'un usager chercheur d'information qui ne serait plus un individu mais un collectif structuré par un projet.
Comme le montrent les travaux de l'équipe de Carole Goble[261] les configurations des outils d'e-science sont encore largement à venir. Cette question croise des dimensions institutionnelles, sociales et communautaires qui sont loin d'être synthétisées, et sur lesquelles il est difficile d'avoir un point de vue unique. Ces questions concernent à la fois les pratiques scientifiques et les modes de fonctionnement des communautés scientifiques, mais également les usages des médias sociaux. La question des pratiques des chercheurs articule des questions à la fois professionnelles comme l'information, la menée de projets, et des questions plus générales, comme l'utilisation des médias sociaux dans le cadre professionnel.
Néanmoins, il n'est question ici que d'outils et de services. L'objectif de l'e-science consiste (et c'est ce qui le différencie des plateformes de mise à disposition d'outils), c'est bien de proposer des inférences. Qu'est-ce que doivent représenter ces inférences et pourquoi ? Une autre question concerne la mémorisation des usages, à savoir de quelle façon la mémoire des consultations, des vérifications, comparaisons, peut être exploitée. D. De Roure et C. Goble se situent dans une perspective altruiste, à savoir le quels que soient les risques, le partage est

plus profitable parce qu'il permet une plus rapide et importante diffusion des idées. La diffusion de méthodes, données, expériences, hypothèses et résultats préliminaires contribuent à la légitimation d'un travail scientifique, au même titre que des publications achevées.

La navigation telle qu'on la propose est largement tributaire des données à disposition. Dans le cadre du projet, elle ne concerne qu'une petite partie de l'e-science, puisqu'il ne s'agit que des liens entre des publications et leurs ressources. Elle ne concerne pas l'expérience même de la recherche, les commentaires qu'elle peut susciter, les questions, les liens entre chercheurs. Cette question renvoie aux développements de l'e-science, sur lesquels nous reviendrons.

En définitive, si notre projet s'inscrit dans le cadre de l'e-science, il possède une dimension documentaire fondamentale. Les relations que nous capturons sont fondées sur des usages, possèdent un ancrage historique (temporel) et s'appuient sur des descriptions. Il n'en est pas de même dans de nombreux projets d'e-science, qui sont marqués par les problématiques du workflow et des réseaux sociaux.

6.2. Complémentarités d'approches culturelles et technologiques. L'orientation usager.

Comment les usages et les usagers s'intègrent dans cette perspective ?

Comme nous l'avons vu, il existe différentes façons de définir les usages. Nous avons essentiellement privilégié une définition par laquelle l'usage caractérise les différentes façons par lesquelles on peut utiliser une entité (une unité de la langue, un outil). L'usage se définit comme la façon dont les communautés humaines mettent en œuvre les potentialités de la langue ou d'un outil, voire d'un dispositif. On postule que l'ensemble des possibilités n'est pas répertorié. L'usage des outils peut être défini comme l'étude de l'ensemble des potentialités de leur utilisation. L'outil étant défini comme un ensemble de fonctionnalités issues d'un processus d'externalisation d'une compétence cognitive, l'usage d'un outil se définit par l'inscription de ces fonctionnalités dans un contexte d'activité. L'analyse des usages sert à définir à quoi sert cet outil dans un contexte défini. Cette approche est inverse et complémentaire de celle qui part des usages attestés des outils dans les communautés humaines et qui cherche à comprendre les connaissances humaines qui sont déposées dans ces outils.

Cette distinction est importante parce qu'elle va permettre de définir la place de l'utilisateur dans l'ensemble du projet. La façon dont les usagers s'approprient les usages des outils constitue une question complémentaire. L'orientation usager se manifeste de façon aujourd'hui assez classique dans la construction de profils utilisateurs et la spécification de vocabulaires. Or la question est un peu plus compliquée : il ne s'agit pas seulement de vocabulaire, mais d'activité et de menée de projets collectifs.

Les ontologies et les métadonnées constituent des outils, insérés dans des dispositifs. En ce sens, la question des usages distingue l'usage de dispositifs comme par exemple une bibliothèque virtuelle, de celle de l'ontologie qui la constitue. Il s'agit alors de façon d'appréhender la diversité des usages possibles des outils, autrement dit d'interpréter une ontologie. Il s'agit alors de distinctions d'usage d'outils.

Nous commencerons donc par la question des usages et des usagers au niveau le plus général, avant de nous pencher sur la façon dont cette question se traduit au niveau plus fin des questions de langage.

Le problème a été déjà posé par M. Schorlemmer & Y. Kalfloglou[262] : « We believe that in order to address the semantic heterogeneity problem in all its complexity, we need to take into account locality and difference not only at the terminological level, but also at the level of the interpretation mechanisms of a community, and the actual scope and use of ontologies by means of particular communities » (, p. 133).

La dimension terminologique est considérée comme le niveau par lequel il est aisé de distinguer des spécificités communautaires : en effet, l'usage de la langue étant différent selon les communautés, le niveau apparaît comme celui dans lequel la distinction entre les communautés est la plus aisée à représenter. Simplement, il ne s'agit ici que de distinctions communautaires d'usage de la langue.

Ces deux perspectives sont corrélées dans la mesure où ce sont des indicateurs de contenu qui sont interprétés. On peut revenir sur la façon dont les profils utilisateurs sont aujourd'hui définis, notamment au sein des CMDI et du DC. Enfin, cette spécificité interprétative est positionnée non seulement dans l'usage effectif de l'outil (au sens de l'outil de navigation par exemple) mais dans la structuration des métadonnées. C'est pour cela qu'une labellisation communautaire des outils est proposée, visant à traduire dans le cadre des usages des outils les spécificités communautaires de vocabulaire.

Comme on l'a vu, les usages peuvent être décomposés de la façon suivante :

- La mise en relation de documents et de ressources.
- Le renseignement de documents par d'autres, autrement dit l'accroissement de la quantité informationnelle associée à chaque document.
- Le fait que l'on propose un outil de navigation.

6.2.1. Usage de dispositifs scientifiques.

La question des usages peut concerner des niveaux très différents des outils techniques. Nous présentons dans un premier temps globalement l'usage de notre système, avant de détailler les différents aspects énoncés plus haut.

Notre outil est destiné à des chercheurs en cours de recherche d'information sur des publications et sur des ressources de type outils et données primaires. Cet outil s'adresse à des chercheurs ayant relativement fréquemment à actualiser leurs connaissances et à faire des choix d'outils et de corpus. Sont concernés les chercheurs, enseignants chercheurs et thésards en linguistique et disciplines proches (TAL notamment).

Les outils de recherche permettent de produire des panoramas des publications, mais pas de leur contexte. Par ailleurs, ces outils servent à construire des bibliographies, mais pas à mener une recherche, avec les implications que cela a en matière d'outils et de données. Par conséquent, pour adapter l'offre informationnelle à la demande des chercheurs, dont le travail passe de plus en plus par des réponses à des appels d'offres, à des travaux pluridisciplinaires, il est opportun de leur proposer un outil adapté de navigation, servant à la fois à proposer une

recherche (répondre à un appel d'offres), à mener cette recherche (en proposant les outils et leurs usages), et à argumenter des résultats.
Le modèle permet de lier les publications et les données/outils qui ont servi à obtenir les résultats publiés et surtout de décrire l'un par l'autre. Pour une publication, cela permet de savoir ce qu'elle a utilisé, pour une donnée primaire, pour quoi elle a été utilisée. C'est une façon de construire la représentation d'un domaine de recherche sur l'ensemble des objets numériques que ce domaine utilise.
Il faut donc proposer un outil de navigation qui permet de visualiser non seulement des titres, mais également des descriptions de façon à ce que le chercheur puisse faire ses choix sur des contenus. Ce peuvent être des choix de lecture, d'utilisation, ou de précision pour une autre recherche d'information.

L'outil doit permettre de visualiser le contexte (données primaires, outils d'analyse) d'une publication (ou inversement, les publications liées à un outil). Il doit aussi dans un second temps relier des publications différentes (qui ne seraient pas systématiquement proches en utilisant une recherche d'information classique).

On propose de distinguer trois types d'usages relativement aux métadonnées proposées :
- dans un premier temps, il s'agirait de lui permettre d'opérer des choix de lecture.
- Dans un second temps, il s'agirait de proposer un outil de navigation complémentaire d'une recherche d'information et une structuration des résultats de cette recherche.
- Dans un troisième temps, il pourrait s'agir de structurer une recherche d'information.

Ces trois types d'usage sont bien évidemment dépendants de la généralisation progressive du modèle.
Dans un premier temps, on considère une recherche d'information utilisant n'importe quel outil, généraliste ou spécialisé. Une publication sélectionnée par l'usager (considérée comme potentiellement pertinente), pourra être soumise à l'analyse.
Dans un second temps, dès lors que suffisamment de documents auront pu être décrits (et le schéma pris en compte par différents moteurs de recherche), il sera possible d'utiliser ces métadonnées pour structurer une navigation entre différents types de données, et donc produire des résultats associant des documents (publications, données primaires, résultats intermédiaires) par usage mutuel.
Dans un troisième temps, nous pourrons effectivement proposer de construire un outil de recherche d'information fondé sur des proximités d'usage, de méthodologie et de cadre de référence.

6.2.1.1. Usage des dispositifs liés aux infrastructures.

On peut envisager une approche relative aux outils les plus généraux, comme ceux qui sont proposés par les infrastructures. On peut également considérer une approche fondée sur le fait que ces infrastructures proposent des outils très différents des bibliothèques traditionnelles, puisqu'ils sont les résultats de mises en relation de ressources hétérogènes.
A ce titre, on peut considérer que l'usage d'EUROPEANA ou d'ISIDORE, par exemple, constituent un sujet d'étude pertinent parce qu'il permet d'envisager les usages d'assemblages de ressources relativement hétérogènes.

Ces projets visent à assembler différentes ressources de façon à assurer une grande exhaustivité au niveau des résultats présentés. Par contre, les ressources, si elles sont très hétérogènes, notamment en ce qui concerne les thésaurus comme dans le cas de

l'infrastructure EUROPEANA (qui articule différents thésaurus) ou encore la structure des collections (et leur volume) comme pour la plateforme ISIDORE, il n'en reste pas moins que l'usage est la recherche bibliographique.

Dans ces deux exemples, il ne s'agit pas de mettre en valeur cette hétérogénéité mais au contraire de la faire disparaître grâce aux outils du web de données. L'hétérogénéité gérée par ces deux outils est bien plus liée à la diversité des structurations (thésaurus, classifications) qu'au contenu des documents numériques eux-mêmes. Notre projet est totalement différent puisqu'il s'agit au contraire de mettre en valeur cette hétérogénéité et de donner à explorer, pour l'utilisateur, la diversité des ressources. Par ailleurs, les outils précédemment évoqués reposent sur une fonction unique, qui est de proposer des résultats pour une recherche d'information. Or, on peut envisager la question différemment : plutôt que de présenter des publications à lire, on peut proposer des ressources hétérogènes en explicitant leurs liens.

Néanmoins, notre projet s'inscrit dans la continuité du développement de ces bibliothèques numériques parce qu'il s'inscrit dans le cadre de l'accès à l'information et qu'il utilise les mêmes outils que celles-ci.

Dans le cadre du TGE-Adonis, les niches proposées ne vont pas au-delà du dépôt de données de la recherche. En ce sens, on est encore loin de la construction d'un dispositif à usage mixte, alliant la recherche bibliographique et l'e-science. C'est la raison pour laquelle la proposition d'une interface ne constitue pas le seul mode de fonctionnement de notre outil. On ne dispose pas d'une infrastructure qui intégrerait cette double dimension pour envisager seulement l'usage lié à l'interface.

6.2.1.2. Usage des mises en relation de données hétérogènes.

L'objectif des outils que l'on propose consiste à structurer des ressources hétérogènes de façon à ce que l'utilisateur puisse avoir accès de façon unifiée à l'ensemble des ressources du domaine de recherche. Ainsi, on ne vise pas à associer des données hétérogènes dans un cadre unique mais au contraire à mettre à profit cette hétérogénéité et de spécifier les liens entre les deux.

Comme on a pu le voir, les bibliothèques numériques et les répertoires de ressources constituent deux univers séparés. Cela s'explique entre autre parce que la mise à disposition des ressources est encore en cours de structuration. Par exemple, dans le cadre français, l'organisation de l'assemblage des corpus de linguistique est en cours[240]. L'infrastructure IR-CORPUS est en cours d'intégration dans le TGE-ADONIS.

Par contre, si l'on regarde maintenant du côté des infrastructures européennes, CLARIN et DARIAH, on a affaire à des offres de services à destination des professionnels de l'information. En effet, dans CLARIN, les outils comme ISO-cat (collection de ressources lexicales pour les données linguistiques), ou encore le VLO (observatoire virtuel des langues) sont à destination des professionnels (chercheurs et ingénieurs) ayant à travailler à la mise en

[240] http://www.corpus-ir.fr/uploads/CorpusIRRapportScientifique2012.pdf

valeur des données et documents concernant les ressources linguistiques. Les outils proposés par DARIAH ont exactement le même usage, même s'ils englobent des thématiques plus larges (comme par exemple la mise en valeur des outils d'analyse des productions médiatiques[241], ou encore le réseau CESSDA[242], qui concerne les données et Sciences Sociales, dont le réseau QUATERLET est partenaire[243]).

Les objectifs sont avant tout le partage des outils de description et d'analyse des données de façon à améliorer leur visibilité sur le web. Cette visibilité du fait d'une description plus précise (notamment au travers de l'intégration de microdonnées (projetDWB[244])) permet de constituer les données fondatrices pour les structures d'e-science.

L'étape actuelle est celle de la mise en valeur des données. Au-delà des données et de leur visibilité, on peut entrevoir une réorganisation de la recherche et de son fonctionnement conforme aux objectifs de l'e-science. Le partage des données crée les conditions pour une réorganisation communautaire.

La conséquence d'un tel effort est l'amélioration de la description des données par la progressive standardisation des vocabulaires et des outils. L'usage de ces outils est pour le moment interne aux communautés du traitement des documents (par exemple, en France, le consortium Corpus écrits[245]). Néanmoins, ces efforts donnent une visibilité aux équipes de chercheurs qui leur permet de valoriser leur travail autrement que par de seules publications.

Nous venons de présenter l'usage actuel des outils de description des ressources publiés par les infrastructures. L'usage des ressources est le fait des équipes de chercheurs, et il n'existe pas de communauté autre qui soit intéressée par ce type d'objet.

C'est justement ce faible usage des ressources que l'on cherche à contrecarrer en proposant un autre usage à ces descriptions : les publications sont lues et utilisées par un public et pour des activités plus larges que le seul cadre des processus de recherche. Les démarches pédagogiques et de formation, qui requièrent à la fois des résultats publiés et des méthodes, sont candidates à l'utilisation des ressources. Dans le cadre des SHS, du fait d'un public relativement restreint tout autant que de la persistance d'un travail individuel, il est difficile d'envisager un usage qui serait uniquement de la recherche, comme dans le cadre des sciences du vivant.

Enfin, les SHS, notamment lorsqu'elles impliquent différentes disciplines et méthodologies (comme par exemple l'informatique et la logique), peuvent présenter sous de mêmes appellations des travaux très différents : la thématique de la structure d'information en est un bon exemple. C'est donc dans ce cadre que l'on a besoin de construire une description plus précise des documents.

[241] http://dighumlab.dk/research-themes/media-tools/
[242] http://www.cessda.org/about/research/
[243] http://www.reseau-quetelet.cnrs.fr/spip/?lang=en
[244] http://www.dwbproject.org/access/call.html
[245] http://corpusecrits.corpus-ir.fr/

6.2.1.3. Renseignement de documents par d'autres : accroissement de la quantité informationnelle.

Jusqu'à présent, la recherche de ressources et celle de documents sont considérées comme totalement distinctes. Les ressources et outils d'une part et les publications d'une autre ne font pas l'objet d'un même usage dans le cadre de la recherche et de l'enseignement. Cette distinction s'explique donc pleinement. Néanmoins, comme on l'a vu, l'e-science offre la possibilité de rendre public l'ensemble des données et des outils utilisés et utilisables. Par conséquent, une publication n'est plus la seule partie visible d'une recherche. Une part essentielle du travail antérieur devient visible et disponible. Cet état de fait transforme considérablement la façon dont on peut appréhender une recherche.

Cette transformation a quelques conséquences sur la recherche d'information. L'effort porte sur la pertinence et le retour. Or, en considérant les ressources utilisées, on peut envisager une démarche par laquelle ce que l'on sait des ressources utilisées structure les résultats de la recherche.

Le problème est que la description des publications relève de standards comme de Dublin Core qui ne laissent guère de place aux spécificités d'usage. A l'opposé, les ressources sont disséminées sur des serveurs distincts et utilisent –éventuellement- des métadonnées spécifiques. L'articulation de ces deux types de métadonnées permet alors de produire un descriptif plus étendu de chacun mais également, par le lien produit, de constituer un objet de recherche au sens de Carole Goble.

On peut critiquer cette position en pointant sur le bruit généré par cet accroissement des descripteurs. Or, les éléments que nous proposons sont relativement peu nombreux, les attributs et les valeurs sont déjà répertoriés soit dans des métadonnées reconnues ou dans les lexiques standardisés.

Par ailleurs, comme on l'a déjà remarqué, la question des contenus est problématique dans les jeux de métadonnées, tout d'abord parce qu'il n'existe pas de critère explicite pour spécifier un contenu. Il reste à l'appréciation de la personne chargée de l'indexation. Comme dans le cadre des bibliothèques et des centres de documentation, les jeux de métadonnées recourent aux vocabulaires contrôlés. Cette solution n'est pas tout-à-fait satisfaisante si l'on considère à quoi sert le renseignement relatif aux contenus. Nous considérons que les contenus servent d'abord à donner des indications sur le cadre de la publication décrite. En d'autres termes, ils servent à construire des indices sur l'objet de recherche. Ce qui est intéressant dans l'approche proposée, c'est le fait de pouvoir expliciter le cadre de travail dans lequel une publication a été réalisée en utilisant uniquement les descripteurs des objets de ce cadre de travail. En d'autres termes, on crée très peu de métadonnées et seulement des relations entre les différents jeux de métadonnées, ceux qui décrivent les documents et ceux qui caractérisent les ressources.

6.2.1.3. Un enjeu : la gestion de la multiplicité des documents et leur spécialisation.

Pour bien comprendre l'usage que l'on cherche à définir, il faut revenir un peu sur la façon dont les résultats de recherche peuvent être aujourd'hui appréhendés. Les problèmes d'ordre méthodologique comme de couverture font qu'il est relativement difficile de disposer de

données incontestables. (L'étude de l'OST (Observatoire des Sciences et des Techniques[246] donne une idée des difficultés à appréhender ce problème). Néanmoins, on peut se fier aux données quantitatives proposées par l'OST et avoir une idée globale de cette évolution. Celle-ci est marquée par un accroissement continu de la quantité de recherche produite, due entre autre à la productivité des pays émergents et des nouvelles puissances économiques (Chine, Inde, Brésil notamment). A cet accroissement, il faut ajouter la mise en ligne des données de la recherche, et la structuration progressive de leur représentation. On peut encore présenter d'autres faits, comme ceux proposés par S. Pettier[263] : « Partly as a consequence of this ability to generate data on such an unprecedented scale, scientists are now publishing more widely and at a greater frequency than ever before: today, life scientists alone are generating more than two peer-reviewed papers every minute. The need for readers to keep up-to-date with, to search, retrieve, analyses, and understand content from this growing body of writing is more pressing than ever, but is increasingly difficult to achieve. Such is the scale of the problem that the doom-mongers have seized upon it, peppering scientific papers with portents of impending catastrophe: from floods and deluges of data, to surging oceans and tsunamis; from icebergs and avalanches of information, to earthquakes and explosions! From this impressive array of disaster metaphors, it isn't difficult to detect a mounting sense of panic."

Aucune de ces études ne prend en compte les ressources, ni n'est capable d'envisager la façon dont les publications sont utilisées, et par qui. Néanmoins, une telle quantité de données remet en cause la notion d'exhaustivité d'une recherche d'information. Et même si l'exhaustivité est atteinte, l'ensemble des ressources obtenues ne peut être traité. Par ailleurs, cette recherche est-elle pertinente par rapport au besoin du chercheur ?

Si l'on répond que le besoin est lié à l'activité de recherche, à un processus en cours, alors la question sera un peu différente : la recherche d'information vise à résoudre un certain problème dans le cours d'action d'une activité de recherche ou d'enseignement. Ainsi, on peut considérer que le besoin d'information possède une dimension pragmatique très prégnante.

On peut alors faire l'hypothèse que l'apport de connaissances ne peut se réduire à des connaissances déclarative, argumentées dans des publications, mais à l'ensemble de leur contexte, comment et avec quels outils ces connaissances ont été obtenues.

On peut considérer un autre usage : l'ensemble des publications concernant un thème est relativement inutilisable du fait de sa quantité. Les ressources et les outils utilisés constituent d'autres moyens pour structurer l'offre de connaissances. En effet, tout d'abord, en Sciences Humaines et Sociales, les résultats sont dépendants des méthodes et outils utilisés. La mise en valeur de ces derniers constitue une façon d'assembler des documents sur la base de leur objet de recherche et non plus sur leur seule thématique.

6.2.1.4. Interdisciplinarité et spécialisation.

Un autre fait n'est pas pris en compte, mais qui par ailleurs interroge à propos de la recherche d'information : les utilisateurs lisent-ils des documents dans leur propre discipline uniquement, ou au contraire, sur de mêmes thèmes mais dans des disciplines voisines, etc. ? A

[246] http://www.obs-ost.fr/sites/default/files/Etude_INFO_rapportfinal_VF.pdf

propos de ces pratiques quotidiennes de la recherche d'information scientifique, nous ne disposons que de peu de données. Simplement, les fonctionnalités des outils de recherche ne permettent pas de représenter des travaux de recherche dans le cadre des démarches qui ont permis ces publications. En ce sens, restituer le contexte constitue un enjeu non seulement pour caractériser certains contenus des publications, mais également pour cerner des objets de recherche.

Le cadre de travail de la recherche est de plus en plus marqué par l'interdisciplinarité et le travail sur projet. Or, la multiplication des publications devrait entraîner une spécialisation de plus en plus marquée des chercheurs : en effet, être compétent sur un thème demande de plus en plus de temps. Par conséquent, l'articulation entre ces deux logiques constitue une nécessité pour la menée de projets. Comment peut-on à la fois conserver une maîtrise de connaissances à propos d'un thème et travailler sur des projets collectifs impliquant des connaissances plus vastes et hétérogènes ?
Une stratégie idéale consisterait à identifier des projets similaires, des méthodes et des choix de travail collaboratif. Dans ce cadre, des informations sur les contextes des publications ne manqueraient pas d'être pertinentes. Si l'on parle effectivement de projets et de travaux collaboratifs (voir l'étude de l'OST sur la question), c'est à la fois du point de vue de la production et de celui des utilisateurs de l'information. En somme, un trait fondamental de la recherche demande à être pris en compte par les outils de recherche d'information. En associant à une publication des données à propos des ressources qu'elle a pu utiliser (mais on peut envisager aussi d'autres relations et d'autres structurations, comme par exemple les liens aux projets de recherches comme commence à le faire l'ontologie VIVO), il devient possible de représenter d'autres données que les seules publications. Cette perspective permet la construction d'ontologies bibliographiques, distinctes des ontologies de domaine parce qu'elles établissent des relations entre des outils de description de documents (ou de ressources). Nous avons déjà mentionné, outre VIVO, BIBO. La Bibliographic Ontology (BIBO) est par exemple un outil permettant de mettre en relation des éléments de description bibliographique de façon à naviguer entre ces ressources. BIBO utilise à la fois ses propres métadonnées et des outils déjà construits comme les propriétés et les classes du Dublin Core, des schémas RDF, OWL ou FOAF.
Il s'agit donc d'un outil qui permet à la fois d'améliorer la description d'un objet documentaire et de lier celle-ci à d'autres entités qui alors le renseignent.
Cet exemple montre de quelle façon les métadonnées, qui sont des descriptions statiques des documents, peuvent s'inscrire, à l'aide d'une ontologie, dans un cadre dynamique. A ce moment-là, on passe de questions de prédication à des questions plus large de navigation à l'aide d'une ontologie.
Par contre, cet exemple aussi montre que l'on raisonne toujours sur des objets homogènes. En effet, les relations entre les concepts sont marquées par un lien partitif ou de généricité. A l'opposé, un raisonnement fondé sur la causalité pourra être marqué par l'hétérogénéité. Les flux tels que nous les envisageons permettent de réviser la nature des relations entre entités dans le cadre des outils de description bibliographique.

Nous pouvons ajouter l'ensemble des outils structurés autour de SPAR[247]. SPAR constitue l'une des tentatives les plus abouties pour associer différents outils de description des publications.

[247] http://sempublishing.sourceforge.net/

Nous reviendrons plus loin sur SPAR, qui constitue le versant « documentaire » des propositions d'e-science. Ces ontologies sont écrites en utilisant OWL et possèdent un module d'interrogation utilisant SPARQL.
L'usage de ces outils est bien le même que celui de l'e-science, mais vu du côté de la recherche d'information. Il s'agit dans ce cadre de lier des ressources hétérogènes (par exemple par le biais de l'utilisation de wikis) et d'associer à des publications l'ensemble du contexte qui a permis leur élaboration et leur diffusion. Comme le présentent S. Perini & D. Shotton [264], "Semantic publishing is the use of Web and Semantic Web technologies to enhance the meaning of a published journal article, to facilitate its automated discovery, to enable its linking to semantically related articles, to provide access to data within the article in actionable form, and to facilitate integration of data between articles. Recently, semantic publishing has opened the possibility of a major step forward in the digital publishing world. For this to succeed, new semantic models and visualization tools are required to fully meet the specific needs of authors and publishers".

Ces propos entérinent le fait que les publications ne constituent plus les seules entités scientifiquement pertinentes pour présenter une recherche. Ce projet est bien plus large que le nôtre, entre autre parce qu'il n'est pas fondé sur un type de données spécifique comme les ressources mais intègre toutes sortes de données différentes (structurés par l'ontologie FaBiO, composant de SPAR[248]).

La méthodologie proposée par SPAR est fondée sur des ontologies crées de toute pièce, même si par ailleurs elles sont fondamentalement interopérables via le principe des LINKED DATA à l'ensemble des autres jeux de métadonnées. En somme, cet outil utilise relativement peu les outils existant, ce qui constitue une distinction majeure par rapport à notre propre projet. Nous nous appuyons en grande partie sur des structurations existantes en matière de description documentaire et sur une dimension communautaire. SPAR se fonde sur des outils généralistes (à commencer par WORDNET pour la caractérisation lexicale et DOLCE pour la structuration de l'ontologie).
La conséquence d'une telle proposition est une faible référence aux formats de description existant et aux vocabulaires de spécialité. Le problème est que l'on ajoute aux descriptions existantes un autre jeu (en l'occurrence plusieurs jeux). En dehors des questions de fonctionnement intrinsèque du modèle, on remarquera que la proposition de SPAR est fondée sur une sémiotique précise, qui est le fondement de l'écriture générale du modèle et surtout, qui distingue les ontologies produites à l'aide de ce modèle de l'ensemble des jeux de métadonnées et des ontologies existantes. Cela explique également le caractère radical des propos tenus sur l'évolution du web par S. Peroni & D. Shotton : « Yet, to date, publishers have not adopted Web standards for their work, but rather employ a variety of proprietary XML-based informational models and document type definitions (DTDs). While such independence was reasonable in the pre-web world of paper publishing, it now appears anachronistic, since publications and their metadata from different sources are incompatible, requiring handcrafted mappings to convert from one to another. For a large community such as publishers, this lack of standard definitions that could be adopted and reused across the entire industry represents losses in terms of money, time and effort."

[248] http://www.essepuntato.it/lode/http://purl.org/spar/fabio

Une telle transformation de la présentation ne concerne plus seulement les données proposées à la lecture, mais la façon de présenter les résultats d'une recherche d'information et les liens opérés entre les publications et tout ce qui y est relatif. Cette question renvoie aux données que l'on représente sur l'interface, entre autre pour éviter des confusions et une surcharge d'informations. Néanmoins, ce problème est considéré comme étant du ressort des plateformes.
Ainsi, une préoccupation centrale des éditeurs d'ontologies bibliographiques est de s'assurer d'une compatibilité avec les outils en place dans les plateformes. Néanmoins, c'est sur la question des choix de plateformes et/ou d'infrastructure que se lit la différence entre les acteurs mais au-delà les choix stratégiques et méthodologiques. Du point de vue des usages, on considère soit que

- l'offre d'e-science pourra se structurer autour de plateformes communautaires qui contiennent entre autre mais pas seulement des bibliothèques (beaucoup de wikis, d'outils et de ressources)
- l'offre d'e-science sera structurée autour des infrastructures, et s'appuiera sur les bibliothèques numériques et les ressources fédérées autour des infrastructures.

Les choix scientifiques articulés sur ces positions sont à l'avenant :

- dans le premier cas, on propose une refondation du web scientifique autour des formats RDF et OWL et de toute une série d'ontologies.
- Dans le second cas, on utilise les outils existant et on les aménage dans le but de construire une offre d'e-science fondée sur des outils interopérables.

Dans la première stratégie, on trouve les acteurs fédérés autour de SPAR, et un certain nombre de plateformes. Dans le second, les infrastructures et les principaux éditeurs de métadonnées et d'outils d'annotation.
Dans notre projet, puisqu'on on se sert de métadonnées existantes pour renseigner sur le contenu d'objets hétérogènes, on est obligé de prendre en compte l'existant. Par ailleurs, on ne voit pas très bien comment les formats des bibliothèques numériques pourraient disparaître, d'autant plus qu'ils sont soutenus par les grandes infrastructures et que d'autres solutions, comme celles mises en œuvre par EUROPEANA, permettent de gérer les relations entre des bases hétérogènes.

6.2.1.5. Usages et inférences.

Si l'usage de notre outil est la possibilité de lier de documents hétérogènes et d'enrichir des descriptions par d'autres, on se distingue par une importance forte accordée au raisonnement. Pour bien comprendre l'enjeu de cette partie, il faut situer les inférences que l'on propose par rapport à FRBR et SKOS. Ils constituent les deux outils d'écriture de relations entre structures de données hétérogènes existant.

FRBR (Functional Requirements for Bibliographic Records / Fonctionnalités requises des notices bibliographiques) est un modèle conceptuel de données bibliographiques (pour la version française, voir la présentation de la BNF[249]). Son intérêt est qu'il situe un objet documentaire dans ses différents contextes. Il s'agit donc d'un outil généraliste (qui s'applique autant aux archives, aux bibliothèques numériques ou matérielles, qu'aux objets de l'e-science). Il caractérise un document par ses différents contextes. Cet outil relationnel

[249] http://www.bnf.fr/documents/frbr_rapport_final.pdf).

permet de façon aisée de construire des objets de recherche, et c'est pour cela qu'il a été adopté par SPAR.

Les relations sont exprimées à l'aide de triples RDF[250] . Il existe aussi une version OWL des FRBR[251] . Ainsi, les mises en relations opérées par les attributs de FRBR peuvent être caractérisées comme des prédications.

FRBR est un langage d'expression permettant de décrire précisément les documents en relation à leur contexte. Néanmoins, il ne s'agit que d'un outil descriptif qui ne permet pas de caractériser l'influence d'un document sur un autre.

SKOS représente des inférences caractérisées entre des concepts mais en reprenant de façon systématique les relations dirigées proposées par le langage des thésaurus.

Ces relations nous apparaissent insuffisantes parce qu'elles ne permettent pas de caractériser le fait que certains descripteurs d'un document peuvent enrichir la description d'un autre. On a un raisonnement au sens où la conséquence est un apport informationnel et de connaissances. Les questions liées au web service et à l'apprentissage pourront être associées à cette dimension du raisonnement.

Les mises en relations de documents que l'on propose sont destinées à enrichir les descriptions de documents. En ce sens, les bibliothèques gardent un rôle fondamental dans notre projet. En effet, les projets come SPAR sont orientés exclusivement vers un usage de type e-science, en faisant en sorte d'inscrire les documents dans des unités plus larges d'activité : c'est en ce sens que l'on parle d'objet de recherche.

Par contre, un tel projet est silencieux sur la façon dont des documents similaires peuvent être structurés les uns par rapport aux autres. L'usage des ontologies FaBIo et CiTO (composant de SPAR) est orienté vers les plateformes intégrant des bases de données (Linked education, Virtual observatory and open citation corpus), vers des web services (CiteUlike, Wordpress) et sont intégrées dans des ontologies déjà existantes comme SWAN (domaine biomédical). (Voir S. Peroni & D. Shotton (op. cit.) pour plus de précisions). Les usages de ce type d'outil peuvent effectivement être relativement larges parce qu'ils sont intégrés à l'intérieur de plateformes et de services ayant leur propre usage.

L'objectif de SPAR est bien d'unifier des collections hétérogènes ou de promouvoir un service en y intégrant l'hétérogénéité des données. Il s'agit d'un outil permettant de structurer des relations entre des documents hétérogènes en fonction de leur parenté dans une activité. En ce sens, ils sont descriptifs du fonctionnement d'une recherche. Ils ne permettent pas de caractériser ce qu'apporte cette mise en relation tout simplement parce qu'elle n'est pas à grain suffisamment fin concernant le lien entre les deux documents.

[250] http://vocab.org/frbr/core.html
[251] http://speroni.web.cs.unibo.it/cgi-bin/lode/req.py?req=owlapi/http:/purl.org/spar/frbr

Le problème qui se pose pour l'ensemble d'outils SPAR est double :

- Pas de visibilité pour l'usager. Les relations qui sont établies sont destinées à être repérées par les moteurs de recherche.
- Pas d'apport pour la structuration des collections. Les bibliothèques numériques ne sont pas prises en compte dans cet assemblage.

Un dernier problème est d'ordre stratégique : si SPAR propose un ensemble d'outils cohérent et adapté à la représentation de données hétérogènes, par contre, en refusant de prendre en compte l'existant, il risque d'augmenter encore la complexité des outils du web, sachant que les acteurs qui maîtrisent le développement du web des bibliothèques et de l'IST défendent des projets totalement différents (BNF, EUROPEANA et les infrastructures).

En caractérisant une ontologie de domaine, plus précisément une ontologie fondée sur les descriptions des objets d'un domaine, on obtient une description précise des informations apportées par une ressource à propos de la publication qui l'utilise.

Cet apport peut être enregistré, ce qui permet d'enrichir une description et par conséquent peut servir dans le cadre de recherches d'informations qui n'impliqueraient pas ou pas seulement les objets de recherche. C'est également la raison pour laquelle la navigation est importante.

6.2.2. Description de documents et information sur les documents.

Nous interrogeons maintenant les pratiques en matière de description de documents. Comme nous l'avons suggéré plus haut, les descriptions de ressources (données primaires, outils) sont totalement dissociées des descriptions de publications. On n'utilise pas les mêmes métadonnées, les données ne sont pas hébergées dans les mêmes serveurs, ce ne sont pas les mêmes acteurs qui sont impliqués.
Le problème d'usage est lié à l'importance que l'on accorde au document. Le développement précédent concernait l'usage des relations entre données hétérogènes. Maintenant, nous développons les usages de la description de documents dans le contexte précédemment évoqué.

Comme nous l'avons déjà évoqué, notre projet porte sur un usage, qui est celui des données hétérogènes, présentes sur le web, relatives à l'ensemble des données, outils et publications de la recherche. Arriver à un tel résultat ne peut être facilement acquis. Cela implique une caractérisation de la place du document dans le cadre des échanges avec l'utilisateur.

A la différence de FOAF ou de VIVO, nous n'avons pas simplement affaire à des liens entre données mais à des relations caractérisées à la fois au niveau des contenus et des descriptions, ce qui présuppose que les utilisateurs ne cherchent pas seulement des réseaux mais des hypothèses argumentées, des méthodes et des propositions. Par conséquent, la structure interne du document envisagée du point de vue des contenus ne constitue pas simplement une question associée au modèle mais à l'usage. L'accroissement des capacités de circulation d'information constitue un point fondamental de l'argumentation de ceux qui proposent d'entrer dans le contenu des documents. S. Pettifer & alii. (op. cit.) proposent de lier certains

contenus des publications à des ressources externes. Cette idée repose sur un traitement particulier du document, fondé sur le marquage par des balises de certains contenus. C'est sur l'idée que l'usage proposé du modèle se situe autour de la publication que l'on va partir, pour ensuite considérer comment cet usage pourra être caractérisé. Le présupposé de l'ensemble de notre propos est bien que l'usager acquiert de la connaissance autour de l'exploration d'une publication. On se distingue en partie des propositions de l'e-science, qui concernent essentiellement la menée d'une recherche, en se focalisant sur l'acquisition de nouvelles connaissances, leur validation et leur utilisation.

6.2.2.1. Document et document annoté.

Grâce à la numérisation, le statut même du document est transformé, tout simplement parce que de nouvelles fonctionnalités apparaissent (apparition de versions, changement de supports, etc.). Par ailleurs, cette même numérisation permet à des objets qui jusque-là ne pouvaient être considérés comme des documents publics (données primaires de la science par exemple) de le devenir. Mais surtout, la numérisation s'accompagne d'annotations de documents qui constituent une description de traits et de composants de ce document.

Des principes relativement similaires caractérisent les microdata. Leur généralisation au travers de la mise en place du format HTML5 permet de réaliser un taggage simple de contenus que l'on souhaite faire reconnaître par des moteurs[252] . Les différents schémas des microdata sont répertoriés dans SCHEMA.ORG. Le taggage des documents sert aux moteurs de recherche (GOOGLE en premier lieu) à repérer des parties de documents contenant des informations plus particulièrement pertinentes dans le cadre d'une recherche.

Ce taggage web constitue la version la plus sommaire des annotations. A l'inverse des outils comme la TEI permettent de caractériser des annotations dans des domaines de connaissances très divers. Néanmoins, on ne peut réduire la TEI à cet usage-là.

La TEI a comme intérêt de constituer un format universel, qui par ailleurs peut s'adapter à des usages très différents, mais autour du repérage de séquences du document ayant un intérêt particulier dans le cadre d'une certaine exégèse. Or, la TEI ne dispose pas d'une approche par extraction ni ne possède liens aux microformats.

Enfin, les annotations linguistiques constituent une autre forme de structuration des données, pour des enjeux qui peuvent être assez différents. Néanmoins, ces annotations linguistiques peuvent être associées à la TEI, comme par exemple LMF (Lexical Markup Framework). Nous reviendrons là-dessus lorsqu'il s'agira de définir le lien que l'on opère entre description de documents et contenu des documents.

Toutes ces techniques d'annotation sont aujourd'hui reprises et insérées dans les projets de structuration de l'édition sémantique. L'édition sémantique est un concept qui représente le fait d'éditer un document en intégrant des liens vers d'autres objets numériques (les données utilisées, les outils, les références). L'édition sémantique repose sur l'annotation du document électronique, mais en intégrant la relation de certains contenus à d'autres objets que sont les données primaires, le cadre institutionnel.

L'édition sémantique n'est pas de l'extraction d'information mais permet d'entrer dans le texte lui-même. Ainsi, le document est loin de constituer une entité imperméable à l'extraction de données (avant de parler d'information). La distinction usuelle entre la description documentaire et l'extraction de contenus peut être aisément dépassée. Dans le cadre de

[252] http://microformats.org/wiki/microformats-2-fr

l'édition sémantique, le propos est limité à la relation entre des entités du texte et des ressources extérieures. Comme nous l'avons déjà évoqué, il existe deux approches distinctes,
C'est le point de vue qui a été adopté par SPAR (DoCO, the Document Components Ontology[253]) et par le collectif UTOPIA[254] . Néanmoins, les propos de ces deux outils sont très différents, puisque pour le premier il s'agit d'une ontologie des parties de document et pour le second il s'agit d'un outil permettant de connecter des contenus scientifiques en pdf avec des contenus en ligne qui lui sont relatifs. A l'aide des métadonnées notamment, UTOPIA va chercher les documents relatifs, les blogs et les réseaux sociaux.

Le système que l'on propose, parce qu'il permet de caractériser comme métadonnées des termes qui par ailleurs appartiennent au document, change quelque peu le rapport qu'il peut y avoir entre le contenu du document et sa description.
Dans le schéma général du web, le document est traité de façon à ce que seules ses métadonnées soient publiées, à savoir accessibles par les moteurs de recherche. Dans le cadre des Linked Data, ces métadonnées peuvent être liées à d'autres ressources et d'autres outils descriptifs. Le document est ainsi inscrit dans un cadre distribué. Comment alors ces compléments, précisions et spécifications d'information peuvent-ils être récupérés et inscrits dans les données elles-mêmes de façon à ce que l'on puisse les décrire ? Cet objectif là est le propos du projet MIM[255], animé par Matthew Gamble. D'un côté, on dispose de documents scientifiques relatant des expérimentations : le domaine présenté est celui des sciences de la vie. D'un autre côté, dans les Linked Data, on dispose d'une ontologie décrivant les conditions et les contraintes d'une expérimentation reproductible.
La perspective se fonde sur une liste de contrôle (« checklist ») qui associe à chaque expérimentation un ensemble d'informations relatif à la construction et l'usage de la ressource publiée, mais également sur ceux qui ont mené cette recherche. Comme ce projet est mené dans le cadre des sciences expérimentales, les informations portent sur les conditions de reproductibilité de l'expérience.

On considère qu'un utilisateur cherche avant tout un certain cadre de travail dans lequel il pourra trouver par sa lecture des informations pour sa propre recherche. Ce cadre ne peut être que partiellement décrit par les jeux de métadonnées habituels parce qu'ils n'ont pas accès au contenu des documents. Or, à partir du moment où on a accès à une partie de ce contenu, relatif à des phénomènes que décrivent les métadonnées (l'origine du document, les acteurs impliqués, le contexte), on peut commencer à articuler les contenus et les descriptions.

Comme on l'a laissé entendre plus haut, les travaux de description interne des ressources d'e-science et des relations que ces ressources entretiennent entre elles laissent envisager la perspective d'une transformation de la description des publications. Ce programme est clairement envisagé par Sean Bechhofer et alii.[265] lorsqu'ils envisagent de caractériser des objets de recherche. Les publications ne doivent plus être envisagées comme des objets linéaires mais se caractériser par toutes sortes de liens entre certaines parties de leur contenu et d'autres ressources.

Cette perspective montre un lien de plus en plus marqué entre les descriptions de document et le workflow, défini comme « The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules."[266]

[253] http://www.essepuntato.it/lode/http://purl.org/spar/doco
[254] http://www.utopiadocs.com/
[255] http://sierra-nevada.cs.man.ac.uk/mim/ns

Ainsi, la perspective de l'annotation se couple avec celle du workflow, à savoir le lien que l'on peut établir entre les objets, publications et ressources, et le processus de recherche. Ces processus ne sont pas inscrits dans une organisation précise mais dans la vaste communauté des chercheurs travaillant sur un même sujet. Ainsi, les procédures, conditions et globalement méthodes de la recherche tendent à devenir transparentes. Il en résulte également d'autres formes d'évaluation de la recherche, au-delà de la qualité d'une argumentation.

L'approche de S. Pettifer & alii. est nettement plus centrée sur le document : elle concerne l'annotation de documents pdf et les liens qui peuvent être établis, grâce à des triples, avec des objets extérieurs à ce document. Les perspectives workflow et annotation sont corrélées : il est possible de représenter la façon dont la recherche a été menée dans le cadre du workflow au sein de la publication des résultats. Une telle perspective permet de rendre visible et consultable une part essentielle du travail de recherche que la concision des publications ne permet pas de rendre compte. Ainsi, une publication courte peut renvoyer à un programme et des travaux scientifiques bien plus longs et conséquents.
Ainsi, le workflow s'articule très précisément à l'annotation : les structures annotées renvoient à des étapes précises et répertoriées du travail des chercheurs qui a alimenté cette publication. Khalid Belhajjame et alii[267] donnent une idée précise de ce que pourrait être cette mise en relation.

Ces perspectives sont quand même relativement distinctes des nôtres dans la mesure où les documents sont essentiellement considérés à partir de leur capacité à être mis en relation à d'autres ressources. Or, ce qui nous intéresse est leur capacité à être enrichis par les relations à d'autres ressources par le biais de métadonnées.

Les perspectives de l'e-science sont relativement distinctes de celles des bibliothèques dans la mesure où elles ne requièrent pas de structuration de connaissances. Or, il s'agit bien là du principe des bibliothèques. Ainsi, l'articulation entre l'e-science et l'univers des organisations de connaissances n'est pas encore attestée.

6.2.2.2. Approches centrées sur la description des documents et projets d'e-science.

Nous aimerions maintenant caractériser plus précisément l'état actuel de l'articulation entre ces deux univers d'outils et positionner notre apport. Nous aimerions montrer ici l'intérêt de nos métadonnées et surtout l'originalité de ce qu'elles représentent par rapport au jeu du DC. En effet, les métadonnées caractérisent un document, mais à la différence des micro-données, ne cherchent pas à identifier des contenus de textes. Dans notre proposition, les métadonnées caractérisent des phénomènes qui ne sont pas de l'ordre du document mais de processus qui sont relatés dans le document ou en relation à lui, mais qui permettent de le caractériser dans son entier. Autrement dit, on se distingue également ici de l'annotation de textes qui le décompose ou sélectionne des entités pertinentes mais ne permet pas de décrire le document

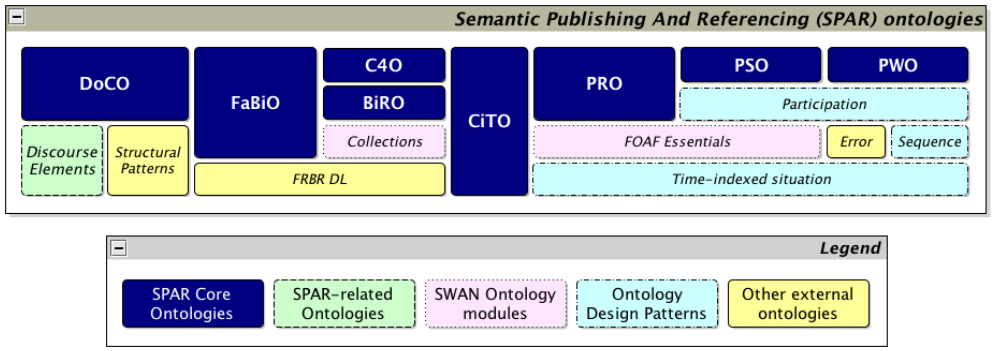
On transforme l'univers d'interprétation des métadonnées, et c'est une conséquence directe de l'adoption à la fois des principes de RDF et de l'idée d'un web structurant des données hétérogènes. On construit un outil qui repose sur une ontologie de domaine et donc qui se distingue ainsi des ontologies bibliographiques. En effet, le domaine est la seule structure de représentation qui nous permet de caractériser des contenus de façon précise. (Les ontologies

bibliographiques permettent de produire des liens signifiants qui eux-mêmes construisent un objet de recherche, mais n'explicitent pas les éléments du domaine concernés).

Pour cela, il nous a fallu partir d'un modèle de domaine dans lequel certes des documents s'intégraient, mais également des connaissances qui étaient relatées dans des expressions appartenant à des jeux de métadonnées spécialisés. Dans le cadre des ressources linguistiques, on pourra citer META-SHARE, IMDI-CMDI, BAMDES et OAI-OLAC. Il en va de même par exemple lorsque les valeurs associées à des attributs du DC proviennent de vocabulaires contrôlés extérieurs. Ces vocabulaires contrôlés sont soit des classifications universelles de type DEWEY ou de la bibliothèque du Congrès que des vocabulaires plus spécialisés comme le MESH (médecine) ou le TGN pour les termes de géographie. Malgré ces limitations, le DC intègre des domaines hétérogènes par rapport à la seule description des documents[256]. .

Nous suivons partiellement le principe de la mise en relation d'outils de structuration de connaissances et créant de toute pièce une ontologie fondée sur des structures de connaissances certes distribuées mais représentant un même domaine. Certes les métadonnées ne peuvent être comparées à des systèmes de classification, mais elles représentent, dans un certain domaine, ce que l'on peut dire des documents spécifiques en utilisant un langage normé et structuré.

Nous présentons maintenant l'approche qui se met en place autour de SPAR[257] . SPAR est une suite de modules d'ontologie destiné à créer des métadonnées RDF, lisibles par une machine, et couvrant l'ensemble des aspects de la publication scientifique et du référencement. L'idée consiste évidemment à dépasser les descriptions bibliographiques usuelles pour proposer des relations entre ces publications et celles qu'elles citent ou qui les citent, aux processus de publication et de validation académique, mais également à des parties du document. C'est en tout huit ontologies qui peuvent être utilisées de façon à accroitre la mise en relations des différentes descriptions des documents.



[256] http://dublincore.org/documents/dcmi-terms/#terms-NLM
[257] http://sempublishing.sourceforge.net/

Chaque ontologie a un rôle particulier dans le cadre de SPAR. Certaines sont liées aux fonctions d'édition comme PSO et PRO, d'autres décrivent les relations à d'autres publications (C40 et BiRO). Dans la caractérisation du contexte des documents, PWO caractérise le workflow dans lequel se situe la publication et FaBiO enregistre et publie les descriptions des entités publiables ou publiées, et plus généralement des documents qui ne sont pas nécessairement des publications académiques. Dans la caractérisation des contenus, DoCO est une ontologie des composants de la publication et CiTO décrit les différentes formes de citations.

FaBiO[258] est plus précisément une ontologie relationnelle qui utilise les liens FRBR pour caractériser les liens entre des documents apparentés. PWO[259] propose de caractériser des liens entre documents dans le cadre de la menée d'un travail, intégrant la dimension temporelle. Ces outils relationnels sont relativement proches des outils d'e-sciences proposés par Carole Goble.

Le service que l'on propose est quelque peu différent et prend en compte de façon plus accentuée la distribution. En effet, l'hétérogénéité des structures et des données permet de mettre en valeur un trait fondamental des métadonnées par rapport aux représentations de domaines (qu'il s'agisse des thésaurus ou des grandes classifications). Par les relations, les versions, les métadonnées de type DC inscrivent le document dans une temporalité. Les métadonnées plus spécifiques, comme celles relatives aux ressources linguistiques, accentuent cette dimension. Par conséquent, les métadonnées fournissent quelques éléments pour inscrire le document qu'elles décrivent dans une dimension temporelle et événementielle.

Cet aspect-là peut être largement accentué dès lors que l'on s'intéresse à la relation entre des données hétérogènes dont l'une précède d'autre. Ainsi, le principe des métadonnées, qui consiste à associer à un document son contexte de façon à construire sa description documentaire, s'enrichit d'une caractérisation accentuée de la temporalité et donc des événements associés à la création du document.

Ainsi, on est amené à considérer le document comme un assemblage de résultats d'opérations et de transferts. Le document se caractérise alors d'abord comme une mise en scène (et en espace) de ces circulations. Autour de la description du document, de son accès et de l'évolution de sa représentation, on trouve la démarche de la production du document. C'est le processus inverse de l'étude de l'impact factor : on ne s'intéresse pas à l'aval d'une publication, mais à son amont.

Mais cet amont peut être caractérisé comme une suite logique d'événements : il s'agit de considérer les différents événements de recherche qui ont amené à l'édition de ce document.
L'équipe de Manchester (autour de C. Goble) propose autour de ces descripteurs de ressource un concept fédérateur : celui de « Research Objects ». Ils constituent une première modélisation des descripteurs utiles pour l'e-science. « A Research Object (RO) provides a container for a principled aggregation of resources, produced and consumed by common services and shareable within andacross organisational boundaries. An RO bundles together essential information relating to experiments and investigations. This includes not only the data used, and methods employed to produce and analyse that data, but also the people involved in the investigation. » Sean Bechhofer & alii., op. cit., p.1).

[258] http://www.essepuntato.it/lode/http://purl.org/spar/fabio#d4e3999
[259] http://www.essepuntato.it/lode/http://purl.org/spar/pwo#namespacedeclarations

Les « objets de recherche » sont considérés comme des réservoirs de connaissances. Ce ne sont pas des métadonnées. Par contre, ils auront besoin de métadonnées (qui peuvent être du Dublin Core simple) de façon à assurer l'accessibilité de leurs réservoirs. Ainsi, les propositions de Goble visent plus à produire des réservoirs de connaissance pour le travail collaboratif que d'assurer la diffusion et la découverte de ces ressources.

C'est là également que l'on propose quelque chose de différent par rapport au modèle proposé par C. Goble : notre projet intègre des relations de conséquences et de contraintes entre les ressources et les résultats et explicite le rôle des contraintes dans le résultat. On cherche à expliquer ce qu'apporte la ressource dans la publication.

On peut également mettre en évidence une autre distinction avec les propositions de l'e-science : les objets de recherche sont des constructions dont la dynamique est seulement marquée par les indices temporels. Dans notre cas, il s'agit d'accompagner ces marqueurs de contraintes et conditions, donc de contenus.

L'objectif à terme de notre projet (dès lors que les inférences auront permis la mémorisation de suffisamment de relations) consiste à proposer un service reposant sur ces connaissances encapsulées. On apprend des relations entre les ressources et les publications certains usages des ressources comme la portée des publications.

Néanmoins, les métadonnées que l'on propose peuvent à priori sembler concurrentielles des descripteurs proposés par l'équipe de Carole Goble. En effet, nos métadonnées caractérisent des processus. Il existe une différence importante entre ce projet et le nôtre : celui du domaine dans lequel il s'inscrit. Le projet de l'équipe de Carole Goble est fondé sur la médecine, et plus globalement les sciences du vivant, pour lesquelles toute expérimentation doit être reproductible. Ce n'est pas du tout le cas des données de linguistiques, pour lesquelles une donnée primaire se doit d'être le plus facilement utilisable pour tester et valider une hypothèse. La construction de la donnée primaire doit ainsi être adéquate par rapport au cadre théorique et méthodologique de la recherche menée. Il ne peut être pour nous question de reproductibilité, ce qui fait que les propositions de Carole Goble sont relativement inopérantes dans notre cas. Par contre, nos trois éléments (`<foundation>`, `<material>`, `<method>`) sont inspirés de la planification des articles en sciences médicales et pharmaceutiques. Ils sont donc très généraux et s'adaptent à un nombre beaucoup plus important de sciences.
La connaissance que l'on veut caractériser s'inscrit à l'intérieur de la relation inférentielle entre les deux ressources. Les métadonnées servent à la rendre visible sur le web.

Quelles sont les possibilités qui s'offrent à l'écriture de métadonnées ?

- Les données présentées ont déjà été utilisées dans le cadre de plusieurs travaux, ce qui permet d'indiquer quelle est la portée de cet outil.
- Par ailleurs, les données primaires ne sont que rarement de seules données collectées mais subissent certains traitements avant de pouvoir être utilisables.
- Enfin, ces traitements ne sont pas (toujours) neutres. Ainsi en est-il des corpus de linguistique : ils sont traités de façon à ce que l'on puisse y trouver certains phénomènes linguistiques.

C'est en ce sens que des métadonnées sont plus particulièrement pertinentes : elles permettent

de caractériser avec précision les opérations et les transformations effectuées sur le document et qui en infléchissent l'usage.
Elles se différencient des choix faits par le DC, dans la mesure où J. Greenberg part d'une caractérisation des métadonnées pour l'e-science par les tâches (« goal-oriented metadata schemes for specific user tasks. » (p.7)).
Le problème de cette perspective est qu'elle ne permet pas de représenter les usages, et en premier lieu la façon dont une ressource peut être décrite par son usage (au sein d'autres ressources).

Néanmoins, la caractérisation de relations entre ressources par des indices d'usage ne peut être envisagée que sur la base des valeurs, voire éventuellement d'attributs. Par contre, il faut leur associer des éléments spécifiques, neutres par rapport à un domaine de recherches.

A titre d'hypothèse, et pour amorcer les propositions qui vont suivre, on proposera un cadre d'usages spécifique à la menée des recherches. En ce sens, si l'on choisit pour sa recherche un outil ou une ressource, on peut le justifier autour de trois arguments :
- C'est un même cadre théorique (champ de recherche et applications).
- Ce sont les observables qui nous intéressent (catégorie d'observables et éléments décrits).
- On partage un même traitement des données primaires (opérations sur les ressources et contenu sémantique).

Ces trois grandes catégories d'éléments pourront être construites en tant qu'éléments. Elles sont complémentaires des jeux à la fois généraux et spécialisés, tout simplement parce qu'elles sont fondées sur la caractérisation d'autres traits que ceux représentés par les jeux existant. Ainsi, par rapport au tableau de J. Greenberg, les éléments que l'on propose s'intègrent autant dans le cadre du contexte scientifique que des métadonnées sémantiques.

6.2.2.3. Directions suivies par les jeux de métadonnées : initiatives pour la prise en charge des questions d'e-science.

Le premier usage de nos métadonnées est de renseigner les publications en leur attribuant des valeurs relatives à leur contenu. Néanmoins, leur usage peut être étendu, notamment en prenant en compte la mémorisation des relations que l'on établit entre les deux ressources hétérogènes. Face aux questions posées par la multiplication des ressources et des documents comme la perspective de l'e-science, le Dublin Core développe à la fois les profils et la relation aux langages contrôlés, et la relation à des ontologies comme VIVO qui font le lien avec les projets d'e-science.
Les développements du Dublin Core sont fidèles à l'idée que les outils d'organisation des connaissances doivent être structurant des descriptions de documents, et donc assurer une cohérence des résultats lors d'une recherche d'information. La recherche d'information n'est pas du tout la préoccupation des chercheurs du domaine de l'e-science.
La préoccupation du Dublin Core reste la structuration des descriptions en conservant le principe d'un vocabulaire commun : les relations de données doivent être intégrées à l'intérieur de la structuration des descriptions d'objets et ainsi obtenir une plus grande homogénéité lexicale.

Notre projet s'inscrit dans cette perspective, entre autre parce qu'il s'intègre dans les jeux de métadonnées existant et qu'il réutilise le vocabulaire déjà constitué dans les métadonnées spécialisées, au sein des communautés. En somme, on propose, par un jeu de métadonnées restreint et un vocabulaire commun, une entrée dans le cadre des connaissances mises en œuvre au sein d'une publication.

On pourrait proposer d'étendre l'élément <relation> du Dublin Core, y compris en le spécifiant de façon à ce qu'il représente pour une publication, les ressources qu'elle utilise, et pour une ressource, les publications y faisant directement référence. Or, justement, ces liens « externes » ne permettraient pas de décrire le contenu commun de ces deux données, et donc leur enrichissement mutuel.
La perspective que l'on choisit consiste à exploiter au maximum les possibilités offertes par les mises en relation de données hétérogènes. Cette démarche implique de ne pas considérer seulement les métadonnées en tant qu'outils de description et d'identification des documents, mais bien en ce qu'ils expriment des propositions sur les objets numériques auxquels ils réfèrent. Les questions des métadonnées sont alors celles de la circulation des descriptions, et des relations entre ces descriptions et les représentations de connaissances. Notre projet rapproche les métadonnées d'une information sur le document au sens où on l'a défini précédemment.

Lorsque l'on définit les métadonnées par leur portée informationnelle, et par extension propositionnelle, on établit un lien avec le contenu du document.
Par ailleurs, les métadonnées apparaissant soumises aux conditions de vérité, elles sont situées et donc associées à une situation d'énonciation, distincte de celle du document par rapport auquel elles prennent sens.

Notre projet s'inscrit dans une définition sémantique des métadonnées, qui n'est pas liée à leur histoire (ce sont au départ des notices bibliographiques numériques), mais à l'adoption de standards qui eux sont fondés sur cette dimension prédicative. Rappelons le principe définitoire de RDF : « RDF is an assertional logic, in which each triple expresses a simple proposition »[268]. .
En adoptant RDF, le Dublin Core a changé la nature des métadonnées en en faisant des expressions au sens logique.

Il y a une autre dimension importante de RDF, qui est sa généralité. RDF ne spécifie pas la nature des objets de référence des URI. Par conséquent les URI peuvent être de nature très différente, ce qui permet de lier des objets totalement hétérogènes. (« There are several aspects of meaning in RDF which are ignored by this semantics; in particular, it treats URI references as simple names, ignoring aspects of meaning encoded in particular URI forms [*RFC 2396*] and does not provide any analysis of time-varying data or of changes to URI references. It does not provide any analysis of indexical uses of URI references, for example to mean 'this document' ». (RDF primer, source citée)

Code de champ modifié

Code de champ modifié

Dans ce cadre, des objets extrêmement variés peuvent être mis en relation. La contrainte est que ces expressions puissent être interprétées à l'intérieur d'un modèle de façon à ce que les inférences produites à l'intérieur du modèle soient valides.
Les deux propriétés de RDF, qui sont d'une part l'absence de caractérisation indexicale et la caractérisation d'un modèle d'interprétation pour valider les inférences, permettent de relier des objets très hétérogènes à condition que l'on ait un même modèle d'interprétation.

Néanmoins, nous devons approfondir cette question du rapport entre les relations entre les descriptions externes (via les métadonnées) et les descriptions internes (via les annotations). Un autre aspect à prendre en compte est le rôle que pourrait jouer la publication par rapport à l'ensemble des données disponibles relatives à une activité de recherche. En effet, comme on l'a déjà évoqué (et c'est d'ailleurs ce qui sert de fondement à la construction des ontologies SPAR), la publication ne constitue qu'une partie de la mise en visibilité des résultats d'une

recherche. Par exemple, une ontologie publiée dans http://purl.org/docs/index.html et un wiki constituent des compléments nécessaires à une publication dans le domaine des ontologies. Les résultats qui sont donnés à lire dépassent largement le cadre des publications, y compris pour la compréhension des contenus. Nous n'avons pour le moment pris en compte que la dimension des ressources utilisées comme données primaires et des outils parce que dans un cadre des ressources linguistiques, elles sont encore les seules ressources numériques représentées.

Nous venons de voir de quelle façon les questions de signification s'intégraient dans la description documentaire. Cette intégration est liée aux formats de représentation utilisés et donc ouvre de nombreuses possibilités pour associer les métadonnées à d'autres outils d'annotation et d'exploration du document. C'est sur ce lien que l'on va se pencher maintenant.

6.3. Langage et description de documents et de ressources.

Les questions relatives au langage parcourent l'ensemble de ce travail et le projet de façon assez particulière puisque nous n'avons pas une problématique linguistique. Les questions relatives au langage caractérisent deux dimensions essentielles :

– la sémantique des descriptions de documents (qui est une condition à leur mise en relation).
– La sémantique des structures qui assurent la circulation d'entités informationnelles au travers de structures hétérogènes.

Les questions linguistiques et plus particulièrement de symbolisation nous permettent d'aller au-delà des relations conceptuelles. L'appareillage sémantique que nous avons présenté plus haut est utilisé pour caractériser ces relations entre objets, descriptions et structures conceptuelles.

Nous avons envisagé ces questions de façon globale dans notre présentation. Maintenant, sur le projet, nous les envisageons de façon nettement plus précise. Dans un premier cas, il s'agit de la construction des relations entre structures conceptuelles et linguistiques, notamment relativement aux ontologies et à la façon dont elles sont articulées à l'extraction d'information et aux flux. Dans le second cas, les questions de structure d'information sont associées à la question de l'extraction. Cette maîtrise des patterns permettra aussi de montrer, sur le projet, le fait que l'on peut effectivement articuler les questions d'information à la fois à propos des flux et des structures.

Une bonne partie de la sémantique que nous avons présenté, et notamment la théorie des situations, sert en premier lieu à représenter la spécificité référentielle de notre jeu et son écriture utilisant RDF. Il ne parle pas des mêmes objets du monde que les autres jeux, notamment tous ceux qui sont issus de la tradition bibliographique puisque nous avons affaire à des structures de ressources hétérogènes.

Nous montrerons donc en quoi les perspectives sémantiques présentées, notamment celles relatives à la structure d'information, sont fondamentales pour notre projet, et plus particulièrement l'analyse et l'élaboration des métadonnées. En ce sens, nous poursuivons à la

fois le travail fait par le Dublin Core à propos de l'utilisation de RDF autant que le soubassement des propositions de SPAR.

Nous distinguons entre des questions sémantiques, qui constituent des modélisations assurant la cohérence globale du propos, de l'adoption de modèles linguistiques partiels, comme ceux qui ont trait à la structure d'information. Ces derniers répondent à des objectifs limités spécifiques à l'extraction. Comme nous le verrons, il existe un nombre de plus en plus important d'outils servant à extraire de l'information. Néanmoins, ces outils ne peuvent remplir leur mission qu'à condition qu'on leur associe des modèles linguistiques caractérisant l'information, tels qu'ils puissent extraire les patterns recherchés. Les modèles statistiques et probabilistes, qu'ils soient supervisés ou non, ne peuvent être correctement utilisés que si on leur indique ce qu'ils ont à extraire. C'est justement ce que nous nous employons à caractériser dans la partie 4.2. .

Nous avions indiqué que les dimensions sémantiques qui étaient associées aux flux, au travers du rôle que ces derniers jouaient dans la théorie des situations, sont au centre du projet. La dimension sémantique s'insère à l'intérieur de la caractérisation du raisonnement fondamental de notre modèle. La question de la prédication sera traitée dans ce sens. Nous nous limiterons ici à la prédication, sans entrer dans les questions de structuration de connaissances, développées ultérieurement.

La distinction que l'on opère par rapport à la linguistique réside dans le fait que cette dernière construira une problématique à partir d'une propriété du langage (comme par exemple la représentation d'événement ou la coréférence) ou d'une de ses spécifications (comme les verbes d'événement terminatifs ou les structures lexicales coréférentes). Les problématiques qui nous concernent ne sont pas linguistiques, mais bien des questions d'information. Or, comme nous l'avons vu, les questions d'information sont fondées sur celles de symbolisation. C'est donc tout naturellement que toute volonté de construire une sémantique à partir des recommandations du W3C se pose des questions de sémiotique. Si les formulations peuvent être différentes entre les auteurs du Dublin Core et ceux de SPAR (parce que les utilisations de RDF n'ont pas les mêmes objectifs), les questions se posent avec une même acuité.

Pour répondre à cette question, on s'intéressera ici plus particulièrement à la prédication et à un outil méthodologique, les niveaux d'abstraction. Dans ce cadre, nous montrerons également comment un modèle de raisonnement comme celui des flux, à la fois en tant que théorie caractérisant les contraintes à la circulation de l'information et en tant qu'outil de modélisation des échanges informationnels, permet d'ajouter des contenus informationnels : le fait de caractériser un contexte d'élaboration d'une publication entraîne une explicitation de son contenu. Ce programme est proche de l'herméneutique. Notre seconde hypothèse est que la meilleure façon d'utiliser au mieux ces ressources, à la fois dans un cadre méthodologique et pratique, consiste à les structurer par niveaux d'abstraction, et ainsi spécifier les héritages de contenus.

6.3.1. Approches du langage et usages dans le cadre de la structuration des objets numériques.

A partir du moment où pour nous il est question d'un usage, la question linguistique est fortement contextualisée : on ne présente pas de spécification des phénomènes propres à la langue ni même au discours mais bien à la capacité du langage de prédire sur des traits d'objets, qui peuvent être des objets numériques. Il y a différentes façons de prédire, relatives toutes à l'usage que l'on fait de cette prédication : classification d'un document, identification de ce document dans le cadre d'une recherche, description des contenus dans le cadre d'une annotation.
Ainsi, cette partie sera dédiée à la distinction entre les représentations d'unités linguistiques, notamment au sein de lexiques, et la sémantique prédicative associée à l'utilisation des langages comme RDF et OWL. Dans la partie suivante, nous traiterons plus explicitement des métadonnées, à savoir de la façon dont on peut rendre compte de la signification de ces descriptions.

Toutes ces questions permettront d'explorer les projets proches du nôtre, en premier lieu SPAR[269] et les travaux d'un de ses promoteurs, A. Gangemi. Nous explicitons la façon dont on envisage les questions de sémantique, et comment on structure théoriquement l'ensemble de données symboliques qui caractérise notre objet de travail.
Nous débuterons par un point sur les structures linguistiques que nous utilisons pour ensuite entrer pleinement dans la façon dont on définit théoriquement l'usage de RDF à l'intérieur de notre objet d'études.

6.3.1.1. Les structures lexicales.

Pour nous, les structures lexicales servent à caractériser un vocabulaire de domaine descriptif des ressources. Dans le cadre des ontologies, les lexiques et terminologies constituent les représentations linguistiques des concepts.

Les ressources lexicales peuvent être caractérisées de deux façons différentes : elles peuvent être fondées sur une sémantique lexicale, comme WORDNET, soit elles sont fondées sur une terminologie (plus ou moins élaborée d'ailleurs).

L'ensemble des ressources lexicales fondées sur les principes des LINKED DATA est présenté par l'Open Linguistic working group en utilisant un nuage (qui lui-même est une partie du nuage des LINKED DATA)[260]. Ce groupe propose un service communautaire qui répertorie des outils et service concernant les ressources linguistiques. (Le projet peut sembler proche de CLARIN : C. Chiarcos & alii.[270] expliquent que la différence entre les différentes initiatives de propositions de répertoires de ressources linguistiques et la sienne réside en son universalité et le principe de collaboration entre des acteurs volontaires, en dehors de toute initiative institutionnelle.
Le nuage est structuré autour de la DBpédia, qui constitue également un lien avec le nuage global (dont elle est le centre). Elle regroupe l'ensemble des ressources et outils utiles dès lors que des unités linguistiques ont à être traitées, quel que soit l'usage : il peut s'agir de projets d'annotation, d'extraction, d'apprentissage, etc.
On peut structurer cette offre en distinguant les outils de traitement des lexiques (LEXINFO, LEXONTO, LEXVO), des outils de traitement des unités linguistiques contenues dans les corpus (PAWLA pour ce qui concerne la description des outils d'annotation de lexiques) et des lexiques eux-mêmes (ISO-cat notamment). En dehors de cette caractérisation commune

[260] http://nlp2rdf.lod2.eu/OWLG/llod/llod.pdf

de respecter les principes des LINKED DATA et d'avoir un usage dans le cadre de l'analyse ou de la description des ressources linguistiques, il est difficile de trouver un principe structurant pour ces outils. En effet, le nuage n'est pas considéré comme une structuration des ressources mais juste comme une représentation de l'interopérabilité des ressources.

Il n'est pas nécessaire pour nous d'entrer plus avant dans le détail de ces ressources. Celle qui nous intéresse est celle qui décrit les catégories de phénomènes linguistiques. En effet, cette ressource nous permet de disposer de l'équivalent, pour les ressources linguistiques, des vocabulaires contrôlés pour le DC. Cette ressource est nécessaire pour identifier l'ensemble des valeurs qui apparaissent dans les renseignements de métadonnées mais également dans les expressions décrivant les corpus (et leur intérêt) à l'intérieur des publications les mentionnant. La ressource qui nous est apparue la plus pertinente est ISO-cat.

Ces différences d'usage des unités linguistiques, et donc ces différences sémantiques, ne sont pas prises en compte dans le cadre de la construction de catégories linguistiques, dont l'un des meilleurs exemples est ISO-cat. Il s'agit de caractériser en utilisant les formats ISO l'ensemble des catégories utilisées pour décrire les phénomènes linguistiques. Grâce à l'interopérabilité, ISO-cat fédère un nombre très important d'outils qui catégorisent les phénomènes linguistiques, leurs approches et les objets qui les recueillent assurent leur préservation et enfin permettent de les analyser. Ainsi, ISO-cat donne l'impression d'un matériau lexical brut, faiblement structuré et ne proposant pas un usage précis de ses données. OSI-cat ne propose pas un ensemble cohérent (entre autre parce que plusieurs entrées existent, avec différentes définitions, pour un seul et même terme), mais un ensemble de données exploitables.
On peut également considérer qu'il s'agit d'où outil permettant justement une structuration de l'ensemble d'un lexique en fonction d'usages différents. Notre approche consiste à structurer ces termes en fonction de leur rôle dans une description documentaire.

L'avantage d'ISO-cat consiste à fournir sans structuration contraignante une panoplie de ressources. Ces ressources sont issues d'outils et de répertoires construits à des moments différents et pour des besoins différents (on y trouve un jeu de métadonnées, une ontologie de domaine, une classification, un lexique terminologiques, etc.). Les unités lexicales sont définies et ces définitions contiennent les mots-clés qui servent à toute recherche de terme.
Cette diversité permet de construire la structure lexicale du domaine du point de vue des catégories de phénomènes linguistiques.
Les catégories linguistiques sont des termes qui désignent des collections de phénomènes linguistiques observables. Par exemple, si l'anaphore constitue un phénomène observable, il s'inscrit dans la catégorie de la sémantique discursive. Cette catégorie ne définit pas le cadre théorique, qui peut être celui de la sémantique cognitive ou de la sémantique formelle. Elle caractérise seulement le cadre dans lequel un type de phénomène linguistique peut être observé.

ISO-cat constitue pour nous un corpus dans lequel il est possible de construire notre représentation lexicale du domaine. Elle concerne les catégories parce que les ressources que l'on utilise sont préparées (annotées, segmentées, décrites) pour que l'on puisse observer certains types de phénomènes linguistiques.
Cette structuration du domaine est pertinente pour les ressources (corpus et outils), mais le sera beaucoup moins lorsque l'on s'attaquera à d'autres types de ressources, comme les outils de travail collaboratif (wikis, sites de projets, etc.).

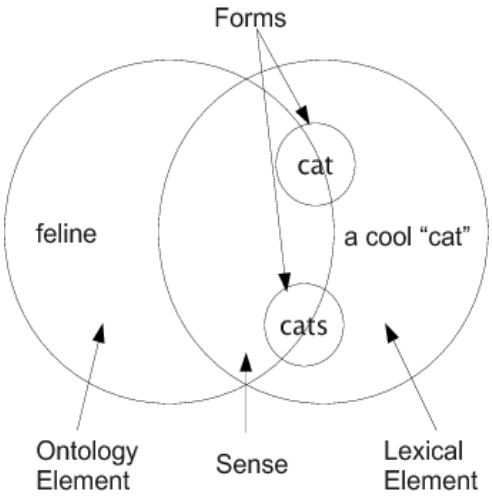
Avec l'ensemble des métadonnées décrivant les ressources du domaine et le lexique des catégories de phénomènes linguistiques, on dispose d'une base de termes utilisables pour notre analyse.
Cette base constitue un ensemble de termes propres au domaine et signifiant, pour les individus experts du domaine, à propos de leurs ressources.

Nous pouvons maintenant entrer plus précisément dans les questions de l'usage des phénomènes linguistiques pour notre propre programme de travail.

6.3.1.2. Objets linguistiques et problématiques linguistiques.
Néanmoins, le problème n'est pas tant celui des structures lexicales en tant que telles mais le cadre dans lequel elles sont proposées, ce qu'elles caractérisent : elles peuvent être liées autant à des vocabulaires que des taxonomies ou des terminologies. En fait, les dimensions lexicales sont toujours considérées dans un cadre (taxonomies, répertoires, terminologies) justifiant leur assemblage. En d'autres termes, les lexiques sont des constructions indépendantes des outils que l'on a précédemment évoqués et qui servent essentiellement d'outils de représentation.

Néanmoins, à partir du moment où l'on parle d'ontologies et de rapport au langage, il reste le problème de la caractérisation des phénomènes linguistiques par rapport aux structures conceptuelles. Pour le W3C, l'ensemble des langages de représentation sont à fondement conceptuel. Seul SKOS, par les labels, réintroduit l'identification des unités linguistiques. Les entités linguistiques sont juste considérées comme des unités lexicales. Dans MONNET, qui constitue un outil de mise en relation des unités lexicales avec des concepts intégrés dans des ontologies, SKOS intervient pour labelliser les termes (`isReferenceOf` to more precisely capture the same semantics as SKOS's prefLabel, altLabel and hiddenLabel) et pour caractériser la forme lexicale (« *lemon* uses the sub-properties of `lexicalForm` »). (Citations depuis : http://www.lemon-model.net/lemon-cookbook/node53.html).
Une solution a été présentée par Jean Charlet et Pierre-Yves Vandenbusshe[271] de façon à modéliser la diversité lexicale dans l'unicité conceptuelle. Comme les auteurs le précisent, la complexité linguistique n'est pas prise en compte. (Ainsi, la proposition ne traite pas de la représentation des unités lexicales comme pourrait le faire LMF).
D'autres solutions se sont développées depuis. LexInfo[272] est un outil qui représente l'information lexicale relative à une ontologie ; il utilise des catégories lexicales, et possède une dimension non prescriptive qui permet des usages souples, comme par exemple dans le projet MONNET. Le modèle LEMON constitue un outil particulièrement complet.
L'ensemble de ces outils est interopérable, utilise les triples RDF : l'objectif étant de produire des relations entre des lexiques et des ontologies, LEMON se doit effectivement de permettre aux lexiques représentés par LMF de pouvoir être intégrés dans cet outil relationnel.

Ces outils sont essentiellement destinés à organiser les relations entre des structures lexicales et ontologiques en considérant leur complémentarité dans la caractérisation de la signification. C'est le fondement de LEMON, dont voici la schématisation de la signification :



Cette schématisation repose sur le principe d'une correspondance purement binaire entre un « élément » lexical et un « élément » ontologique. Elle ne prend pas en compte le fait que le concept existe et est structuré dans le cadre d'une activité, d'un protocole ou d'un cours d'action. Le concept sert d'unité de pensée dans le cadre d'une opération qui ne fait pas partie de l'activité de communication (d'un résultat scientifique au travers d'un article par exemple). En ce sens, la signification que peut avoir une entité symbolique dans un cadre communicationnel (donc linguistique) ne peut être que distincte de la signification de cette même entité dans le cadre d'un protocole.
Par ailleurs, la correspondance entre « chat » et « félin » pose un autre problème. Elle repose sur l'indifférenciation du sens linguistique et du sens conceptuel. L'indifférenciation des dimensions ontologiques et lexicales dans le sens pose le problème fondamental de l'usage, mais également celui de la seule existence des catégories classificatrices comme `félin` (le `chat est_un félin`) dans l'ontologie.
Les flux d'information permettent de considérer chaque contexte de réalisation (lexical et discursif, informationnel et connaissances liées à une activité) comme distinct de l'autre. Les propriétés classificatrices des flux permettent de spécifier les relations (dans le cadre particulier de notre projet) entre les différents contextes.

6.3.1.3. Langage et information.

Comme cela est stipulé dès l'introduction de la recommandation, RDF caractérise l'information qui est associée à une donnée : « The Resource Description Framework (RDF) is a framework for representing information in the Web »[261] . Or par rapport au langage, la question de l'information relève de la partialité des problématiques : on s'attache à représenter des phénomènes partiels relativement au langage ou au discours dans son ensemble, mais qui peuvent apparaître centraux (ou exemplaires) dans la production du discours et son interprétation.

L'information constituant un phénomène relativement partiel du langage puisqu'il ne s'agit que de l'une de ses fonctions, elle constitue également un outil qui permet de passer aisément des questions de langage à des représentations logiques. Dans le cadre du web de données,

[261] http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

des représentations formelles du raisonnement comme les logiques de description qui sont fondées sur des logiques prédicatives peuvent être articulées à des représentations de la prédication fondées sur le langage naturel.

En linguistique, l'information constitue une façon de poser linguistiquement les problèmes relativement différente des autres modèles (Cf. : Z. Harris, A. Joshi, op. cit.) : on part d'un principe externe au langage pour en déduire des règles de prédication. Celles-ci servent alors à produire une grammaire reposant sur l'explication de phénomènes locaux (notamment les phrases nominales). Un tel principe peut effectivement être étendu peu à peu à des problématiques de discours plus étendues, comme les structures d'information par exemple.
L'information, vue comme fonction fondatrice du discours, donne lieu ainsi à une grammaire qui explique comment l'information est réalisée au travers du langage (du lexique plus particulièrement). La dimension sémantique, qui associe alors à ces structures une signification, est une représentation utilisant des types : à telle structure syntaxique correspond un type sémantique.

Ces deux approches relativement différentes montrent comment aborder le langage à partir d'un objet partiel ou d'une des fonctions du langage. Les grammaires de Z. Harris proposent des traitements relativement proches de ceux que les flux rendent possible. Un certain nombre de travaux proposent effectivement de concilier une théorie de types et les principes fondamentaux du distributionnalisme de Z. Harris[273]. La différence entre cette approche et la nôtre réside dans le postulat de la langue (comme véhicule de la fonction d'information) et l'indistinction entre information et discours.

Par rapport aux approches précédentes, les flux et les structures d'information spécifient la fonction informationnelle au sein du discours. Concernant ces dernières, l'approche de H. Kamp prend comme point de départ le discours et les structures qui produisent sa cohérence. En ce sens, l'information s'approche des structures d'information définies dans un cadre rhétorique. Ces approches intègrent totalement à la représentation de l'information dans le média. (Comme on l'a vu, une telle approche repose sur des figures de discours et par conséquent est peu adaptée à notre approche).
Egalement, les flux et les structures dotent cette fonction informationnelle d'un ensemble d'outils de représentation ; par conséquent, le langage perd le monopole de cette propriété au profit d'artifices techniques opératoires comme les opérations distribuées. Nous nous sommes servis de cette dimension afin de valider l'hypothèse d'opérations externes par rapport à l'esprit humain ensuite intégrées dans des structures symboliques constituantes des discours. Par conséquent, on a considéré que la prédication constituait une représentation de phénomènes d'attribution de sens dépassant le cadre des discours. Une conséquence de cette approche est la capacité à distinguer, à propos de l'information, les dimensions de transfert distinctes de celles de communication (incluant les dimensions rhétoriques).

6.3.1.4. Sémantique conceptuelle et ontologies.
Les propositions de LEMON ne répondent que de façon technique à une question fondamentale de la relation entre structures lexicales et ontologies : « comment relier la sémantique des entités lexicales aux ontologies ? »
Cette question ne se pose pas pour des ressources comme ISO-cat mais de façon très précise pour des outils comme les FRAMENET, WORDNET, c'est-à-dire des ressources numériques proposant des descriptions sémantiques élaborées des unités linguistiques. Ce cadre théorique se distingue de celui qui fonde la sémantique associée à RDF sur des principes sémiotiques[274]. Cette modélisation permet au mieux de caractériser les relations entre les unités lexicales et

les concepts. Elle est dénommée Linguistic Meta-Model (LMM) et constitue un méta-modèle permettant de connecter des objets hétérogènes. La proposition d'A. Gangemi repose essentiellement sur une sémantique conceptuelle, et cognitive. Par conséquent, les sémantiques concernées reposent en premier lieu sur des représentations des mots comme concepts et non des unités linguistiques inscrites dans le cadre d'une communication.

Cette perspective pose un certain nombre de problèmes dans notre cas :

- Celui de l'extraction des structures d'information tout d'abord et du modèle linguistique utilisé.

- Plus généralement, le lien entre les activités et productions humaines d'une part, et la conception des outils du web d'autre part.

Nous préférons donc explorer d'autres voies, et principalement celles proposées par les grammaires d'unification, qui permettent de traiter en profondeur les structures linguistiques. Les HPSG constituent des outils qui pourraient apparaître relativement pertinents notamment parce qu'ils articulent les dimensions syntaxiques et sémantiques. Par ailleurs les HPSG[275] positionnent la théorie des situations dans leur modèle de grammaire.

6.3.2. Sur la signification associée aux langages du web de données.

Nous nous proposons maintenant de préciser les conséquences de notre propos sur les descriptions sémantiques d'outils de description documentaire et d'organisation des connaissances, principalement de métadonnées. Comme ces dernières utilisent des langages contrôlés, donc des organisations de connaissances, nous aborderons ici les questions liées à l'intégration des outils d'organisation des connaissances dans les structures décrivant des objets du web. La question est loin d'être simple : « comment décrire la signification des représentations du web de données ? » se dédouble dans celle de structures ayant elles-mêmes leur propre signification, notamment par exemple les grandes classifications.

Le fait que les langages à balise ont en eux-mêmes des propriétés sémantiques est défini dans les vocabulaires eux-mêmes : « Nowadays, a large amount of content stored in digital libraries is encoded with XML. XML, as any markup (meta-) language, provides a machine-readable mechanism for defining document structure, by associating labels to fragments of text and/or other markup. This association has a particular meaning, since each markup element asserts something about its content. What is asserted by the markup is not an issue of the markup itself. In fact, one of the goals of markup meta-languages is to avoid imposing any particular semantics: they express mere syntactic labels on the text, leaving the implicit semantics of the markup to the interpretation of humans or tools programmed by a human. Of course, a lot of markup languages, such as HTML, TEI and DocBook, are accompanied by natural language descriptions of their markup, but those descriptions are not machine-readable; in other words, there is no formal mechanism to embed markup semantics within markup language schemas. »[276]

Dans le projet, les métadonnées sont associées à la description d'une ressource dans sa globalité et relativement à son contexte. L'ensemble de ces descriptions possibles et réelles demande à être associé à une représentation unique. L'ontologie a comme fonction de

structurer les descriptions de ressources. La description de la composition d'une ressource et des fonctionnalités précises qu'elle peut offrir relève du bas niveau d'abstraction de l'ontologie. Le niveau moyen est celui des éléments et attributs des jeux de métadonnées. Enfin, le processus global de production d'une ressource est considéré dans le haut niveau d'abstraction.

Ces questions sont traitées dans notre cas par les classifications des flux d'information. Les classifications sont considérées comme les fondements de la signification dans la théorie de F. Dretske.
On proposera donc de caractériser la prédication au travers des raisonnements associés aux flux et les distinctions signifiantes de niveau (entre les entités lexicales et conceptuelles) au travers des classifications.

L'ontologie que l'on propose intègre des structures différentes : un lexique du domaine et des métadonnées. Au vu de tout ce que nous venons de dire, on ne peut se contenter de l'établissement d'équivalences. Elle repose donc sur une hétérogénéité sémantique des composants des différents niveaux. Les principes de classification des flux sont ainsi fondamentaux pour la mise en relation entre ces niveaux. (Plusieurs items lexicaux peuvent être classés relationnellement dans un type informationnel puis conceptuel en vertu de la capacité de la pensée à schématiser et modéliser).

Les métadonnées sont considérées comme des structures conceptuelles. Ainsi les métadonnées du DC ont pu être reliées à une ontologie de façon à optimiser la caractérisation des sujets dans le domaine médical[277]. Néanmoins, cette proposition ne s'interroge pas sur la signification de cette mise en relation.

Cette approche par la prédication distingue notre projet des propositions d'A. Gangemi (op.cit.). Le modèle qu'il propose pour structurer la sémantique de RDF et OWL est fondé sur la sémiotique peircienne. Son constat de départ est le même que le nôtre, à savoir qu'il est nécessaire d'adopter un cadre théorique pour structurer les triplets RDF. Notre solution consiste à centrer les triples RDF autour d'une activité prédicative.

6.3.3. Remarques sur les niveaux d'abstraction et la construction du projet.

Cette partie s'intéresse à la relation que l'on établit entre les structures lexicales et les différents niveaux de la conceptualisation. Il s'agit ici de l'usage de la théorie des niveaux d'abstraction de L. Floridi. Cette partie permet de rendre compte des aspects les plus fondamentaux de notre modélisation.

Si terminologie et lexicologie diffèrent fondamentalement dans leurs objectifs, leurs objets et leurs fondements, elles ont en commun de construire des structures à un faible niveau d'abstraction, notamment par rapport aux flux et aux structures d'information. Néanmoins, de telles analyses sont pertinentes parce que leurs résultats peuvent être utilisés dans les niveaux qui nous concernent. En premier lieu, l'existence de concepts communs, comme celui de domaine, de concept, de relation, permet de définir des objets d'étude communs. En second lieu, la finalité permet de diriger la traduction des résultats d'un niveau vers un autre : on

considère qu'une représentation de connaissances ou une terminologie, un lexique, émergent de l'analyse d'une pluralité de documents, sensés les contenir.

Taxonomies, thésaurus, lexiques et terminologies sont des produits des démarches de sciences concernées par l'information comme par ailleurs la terminologie. Il faut éviter une confusion qui limiterait à la fois la dimension scientifique et celle d'inter-science : les spécificités de niveau d'analyse et les traductions entre les niveaux d'abstraction constituent les meilleures garanties de pertinence de chacune des approches scientifiques. Par conséquent, et c'est ce que l'on cherche à mettre en œuvre dans les projets, c'est la complémentarité de ces deux ensembles de travaux : la sémantique formelle « élargie » et les analyses fonctionnelles du langage (si cette dénomination convient pour caractériser à la fois la lexicologie, la terminologie, l'extraction de contenus et la conception de thésaurus).

Du côté de l'usage, dans le projet, les questions de niveau d'abstraction ont un rôle particulier puisqu'il s'agit d'abord de caractériser comment positionner les flux dans le cadre des outils lexicaux et ontologiques du web.
Une question fortement liée est celle de l'articulation entre les ontologies et les métadonnées.
Une dernière est la façon dont les ontologies peuvent être reformulées et réorganisées par l'usage qui peut en être fait, notamment relativement aux flux et aux métadonnées.

L'articulation que l'on propose entre d'une part les outils classificatoires et d'autre part les flux peut sembler paradoxale. En effet, les flux reposent sur des classifications. La différence réside dans le fait que les flux ne classent pas des entités de même niveau d'abstraction (à la différence des outils documentaires), mais des entités appartenant à des niveaux d'abstraction différents. C'est cette distinction de niveau d'abstraction qui permet de fonder le lien entre entités hétérogènes.

Les niveaux d'abstraction que l'on utilise sont limités aux dimensions informationnelles. On limite la conceptualisation à sa nécessité pour la production, la circulation et l'interprétation de l'information.

Il faut bien comprendre que les bibliothèques et ensuite leur mise en réseau se sont constituées sur la base de l'assemblage d'objets plus ou moins hétérogènes. L'idée de langage commun classificatoire répond à cette demande. Aujourd'hui, entre autre parce que l'on dispose de documents numériques de même format, ou tout du moins avec des formats interopérables, les problèmes se posent un peu différemment.

Dans le projet, l'approche se caractérise par l'articulation de plusieurs points de vue sur le phénomène de mise en relation de documents : au même titre que l'activité dans notre exemple de l'adaptation, ces perspectives (usages sociaux, lexiques sémantiques, ontologies, recherche et extraction d'information) peuvent être utilisées indépendamment et pour des finalités différentes. Elles sont fédérées dans un objectif unique et ensemble elles permettent chacune la caractérisation d'un cadre dynamique dans lequel les flux s'inscrivent.

A partir du moment une unité documentaire peut être appréhendée par de multiples réseaux, que devient donc son statut ? En d'autres termes, l'appartenance à une collection, qui garantissait auparavant une identité au document, n'est-elle pas dissoute par les accès et lien multiples ?

6.4. Description de documents et représentation des connaissances.

La question des connaissances est présente dès qu'il s'agit de caractériser des inférences entre des structures hétérogènes. On considère alors des structures conceptuelles destinées à représenter le domaine dans lequel ces inférences se réalisent.

Il est nécessaire en premier lieu de caractériser précisément comment on entend le terme de structure conceptuelle, notamment dans le cadre sémantique que l'on s'est donné.

A la différence de la partie précédente qui s'intéressait aux aspects sémantiques, nous centrons cette partie sur la façon dont les principes sémantiques définis précédemment peuvent se traduire dans la conception d'ontologies, et notamment d'ontologies de différents niveaux d'abstraction.

Les propositions de l'ensemble des auteurs qui se sont préoccupés de la question de la sémantique des langages et des outils du web de données ont structuré une ontologie. Le problème que pose un tel outil est qu'il s'agit d'une structure de connaissances, donc, comme on l'a vu, d'une représentation d'état, et non de processus, comme l'est l'interprétation. Les flux peuvent être alors considérés comme un modèle permettant de caractériser la sémantique dans le cadre de l'interprétation. Comme on a pu le voir, les flux étaient utilisés dans la dimension logique du web, essentiellement pour l'alignement d'ontologies. Nous avons également vu que les flux s'inscrivent dans le web de données dans la brique « logique ». Nous privilégions non pas tant la dimension logique mais les implications de ces représentations formelles sur des structures de données hétérogènes.
Pour cela, il faut considérer que l'interprétation est une suite d'inférences depuis une structure de données vers une autre, considérant l'hétérogénéité des deux. L'inférence ne serait pas produite dans les structures de données, mais entre elles, par un flux d'information qui permet d'enrichir la seconde structure par la première. Par exemple, si l'on décrit un document en lui associant des liens à une classe de la DEWEY et en utilisant des mots-clés issus d'un thésaurus quelconque, on n'enrichit aucune des descriptions par l'autre. Par contre, on pourrait considérer que dans le cas de ce document, la classe DEWEY peut être enrichie par les descripteurs : la classe s'applique à des documents pouvant être décrits par ces termes du thésaurus. Ainsi les descripteurs informent sur la classe elle-même et les connaissances qu'elle décrit. L'ensemble est interprété dans la situation d'un certain type de documents.

Nous sommes parti d'une hypothèse forte : les ressources numériques que l'on considère sont avant tout des outils, à savoir qu'il s'agit de connaissances inscrites dans de la matière (numérique dans notre cas). Simplement ces connaissances sont associées à des opérations réalisées automatiquement et qui se traduisent dans les données elles-mêmes. Le cas le plus simple de l'utilisation d'un outil est l'annotation.

L'ontologie que l'on propose trouve son unité dans la mise en relation classificatoires entre des structures ayant leur propre cohérence interne. Parler d'une seule ontologie est quelque peu fallacieux : chaque niveau de structure est autonome et les entités sont liées les unes aux autres par des principes classificatoires.

Cette hypothèse de structuration pose un certain nombre de problèmes techniques, que la seule référence aux standards de représentation ne permet pas de résoudre. Pour rendre compte de l'ensemble des relations, on utilisera SKOS, qui constitue le standard le plus opératoire pour lier deux structures de données hétérogènes.

6.4.1. Elargissement et convergence : vers des raisonnements automatisés sur le web.

La navigation à l'aide d'outils du web sémantique ne constitue pas le seul enjeu de l'articulation des outils de description d'information et de connaissance sur le web. L'objectif, plus précisément, consiste à proposer des raisonnements automatisés entre descriptions différentes de documents ou entre documents hétérogènes par le biais des conséquences de la description de l'un sur la description de l'autre. Cette question s'intègre dans celle, plus générale, de la place et de la formulation des raisonnements sur le web. Comme on l'a vu, la brique « logique » de la structuration du web se situe à un niveau d'abstraction qui se situe à un niveau encore plus abstrait qu'OWL. Or, relativement à tout ce que l'on vient de dire, on peut difficilement considérer le raisonnement comme une activité de haut niveau d'abstraction.
Néanmoins, conformément à la notion de niveau d'abstraction, on peut considérer que les raisonnements se déploient sur différents niveaux, c'est-à-dire qu'ils peuvent être représentés de façon distincte entre différentes formes de manifestation.

Cette question est essentielle, tout simplement parce qu'elle implique des choix à la fois techniques et intellectuels. En effet, on ne peut pas systématiquement associer la mise en relation de données structurées hétérogènes à un raisonnement. Par exemple, les équipes proches du Dublin Core proposent des relations entre structures des données fondées sur des protocoles FRBR. Ces relations sont sémantiquement faibles. Elles marquent des relations entre objets documentaires, mais ne caractérisent pas des inférences, à savoir des représentations du raisonnement humain.
Par ailleurs, on peut utiliser les mêmes langages de représentation pour caractériser des phénomènes différents : RDF et ses déclinaisons sont conçus pour être neutres quant à la nature des phénomènes représentés. Les raisonnements doivent donc être caractérisés en dehors des langages de représentation, par exemple en utilisant les logiques de description.

Dans le cadre du projet, nous utilisons une ontologie dans laquelle s'insèrent les flux. Ces derniers constituent le type de raisonnement qui permet de lier deux objets hétérogènes. Le raisonnement que l'on veut modéliser emprunte à la fois aux relations entre objets de recherche et aux processus de recherche.
Nous allons caractériser en quoi les flux constituent des outils permettant des représenter des inférences entre des entités. Ces entités ne sont pas seulement des concepts, mais d'une part, des représentations conceptuelles des documents, et, d'autre part, des entités extraites de ces documents eux-mêmes. Ces unités symboliques n'ont pas du tout le même statut.
Cette distinction se retrouve sur deux autres problèmes :

- d'une part la distinction entre le typage classificatoire tel qu'on l'envisage à propos des flux et les classifications telles qu'envisagées dans le cadre de l'organisation des connaissances,
- d'autre part le lien entre la prédication et les raisonnements tels qu'on peut les proposer dans le cadre des flux.

Cette question n'est pas seulement essentielle pour notre projet. A l'heure où des outils très différents sont disponibles dans les LINKED DATA et utilisant les mêmes langages de représentation, des confusions et des mises en relation abusives peuvent apparaître.

6.4.1.1. Classes et types.

Nous aimerions dans un premier temps revenir sur les classifications et la façon dont elles s'insèrent dans le cadre de notre projet. On doit articuler des classifications considérées dans les systèmes d'organisation des connaissances (classifications documentaires et thésaurus) et dans ceux d'organisation de domaines de connaissances (représentations de connaissances et ontologies).

Nous caractérisons maintenant les classifications utilisées dans le cadre des activités d'information. En effet, jusqu'à présent, un document était classé en vertu soit

- d'un système préexistant de classification, ou indexation pré-coordonnée (de type DEWEY), soit
- on établissait des appartenances et des distinctions en fonction d'un ensemble de descripteurs disponibles et organisés en fonction d'une certaine « sémantique ». L'indexation s'apparente alors à un ensemble de choix dans un système : il s'agit alors d'une indexation post-coordonnée.

Cette distinction doit être largement modulée, entre autre du fait des systèmes de classification originaux, et pouvant servir pour les thésaurus, comme la classification Colon de S. Raganathan et son système de facettes.

Ces organisations de connaissances sont fonctionnelles, au sens où elles permettent de structurer des collections de façon pérenne et sans ambiguïtés). Elles ne se caractérisent pas par un typage classificatoire mais par une appartenance dans le cadre des classifications, et l'adjonction d'une propriété descriptive dans celui des thésaurus.

Les types tels que nous les envisageons ne peuvent être qu'hétérogènes par rapport aux systèmes d'organisation des connaissances. Par conséquent aussi, ils peuvent être mis en relation à l'aide de flux, et représentés à l'aide de SKOS.

Le typage pourrait alors s'inscrire à un niveau d'abstraction supérieur par rapport aux organisations de connaissances. Les théories (telles qu'envisagées dans le cadre des flux) pourraient s'intégrer dans la brique « logique » de la structuration du web.

6.4.1.2. La question de la prédication : renouvellement de la perspective et application à l'aide de flux.

Pour bien comprendre l'enjeu de la question de la prédication, il faut revenir à la conception que les tenants des ontologies ont de la dimension linguistique. Elle est seulement considérée au travers des lexiques et des terminologies, alors que nous la considérons au travers d'une dimension discursive et interprétative, celle de la structure d'information.

De façon à mieux situer notre travail, nous commençons par le reformuler en utilisant un vocabulaire classique qui permettra de l'inscrire dans la continuité des projets élaborant des outils de type analyseurs fonctionnant sur des modèles linguistiques. Ces outils servaient ensuite aux projets d'indexation automatique, de recherche d'information et de synthèse. Ils constituaient les bases d'approches articulant des préoccupations de description linguistique et des objectifs documentaires. L'idée consistait à élaborer un analyseur qui permettait d'extraire des unités et des structures pouvant être pertinentes dans le cadre d'une activité documentaire.

La difficulté intrinsèque à ce type de projet (et sans parler des difficultés techniques), résidait dans l'hétérogénéité entre les opérations linguistiques réalisées dans la génération ou l'interprétation des textes d'une part, et les activités documentaires d'autre part. Dans un premier cas, on a des unités du langage interprétées dans un contexte, et de l'autre des unités descriptives des contenus. On ne se situait donc pas au même niveau d'abstraction et la relation entre mots du discours et termes, concepts et descripteurs est loin d'être explicite.

Cette difficulté, intrinsèque aux analyseurs, est contournée dans le cadre des outils du web parce que l'on ne cherche pas à analyser le document, mais seulement à en produire une description. Cette question est également présente dans les problématiques de l'annotation de documents. On ne cherche pas à extraire des unités ou des structures linguistiques, mais à produire des descriptions de ces discours, partiellement dans le cas de l'annotation, en totalité dans celui des métadonnées. Ainsi, l'usage des concepts de l'analyse du langage est différent : les modèles de la prédication, binaires et ternaires, sont intégrés dans des syntaxes de renseignement, par lequel l'usage attribue des valeurs à des entités prédicatives prédéfinies.

Cette dimension linguistique ne disparaît pas ; elle est néanmoins limitée ou paramétrée dans le cadre des métadonnées comme le remarque T. Baker : « Resource description is inherently linguistic in nature. Over the past decade, the bibliographic and subject standards used in the library world for bibliographic control – in Svenonius's terms, its languages of description – have progressively been translated into the language of the semantic web and linked data"[278]. Ainsi, on considère que les phénomènes du langage naturel qui nous intéressent sont traduits dans les langages de représentation du web.

La prédication (que l'on considère comme une expression logique de la classification) constitue une forme de raisonnement permettant de caractériser un document par la possibilité, ou pas, de lui attribuer une propriété. En ce sens, la prédication est une articulation logique entre un objet document et les outils symboliques que l'on se donne pour le décrire. Les flux permettent de relier l'ensemble des possibilités de prédication sur un objet en vertu de la contrainte que cette prédication devait donner à connaître un état de l'objet.

Ces propositions sont associées à un processus à fondement documentaire, mais dont l'importance dépasse depuis ce cadre ; il s'agit des métadonnées. Nous reviendrons sur ces outils essentiels rapidement puisqu'ils constituent les moyens dont disposent les producteurs d'information (et les éditeurs) pour décrire des propriétés associées aux documents qu'ils diffusent. Les thésaurus et autres langages documentaires étant des moyens utilisés par les médiateurs pour structurer l'offre documentaire, mais également, via les utilisateurs, pour produire de nouvelles mises en relation.

L'articulation que l'on vient de proposer présuppose des outils dont la fonction est d'effectuer la circulation de l'information. On considère la prédication dans le contexte des contraintes fixées par les réseaux et systèmes de production, circulation et réception de l'information. On intègre dans ce travail le rôle contraignant des technologies de ces systèmes sur la description des documents, ces contraintes étant aussi techniques (questions de format, de compression) que relatives aux contenus (langages de représentation notamment). L'ensemble des métadonnées (DUBLIN CORE mais également les métadonnées plus spécialisées comme IMDI pour les corpus de linguistique) peuventt être considérés comme des choix de prédication. En ce sens, ces jeux sont signifiants et transmettent une information.
Ces contraintes sont le fait des médias de transmission (autrement dit des règles communes associées à chaque jeu de métadonnées), ce qui permet de spécifier une approche de la prédication distincte de celle qui est pratiquée en linguistique. (Le problème se pose dans les mêmes termes concernant les sciences de la communication, pour lesquelles un discours est saisi dans des termes et avec un contexte distinct d'une approche linguistique traditionnelle).

Dès lors que l'on postule des contraintes régulières, en partie techniques, permettant la transmission d'un contenu, on obtient la caractérisation de formats d'information ; comme ces

contraintes sont régulières, elles peuvent être groupées sous l'appellation de contraintes de flux. Alors ces formats seront les résultats des contraintes des flux. La dimension des réseaux et systèmes spécifie la prédication : cette dernière est contrainte par la transmission de l'information. Prenons, un exemple concret ; un document renseigné par des métadonnées est classé par celles-ci ; dans le cadre d'une recherche d'information, cette métadonnée renseignée sera acceptée comme le résultat d'une inférence relativement au type de contenu du document recherché. Ce dernier type de contenu spécifie alors le canal par lequel au type de la requête correspond le type de la métadonnée inférée.

Enfin, nous insistons sur le fait que si notre projet ne traite pas directement la question de la recherche d'information, c'est en partie parce qu'il nous apparait primordial de mettre en évidence une propriété spécifique des flux immédiatement exploitable, à savoir la mise en relation de données hétérogènes.

6.4.1.3.. Processus, domaines et outils

Un outil constitue une construction cognitive et matérielle fonctionnelle qui puisse être adapté à des contextes différents et repose sur un modèle caractérisant les propriétés associées à lui. En ce sens, les propositions de G. Hutchins et de l'ensemble des auteurs liés à l'anthropologie cognitive seront utiles pour structurer ces lexiques parce qu'elles vont ordonnancer les données en fonction des processus à l'œuvre. Ainsi, on pourra construire une structuration dynamique des phénomènes à l'œuvre.

Néanmoins, si ces mondes sont extérieurs matériellement à l'activité et que les phénomènes qui s'y déroulent ne sont pas prédictibles, ils sont néanmoins structurés par les acteurs. D'Andrade (op. cit. p. 795) considère que le premier apport de l'anthropologie aux sciences cognitives réside dans la structuration des domaines comme constructions sociales permettant à une population de construire un rapport spécifique au monde explicable au travers de ses activités.

6.4.1.4. Quels sont les produits disponibles à la suite de ce travail ?

Notre projet a quelques conditions de succès à partir du moment où il rencontre les intérêts d'acteurs du domaine, lesquels sont relativement en concurrence, comme on a pu le voir à l'intérieur des infrastructures et des plateformes.

Dans la mesure où l'on propose un outil, représenté par un schéma, nous n'aurions en fait qu'à nous préoccuper de la diffusion de ce schéma auprès des bibliothèques numériques, comme le font par exemple le BIBO ou le DC.

Néanmoins, on se doit de caractériser la portée de la fonctionnalité de l'outil, à savoir l'enrichissement de métadonnées à l'aide de la mise en relation ce celles-ci. Ainsi, par rapport à l'offre existante de modèles de mise en relation, notre proposition se distingue par le fait que la mise en relation s'accompagne d'opérations sur les contenus. Autrement dit, la mise en relation ne constitue pas pour nous une finalité, à la différence de l'enrichissement.

On peut considérer, comme on l'a fait jusqu'à présent, que les jeux de métadonnées constituent les destinataires les plus immédiats de notre travail. En effet, le schéma proposé pourrait s'intégrer dans les jeux existant.

Néanmoins, cette application immédiate ne concerne qu'indirectement les bibliothèques numériques. Les flux, fondés sur la mémoire d'un web service, permettent de concevoir une application très générale, la structuration des collections en vertu des relations enregistrées entre types de documents.

Un tel outil aurait le mérite d'éviter le flou des classifications recourant aux mots-clés déterminés par les auteurs et à la difficulté à utiliser les outils documentaires pour d'autres objets que les publications.

Nous avons présenté rapidement le modèle de recherche d'information élaboré par P. Ingwersen. Nous pouvons maintenant montrer en quoi l'adoption d'un cadre théorique différent relatif à la cognition modifie ce modèle de la recherche d'information. Cette transformation permet également de proposer une adaptation pour la modélisation de la recherche d'information sur le web. En effet, le problème de la quantité d'information obtenue amène un certain nombre d'acteurs à proposer des outils de filtrage reposant, par exemple, sur une extraction d'information à partir des résumés des publications associées à des facettes d'interrogation[279]. (Ce projet est intéressant parce qu'il permet à ELSEVIER de proposer un moteur de recherche bien plus précis que les moteurs de recherche généralistes). Ainsi, la nécessité d'une description plus fine des contenus apparait comme un point de convergence entre les bibliothèques numériques et les éditeurs.

6.4.2. Modèle sémantique et structures conceptuelles.

Les structures conceptuelles sont caractérisées par leur dimension intensionnelle. Néanmoins, ce positionnement ne permet pas de répondre à la question du lien nécessaire entre ontologie et monde. On vise ici à clarifier le lien que l'on établit entre la sémantique associée aux structures et aux flux (plus globalement à l'information), et celle qui d'une part est associée aux ontologies et d'autre part aux langages de représentation de ces ontologies.
Cette question peut sembler à priori éloignées des questions liées à un projet. Néanmoins, l'articulation entre des structures informationnelles et conceptuelles constitue un problème nettement plus important qu'un lien entre des modèles de représentation de données.
De plus, aux questions traditionnelles de détermination entre langage et pensée, s'adjoint celle de l'activité.
L'hypothèse dominante a longtemps été celle de l'internalisme, à savoir le fait que les structures conceptuelles préexistent à l'expérience, et donc peuvent être représentées indépendamment de celle-ci : « the view that social facts about public language meaning are derived from facts about the thoughts of individuals, and that these thoughts — and hence, on this picture, also indirectly facts about public languages — are constituted by properties of the internal states of agents"(J. Speaks, p. 429)[280].
L'externalisme propose au contraire une caractérisation des états mentaux par des relations entre agents en interaction et en lien au monde : « Externalism: facts about the contents of the beliefs of agents are partly determined by relations between those agents and facts external to them. »

La question pour nous consiste à caractériser de quelles façons les phénomènes informationnels (externalisés) et les structures de connaissances (notamment les ontologies, qui représentent fondamentalement des phénomènes internes), peuvent être articulés de façon fondée. Cette question amène à se pencher de façon plus précise sur la nature des ontologies que l'on cherche à caractériser.

6.4.2.1. Positionnement du problème.

Traditionnellement, les structures conceptuelles sont considérées comme des univers d'interprétation. On caractérise comme entité intensionnelle les énoncés modaux, les conditionnels, et "tout ce qui aurait pu avoir lieu si les conditions auraient été autres".
L'idée, défendue notamment par Cresswell, consiste à dire que le langage naturel est fondé sur des ontologies qui fondent la dimension intensionnelle du langage.

La nature même des entités linguistiques est d'être intensionnelle. De cette façon, les états mentaux ou structure cognitive constituent le centre de l'élaboration de modèles sémantiques formels. (Cf. Cresswell, Stalnecker, etc.)

A l'opposé, on assiste au développement des théories fondées sur la référence directe, pour lesquelles le rôle du langage est avant tout de désignation. On inverse la perspective par rapport à Frege : on considère la conceptualisation comme un recours lorsqu'une désignation directe est impossible.

Ainsi, dans le cas des descriptions définies (" le chat "), on n'a pas un concept limité par un déterminant, mais une désignation utilisant un concept pour signifier.

Ce cadre permet facilement de distinguer autour des unités linguistiques les dimensions intensionnelles et extensionnelles, et par conséquent les structures conceptuelles.

Le problème est légèrement différent lorsqu'il s'agit de représenter directement des structures conceptuelles sans considérer qu'il s'agit d'entités linguistiques. Une telle proposition explique le fait que l'on puisse disposer de structures hétérogènes permettant de représenter les domaines de connaissance, alors que dans le cadre de la langue, les univers intensionnels ne sont pas structurés autrement que par des relations entre entités lexicales ou entre classes de mots ayant une distribution similaire. FRAMENET constitue une exception particulièrement intéressante, tout comme la théorie des situations utilisée par les HPSG. Par ailleurs, la dimension linguistique, en tant qu'elle caractérise la langue, ne peut spécifier des structures fonctionnelles, associées à un usage, comme le sont par exemple les thésaurus.

Cette distinction entre ces deux acceptions de ce que sont les structures conceptuelles n'est pas inutile : elle permet d'envisager les ontologies au travers de plusieurs niveaux de structuration, relativement à la prise en compte ou pas des unités linguistiques.

Cette clarification permet également de prendre position sur des questions récurrentes des ontologies, à savoir le rapport aux unités lexicales et terminologiques.

Deux positions existent relativement aux liens existant entre ontologie et langage : soit on considère des structures purement conceptuelles, et alors une ontologie constitue une structure qui n'a pas de lien à la langue (ce qui est le cas des ontologies théoriques comme les propositions de N. Guarino et son équipe), soit au contraire on propose des ontologies de domaine, liées à un domaine de connaissances, et à ce moment-là en lien avec le vocabulaire du domaine : les propositions de B. Smith se situent dans ce cadre. Elles opposent conceptualisme et réalisme. Pour le conceptualisme, les terminologies dérivent de structures conceptuelles préalables, et constituent des réalisations linguistiques d'unités de pensée. La signification d'un terme est un concept, une construction de la pensée.

Pour le réalisme au contraire, un concept est une représentation mentale d'un terme qui dénote une entité du monde. Si les concepts sont des unités psychologiques, elles se caractérisent par un emploi dans le cadre des échanges entre experts du domaine. Elles ont une dimension linguistique, laquelle dénote effectivement des phénomènes du monde. Le réalisme s'appuie fondamentalement sur des expériences dans le domaine biomédical ; les universaux sont des objets du monde.

Notre approche intègre d'autres niveaux de réalité : elle prend en compte des outils existant (les métadonnées, ce que ces dernières représentent) tout autant que le vocabulaire d'un domaine. Nous utilisons des structures déjà constituées et organisées par rapport à un usage. Même s'il s'agit d'un même domaine, ces différents outils relèvent de connaissances et de raisonnements appartenant à des compétences scientifiques et techniques relativement différentes.

L'hétérogénéité se fonde sur cette distinction entre des unités qui relèvent de domaines, de compétences et d'usages différents. L'intérêt de l'approche anthropologique cognitive se manifeste ici. A l'intérieur d'un domaine donné, les outils manifestent des connaissances propres, qui ne sont pas nécessairement répertoriées dans le vocabulaire de ce domaine.

Généralement, les domaines de connaissance sont établis sur des bases essentiellement lexicales, considérant des connaissances déclaratives. Par contre, les flux dans leur généralité posent les fondements d'une approche fondée sur l'hétérogénéité des données. Plutôt que de considérer un objet unique et homogène (un texte, une langue, etc.), les flux rendent possible l'observation de phénomènes hétérogènes, mais reliés entre eux par des inférences particulières. En ce sens, toute structure est un point de vue, une représentation d'un domaine qui ne l'épuise pas.

L'originalité de chaque ontologie de domaine explique le relatif insuccès des bibliothèques d'ontologies. En effet, mis à part relativement au haut niveau (« upper-level »), les structurations de domaines font peu apparaître de structures communes ou ré-employables. Comme nous l'avons noté dans notre première partie, les bibliothèques d'ontologie ne sont véritablement actualisées et exploitées qu'à partir du moment où soit elles concernent un domaine précis, comme le domaine médical, soit parce qu'elles accueillent une ontologie de haut niveau. Ainsi, des outils comme la BFO (Basic Formal Ontology), DOLCE, voir même CyC constituent des outillages conceptuels communs à de nombreuses ontologies. Plus on entre dans la description des domaines et surtout dans des points de vue différents, moins les expériences peuvent être partagées.

Enfin, les langages du web offrent la possibilité d'autres écritures du rapport entre un document et une classe ; en témoigne par exemple le renouveau des classifications à facettes, mais également l'essor de la FCA et des logiques de description. En ce sens, on essaie d'intégrer à l'intérieur des problématiques documentaires les outils logiques de théories de types.

L'Analyse Formelle de Concepts (Formal Concept Analysis), étendue par l'Analyse Relationnelle de Concepts (Relational Concept Analysis) constituent une famille d'outils permettant d'établir des relations entre les ensembles de données structurées. Si elles offrent les mêmes possibilités de mise en relation que les flux, elles ne permettent pas de caractériser un raisonnement et donc le transfert de contenus. En ce sens, elles n'offrent pas les mêmes possibilités que les flux. Nous expliquerons cela par l'absence d'une théorie de l'information

dans les propositions des FCA, et à l'opposé, le fondement informationnel de la théorie des flux.
En définitive, elle ne travaille pas à un niveau d'abstraction qui permette de caractériser l'information.

6.4.2.2. Conséquences de l'externalisme sur la caractérisation du document et des ressources.

Nous aimerions maintenant montrer quelle sorte d'ontologie il est possible de construire à partir des ressources scientifiques. Cette ontologie a comme domaine la description des documents et donc l'ensemble des propos qui peuvent être tenus sur des documents. Ainsi considéré, le domaine représenté est celui des descripteurs d'objets documentaires. Il se caractérise par son hétérogénéité : les descriptions étant fondées sur des facettes, elles considèrent le document par la diversité de sa réalisation.

Néanmoins, on ne peut se limiter à cette caractérisation. La description étant toujours à propos d'un objet, celui-ci est nécessairement représenté dans la structure. On peut caractériser l'unité d'un document par sa singularité. En d'autres termes, si l'on regarde les feuilles d'adaptation, le seul élément caractérisant une unité au document est la référence à un individu unique. On peut donc envisager de caractériser par la référence directe, celle de l'individu et du nom propre, l'unité du document. Le reste peut être effectivement considéré comme étant n'importe quelle information relative à cet objet et cette représentation.

Les principes de la structure d'information peuvent alors être utilisés pour construire les bases de l'ontologie. Ils mettent en évidence un principe de structuration de la référence qui peut être repris dans le cadre de la structuration de l'ontologie en niveaux.

En vertu de tout ce que l'on a dit précédemment, on pourra définir de façon structurale le document comme ce qui rend possible une circulation d'information en se caractérisant comme artifice de représentation (au sens de G.Hutchins).

Ce document peut être parcouru par un outil (de type annotation notamment) ce qui en fait une ressource. En tant que ressource, n'est donc plus seulement informatif à propos de l'individu auquel il réfère, mais par rapport aux fonctionnalités et aux usages qu'il permet.

L'idée générale que l'on a retenue est qu'il est pertinent de structurer un domaine scientifique du point de vue de ses ressources. Les ressources sont ainsi définies comme des objets individuels structurés pour un usage défini.

Du point de vue plus général de l'e-science, ce ne sont pas des données qui sont rendues disponibles, mais des données traitées de façon à pouvoir servir de ressources. Par ailleurs, les outils qui servent à traiter ces données sont disponibles, ce qui permet de relier ces entités et ainsi de structurer des collections.

Par conséquent, nous pourrons structurer les ontologies en reprenant ces idées-là. D'un point de vue réaliste, les ontologies sont fondées sur des taxonomies et des terminologies (on reviendra sur ces questions ultérieurement) qui caractérisent comment des mots réfèrent aux

choses. Par conséquent, un effort doit être porté sur la caractérisation des objets. Dans notre cas, ces ressources linguistiques sont des outils et peuvent être définies par leur usage dans le cadre de recherches. Ce ne sont pas des objets observés par la science, mais des objets construits pour faciliter (voire même rendre possible) le travail scientifique.

Ces ressources ont une structure, ils ont aussi des fonctions et s'inscrivent dans des processus. Une structuration des objets autour de domaines distincts a déjà été tentée par la GENEONTOLOGY[262]. Les propriétés des produits génétiques ne peuvent être décrites par un seul domaine : les composants cellulaires caractérisent les parties d'une cellule ou son environnement externe, la fonction moléculaire décrit les activités élémentaires d'un produit génétique au niveau de la molécule et le processus biologique les opérations ou les ensembles d'événements moléculaires. Ces derniers sont caractérisés par un début et une fin, et caractérisent le fonctionnement d'unités vivantes intégrées : cellules, tissus, organes et organismes.

Que plusieurs points de vue apparaissent complémentaires pour caractériser un objet évolutif, vivant, a une conséquence essentielle : les relations entre les concepts demandent à être spécifiées pour cette représentation, à savoir qu'il a été nécessaire de produire une ontologie de relations. Ces relations sont différentes des propriétés d'objets.

Les relations caractérisent chaque lien possible entre deux entités. La collection de ces relations, valable pour la GENEONTOLOGY, est disponible dans la bibliothèque d'ontologies d'ONTOBEE[263]. Ces relations sont considérées comme des propriétés d'objets.

Il existe de nombreuses ontologies permettant de décrire des processus, et notamment des tâches impliquant des objets[264]. Néanmoins, notre propos est légèrement différent puisqu'il s'agit de structurer des catégories descriptives d'objets scientifiques qui présupposent une activité.

Notre ontologie s'inscrit dans le cadre des descriptions de documents, sauf qu'il s'agit ici de dépasser le seul cadre bibliographique pour caractériser les opérations effectuées sur ces objets et les usages prévus de ces objets.

On explicite l'activité de production d'une ressource à l'intérieur d'un haut niveau d'abstraction. Le niveau médian caractérise la façon dont la ressource est décrite. Dans le bas niveau d'abstraction, on peut envisager une structuration autour des opérations qui transforment l'objet numérique. Ainsi, la ressource est caractérisée par un processus dont elle est un aboutissement, par un contexte qui la décrit et par la succession des opérations internes qu'elle subit de façon à être opérationnelle.

On envisage ainsi un autre mode de description des ressources : en deçà des métadonnées, on trouve des marques de traitement qui à première vue font penser à des annotations de type TEI. Sauf que notre objectif consiste à décrire ces annotations de façon à rendre l'objet accessible aux moteurs de recherche.

---

[262] http://www.geneontology.org/GO.doc.shtml

[263] http://www.ontobee.org/browser/term.php?o=RO&iri=http://www.w3.org/2002/07/owl%23ObjectProperty&graph=http://purl.obolibrary.org/obo/merged/RO

[264] http://www.ecolleg.org/trms/ontology.html

6.4.2.3. Intégration de la dimension de l'activité dans l'ontologie.

Comme on l'a déjà évoqué, en termes d'ontologie, deux postulats (réalisme et conceptualisme) s'opposent. Chaque postulat a donc proposé un modèle de « upper ontology ». Pour le conceptualisme, il s'agit de DOLCE et pour le réalisme de BFO. Au niveau des relations, chacun exprime différemment la façon dont peut lier deux entités (concepts et instances).

Si l'on observe de façon un peu plus précise, on s'aperçoit que les « upper ontology » ne sont pas fondamentalement indépendantes de tout domaine. Autant la BFO, sous l'impulsion de B. Smith, a un fondement biomédical, autant DOLCE se développe en intégrant l'activité économique.

L'équipe de DOLCE, et principalement Borgo, vont s'intéresser à l'activité. Ils reprennent le travail fondateur de Chandrasekaran, lequel vise essentiellement à construire un modèle inférentiel de l'activité professionnelle humaine.

Il existe aussi une différence importante entre les deux modèles, considérant ce que l'on peut considérer comme étant des concepts. Dans la BFO notamment, les concepts représentent des phénomènes de type événementiels. Nous pouvons reprendre les propositions de B. Smith : « The terms 'continuant' and 'process' thus correspond to what, in the literature of philosophical ontology, are known respectively as 'things' (objects, endurants) and 'occurrents' (activities, events, perdurants) respectively. A continuant is what changes; a process is the change itself. The continuant classes relevant to biological ontologies include *molecule,cell, membrane, organ;* the process classes include *ion transport, cell division, fat body development, breathing. » (*[281] p.4)

Il s'agit là de la plus grande différence avec les documents et les descriptions bibliographiques. Les activités humaines constituent des plans et des actes qui ne sont pas liés à des objets en extension. Par contre, les objets que l'on décrit sont systématiquement associés à des traces matérielles (même si elles sont numériques). Les termes décrivant des actions sont absents de la désignation d'objets.

Techniquement il n'est pas interdit de modéliser l'ontologie selon les principes de l'un des outils puis de représenter les concepts et relations en utilisant les outils de l'autre.

Les bibliothèques d'ontologies constituent dans ce cadre des outils extrêmement importants. On peut les caractériser de deux façons : celles qui sont fondées sur des ontologies de haut niveau, et qui donc donnent un aperçu des applications possibles (c'est le cas notamment de la BFO et de la bibliothèque OBO[265]) ou des collections comme celle proposée par VIVO [266].

[265] http://www.obofoundry.org/index.cgi?show=mappings
[266] http://swl.slis.indiana.edu/repository/index.html

Notre ontologie est fondée sur la volonté de décrire des relations dynamiques entre des concepts qui représentent eux-mêmes des parties ou des phénomènes apparaissant à l'intérieur de documents.

Les relations entre les blocs descriptifs de traits de la ressource peuvent être caractérisées par des concepts indiquant des processus et des actions. Quel que soit la qualité de la représentation de l'action dans l'ontologie de départ, elle ne permet pas de caractériser une ontologie dynamique.

Comme nous l'avons évoqué, notre modèle du haut niveau d'abstraction de l'ontologie est fondé sur la notion d'activité. Nous l'empruntons en grande partie à G. Hutchins ; néanmoins, d'autres modélisations, notamment dans le cadre des ontologies, ont été élaborées. L'une des plus précises est celle de R. Ferrario et N. Guarino[282]. Elle postule deux niveaux (p.4/13) : "an *abstract* level, where functional capabilities find their place, and a *concrete* one, where the functionalities are *realized"*. R. Ferrario et N. Guarino pensent que le niveau abstrait gagne à être décrit comme un engagement (« commitment ») et non la seule description d'une suite d'opérations. Au niveau le plus abstrait, un service est considéré comme une suite d'événements.
La différence entre un service et le cadre de l'e-science réside dans la distinction des acteurs et l'impossibilité de prédire qui va utiliser telle ressource et pouquoi. La notion de distribution et d'hétérogénéité apparait ainsi plus fondamentale : l'usage prescrit d'une ressource permet de prédire sur elle mais n'épuise pas ses différents usages possibles.

Une autre conclusion peut être déduite de la distinction entre la modélisation de services et celle d'activité de recherche : non seulement l'activité est distribuée entre des acteurs différents mais également les finalités de ces différents acteurs peuvent être très éloignées.

On peut déduire de tout cela le fait que l'on ne peut pas calculer avec certitude la totalité des usages d'une ressource (voire d'un type de ressources). On ne peut pas contrôler les usages à l'aide d'un modèle de l'activité de recherche. Par contre on peut indiquer ce quoi a pu servir une ressource et pourquoi. Cette idée justifie l'extraction d'information et la construction d'un web service intégrant la mémorisation des données et des relations. Il est alors possible d'associer des calculs (fréquence, poids, etc.) à ces mémoires, de façon à spécifier plus précisément ces usages. Ainsi, on enrichit l'ontologie.

6.4.2.4. Place et rôle des bases de données dans les ontologies. Spécificités d'une approche par l'information vis-à-vis des contextes usuels de développement d'ontologies.

Nous considérons l'ontologie essentiellement sous la forme de structure permettant de lier des descriptions, à savoir des unités symboliques. Or la caractérisation d'ontologies dépend aussi des applications visées. Ainsi, on peut se servir des ontologies pour structurer des bases de données, et notamment des relations entre des bases de données différentes. On entend alors des ontologies relativement statiques, parce qu'elles sont fondées sur des stocks constitués de données. Il peut s'agir de bibliothèques numériques comme DBLP, ou d'ontologies de domaines comme, GOLD en linguistique.

En envisageant par les ontologies des relations entre des concepts, on adopte un tout autre point de vue. L'ontologie sert à établir des relations entre des descriptions de documents, qui sont eux stockés par ailleurs. L'ontologie permet ainsi de caractériser des inférences entre structures de concepts, et donc ne s'inscrit pas dans la structuration d'une base de données. Elle s'inscrit dans des structures descriptives de données et des informations extraites, ce qui implique de définir l'ontologie comme la caractérisation de relations entre des descripteurs (et les descriptions qu'ils proposent).

On propose donc une ontologie fondée sur des descriptions, et non directement sur une base de données.

Néanmoins, il ne s'agit pas là du rôle fondamental de l'ontologie : si l'on raisonne en termes de fonctionnalités, l'ontologie sert à structurer les concepts du domaine de façon à ce que l'on puisse ensuite effectuer des inférences valides. Le fondement de l'ontologie sur des descriptions entraîne le fait que l'on traite essentiellement des structures symboliques. De par les flux (à la fois dans la dimension de la signification par le typage et le transfert d'information), on introduit dans les ontologies des dimensions de sémantique linguistique et de transfert d'information au travers d'un raisonnement.
Pour cela, on aimerait faire quelques commentaires sur la portée des exemples que l'on a choisis et surtout en quoi ils peuvent être caractérisés aussi dans une problématique de représentation des connaissances et d'ontologies. En effet, les questions de connaissances (les propositions de G. Hutchins s'inscrivent dans le cadre des Sciences Cognitives) sont considérées au travers de dispositifs et donc de connaissances procédurales. L'élaboration d'une ontologie fondée sur les principes des flux introduit des dimensions procédurales à l'intérieur d'un cadre de travail jusque-là considéré comme fondamentalement déclaratif.

Les flux peuvent être caractérisés comme un raisonnement. En d'autres termes, leur caractérisation en situation naturelle requiert des outils d'observations tels que ceux mis en place par G. Hutchins notamment.

Les flux contribuent à contextualisation du document parce qu'ils permettent à la fois de caractériser des structures internes du document comme les structures d'information, et des liens entre le document et d'autres objets structurés comme les données primaires, ou les publications en sens inverse. Il peut être effectivement tentant de connecter ces deux applications des flux. Ce qui importe ici est l'articulation entre

- les structures stables qui explicitent une dimension des contenus, ou plus précisément un format à l'information (à savoir les structures d'information), et

- les structures permettant la circulation de cette information (à savoir les flux entre descriptions de documents).

Dans le premier cas, la structure d'information constitue un lien entre des traits descriptifs hétérogènes, dans le second les flux permettent de connecter des documents hétérogènes.

Paradoxalement, sur le web, c'est la seconde perspective qui apparaît la plus immédiatement accessible. Effectivement, elle prend appui sur des données déjà largement structurées.

Conclusion.
L'articulation entre les deux perspectives peut s'envisager autour de la connexion entre les flux comme représentation de raisonnements entre documents hétérogènes, et les structures d'information comme composants des documents et permettant d'identifier l'origine des entités symboliques (c'est le produit d'opérations menées dans les laboratoires, au travers de procédures très distinctes, loin à la fois du patient référent comme du pharmacien, qui est inscrit sous le nom du patient et de son traitement). Cette articulation peut également structurer des relations entre des corpus linguistiques, des articles de linguistique et la façon dont les publications utilisent ces corpus.
Cette articulation permet de distinguer très concrètement entre annotation et métadonnées, et de les lier. Il s'agirait ainsi de connecter les perspectives des métadonnées et celles de la TEI.

Toutes les entités qui sont créées sur le web, et notamment ces unités abstraites que sont les URI, ont comme avantage de ne pas être contraintes par le monde des discours et des textes puisque l'on travaille essentiellement sur leur description. Or cette distinction d'univers ne pourra pas être éternelle tout simplement parce que l'accès aux contenus du document constitue un enjeu essentiel pour l'avenir du web.

La clé de cette articulation entre contenu structuré des documents et description des documents, au-delà des questions techniques, c'est la façon dont le discours et les différentes formes de la représentation des connaissances (classifications, thésaurus, ontologies notamment) peuvent être mises en relation.

6.4.3. Contraintes à la mise en relation de concepts avec des unités symboliques

La dernière partie de cette présentation a trait de nouveau aux questions de concepts, et plus particulièrement aux relations que l'on peut établir entre les différents niveaux d'abstraction et l'écriture des métadonnées. En d'autres termes, comment passer d'une structure hiérarchique à une structure prédicative.
Cette question sous-tend l'articulation entre une sémantique des unités conceptuelles et leur traduction dans une sémantique fondée sur l'interprétation.
Nous montrons maintenant de quelle façon les questions que nous nous posons relativement à la sémantique commencent à se poser à l'intérieur des problématiques d'ontologies.
En effet, nous voudrions d'abord montrer que les fondements des ontologies dans l'IA éludent les questions de sémantique interprétative. Ensuite, nous montrerons que ces questions redeviennent d'actualité dans le cadre du développement des ontologies. Enfin, nous montrerons ce que peut apporter sur le sujet une sémantique interprétative.

La multiplication des représentations de connaissances, ontologies, terminologies, etc. qui sont connectés au sein du web de données ne peut qu'entraîner des confusions importantes, parce que la définition de la sémantique associée à ces outils est fondamentalement différente. Dès lors les unités symboliques se voient dotées de statuts et de rôles sémantiques très différents à l'intérieur du web de données. Comme on l'a vu, la dimension interprétative requiert un usager au moins au titre de position dans le dispositif.

Elle oblige le web de données à sortir d'une neutralité jusqu'à présent avantageuse parce qu'elle évite de prendre en compte la notion d'objets de référence et de contexte. Or, une telle position ne peut être tenue que tant que l'on ne se préoccupe pas de l'activité menée par les usagers.

Néanmoins, nous ne proposons pas une remise en cause de la structuration du web. Nous proposons simplement d'insérer dans les outils du web des aspects qui jusque-là n'ont pas été pris en compte : l'interprétation des discours, le transfert d'information (comme composant de la communication), le raisonnement dans le cadre d'une activité.

la notion de niveau d'abstraction nous apparait comme l'une des plus pertinentes pour caractériser ces distinctions. La structure d'information, vue comme l'articulation logique d'entités de niveau d'abstraction différent, pourrait justement contribuer à caractériser l'information portée par la relation entre des entités de niveaux différents.

Fondements de conception hérités de l'IA.

Si l'on suit les langages du web ou par exemple l'ontologie bibliographique DBLP, mais également METASHARE[283], les ontologies servent à structurer les relations entre des ensembles de métadonnées hétérogènes. Dans le cadre bibliographique, les ontologies constituent un niveau relationnel abstrait qui sert à lier des éléments descriptifs, considérés comme moins abstraits, et utilisant souvent des vocabulaires contrôlés. Cette conception des outils est assez proche de celle des théoriciens des ontologies, comme Barry Smith.

L'IA distingue conception et élaboration ; la conception constitue un processus d'ingénierie qui est lui-même inscrit dans le cadre d'une élaboration. Celle-ci comprend des phases d'analyse, de spécification et enfin de conception. Cette distinction est très largement partagée dans le cadre des ontologies, structuré autour la distinction entre ontologie générale, qui regroupe des outils d'élaboration, et ontologie de domaine, qui caractérise la conception d'une ontologie relative à un domaine précis.

Pour l'élaboration, un certain nombre de méthodologies, ou boîtes à outils, ont été conçues, d'abord en IA (KOD, KADS) puis aujourd'hui du côté des ontologies (DOLCE, ONTOWEB, SWED notamment, et toutes les ontologies groupées autour de la BFO). Elles fournissent des règles communes permettant de réaliser des outils en garantissant à la fois homogénéité et rigueur. Si elles peuvent être critiquées du fait de leur généralité même et des contraintes qu'elles posent à des situations particulières (notamment par exemple, pour élaborer des ontologies de domaine), elles fournissent un cadre de référence. Ainsi, les ontologies montrent une certaine unité théorique du fait d'un recours à de mêmes références non seulement bibliographiques, mais également d'outil de travail. Cette unité repose également sur le principe du recours à des outils lexicaux et terminologiques, utilisant évidemment les standards du web.

Or, dans ce cadre relativement consensuel, un problème se pose : celui d'une ontologie de l'information. Il s'est également posé lors de la caractérisation de liens entre les perspectives d'ontologies et les dimensions linguistiques. Les travaux d'A. Gangemi[284] visent effectivement à caractériser cette relation, notamment lorsque les ressources linguistiques utilisées comportent une dimension sémantique, comme FRAMENET notamment.

Représentations conceptuelles et ontologies. Caractérisation de la structure d'information dans le cadre d'ontologies.

Nous aimerions explorer une ontologie, qui a été mise en place à l'intérieur du cadre d'OBO. Il s'agit de l'Information Artifact Ontology (IAO), dont l'objectif consiste à représenter des données scientifiques à des niveaux plus précis que ceux caractérisés par les ontologies de domaine. Dans sa présentation au workshop fondateur de IEO (Information Entity Ontology), Alan Ruttenberg[285] distingue trois niveaux de caractérisation de la communication scientifique :

« Record level: Represent database records. Inconsistent if two sources disagree about contents of a field.

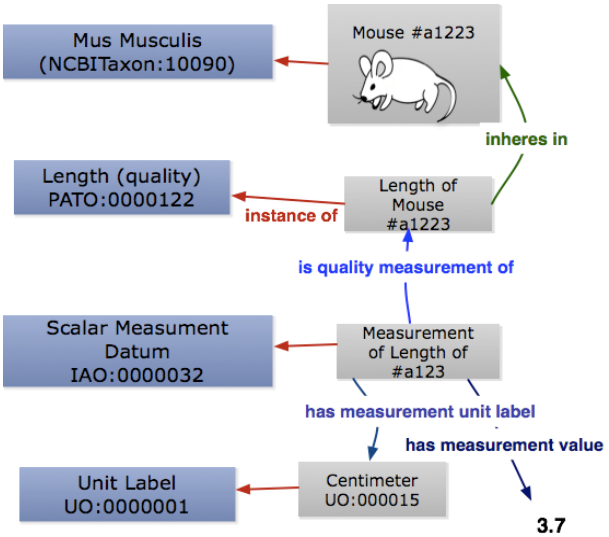
• *Statement level*: Represent what researchers say. Inconsistent if two people disagree about what a paper said

• *Domain level*: OBO Foundry approach. Represent your best understanding of consensus. Inconsistent if facts contradict. » (P. 4)

Le projet d'IAO consiste à caractériser par une ontologie les contenus informationnels. A l'origine, OBO (ontologie des produits biologiques, dont dérive la GENEONTOLOGY) caractérise des phénomènes naturels. Or, ces phénomènes sont issus de protocoles d'analyse, donc d'outils, qui sont des phénomènes artificiels. Ruttenberg [286] propose alors de caractériser un « plan de spécification » et un « objectif de spécification » afin de caractériser le passage des unités du monde vers la façon dont elles se réalisent dans l'univers, chez un patient par exemple. Ces plans et ces objectifs nécessitent un processus qui lui-même peut être modélisé. (Ce modèle a permis l'élaboration de l'ontologie OBI -Ontology of Biomedical Investigation). Il est donc cohérent d'envisager une représentation des résultats de ces analyses.

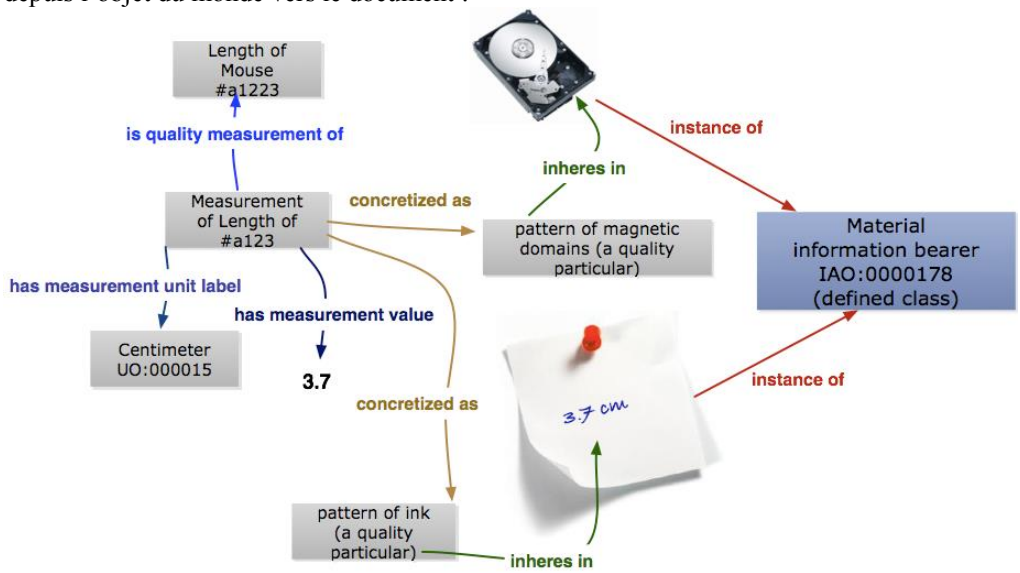
Notre caractérisation de la structure d'information fondée à la fois sur la structure d'information et les flux n'est pas la seule proposition de modélisation qui ait pu être formulée. L'une des plus intéressante est celle proposée dans le cadre de la BFO et de l'une de ses extensions, l'OBI -Ontology of Biomedical Investigation. A partir du moment où les processus d'analyse médicaux et biologiques sont décrits, on peut envisager la description des résultats de ces processus. C'est l'objectif qui a été assigné à l'IEO. Il s'agit dans ce cadre de caractériser l'information obtenue à propos d'un objet, à l'intérieur d'un processus expérimental.

Dans sa proposition, A. Ruttenberg relie de la façon suivante les composants de l'information :



Cette schématisation est la plus proche de notre caractérisation de la structure d'information. Elle n'en reste pas moins qu'elle s'inscrit dans une réflexion à propos de la symbolisation, puisqu'effectivement on est dans une perspective réaliste.

L'intégration des questions de symbolisation amène à considérer la totalité de la trajectoire depuis l'objet du monde vers le document :



En définitive, l'IEO constitue une ontologie des opérations et des résultats d'opération dont la première difficulté est le fait qu'elle ne structure pas les constituants de l'information à l'intérieur d'une expression.

En effet, le principe fondateur de l'IEO étant de réalisme de la BFO, les processus de symbolisation sont relatifs aux objets et ne présupposent pas de caractérisation du domaine d'appartenance de ces objets. Ou tout du moins, elles n'ont pas à être représentées dans le cadre de la modélisation des structures d'information. Une telle position ne peut être tenue à partir du moment où l'on distingue des univers différents.

Notre proposition relative aux mondes permet de poser différemment le problème et d'externaliser une partie des relations identifiées dans le cadre du rapport de résultat d'analyse.

Ontologies et sémantique des ressources linguistiques.

Nous nous intéressons maintenant à la perspective développée par A. Gangemi[287]. Au départ, le projet a consisté à représenter et rendre interopérables les dimensions sémantiques associées aux descriptions lexicales. La sémantique dont il est question excède la sémantique lexicale généralement associée aux dictionnaires électroniques pour concerner les sémantiques « formelles » : « The amount of lexical resources that are developed either as long-term repositories, or as short-term products of NLP techniques, is growing significantly, posing the problem of understanding their commonalities and their potential for reusability and interoperability. A major concern for reusability and interoperability is the ability to control, both intellectually and computationally, the semantics of the data that are contained in a lexical resource. Lexical data semantics is usually left implicit, either because there is not a shared agreement on how to represent that semantics, or because developers of lexical resources do not customarily employ formal methods" (p.1).

A. Gangemi propose un méta-modèle qui permettrait d'articuler les différentes caractérisations des entités (comme concept, entité informationnelle, interprétée par un agent et insérée dans une certaine situation). Dès lors, il construit une ontologie qui représente ces différentes dimensions :

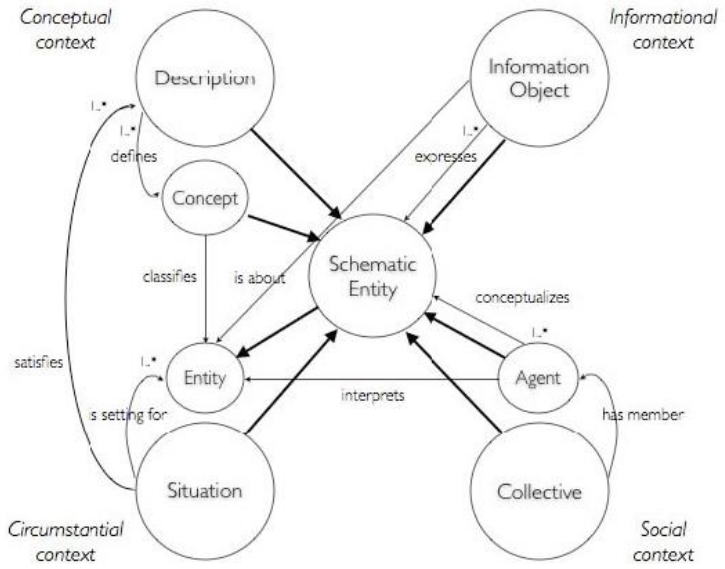


Figure 1.1 **The contextual bindings for the representation of a conceptualization in c.DnS (following the OWL version of the ontology). Ovals denote classes, bold arrows denote subclass relations, regular arrows denote relations holding between members of the linked classes. The cardinality of a relation and its inverse is by default 0...*, except when indicated explictly.**

On ne peut manquer de mettre en relation cette structuration avec FRBR. Néanmoins, le lien entre ces différents modèles n'a pas encore été exploité. Le modèle FABIO[288], qui constitue

un outil de représentation bibliographique fondé sur une ontologie et utilisant FRBR, ne réfère pas à une sémiotique.

On pourra remarquer que ces deux propositions ne prennent en compte l'interprétation que comme un composant du modèle. En d'autres termes, la dualité que l'on a précédemment explorée, entre l'observé et l'interprété, notamment à l'aide des paires de D. Chalmers, peut être exploitée de façon à produire une modélisation plaçant les différentes situations au centre du dispositif.

Une autre stratégie est suivie par A. Gangemi[289] : la solution au problème consisterait à construire un modèle cognitif, basé sur les affordances de J. Gibson. Un lien apparait entre ces propositions et celles de la théorie des situations. L'objectif d'A. Gangemi consiste à modéliser ces relations entre différents niveaux par un principe d'assignation sémantique entre différentes formes de réalisation. Ce niveau de pertinence cognitive permet de construire un modèle appelé FRASL qui relie les différents niveaux de réalisation d'une entité symbolique et ainsi permet de connecter les réalisations lexicales et conceptuelles. L'apport supplémentaire par rapport aux modèles comme LEMON réside dans la construction de formules permettant de caractériser un contenu propositionnel. Ce contenu propositionnel est représenté aux différents niveaux de la modélisation, sous forme soit de représentations d'événements, d'expression (utilisant les grammaires de cas de Fillmore) ou de DRT. Par ailleurs, la ressource théorique de cette proposition est la théorie des événements de Davidson[290]. Cette représentation de l'événement sert d'appui aux relations établies entre les différentes classes.

Un concept central dans ce cadre est celui de « Knowledge Pattern », qui caractérise un ensemble de caractéristiques récurrentes, ou typiques, qui vont ensuite être instanciés dans des réalisations différentes. On peut également citer dans ce cadre les travaux de Laure Vieu[291], qui vient à structurer à l'aide des concepts fondamentaux de DOLCE, des structures descriptives de parties de discours. Même si ce travail est beaucoup plus centré que le nôtre sur le traitement automatique des langues, il propose une caractérisation des frames (fondée en partie sur la théorie des situations) qui permet d'articuler des structures conceptuelles et des modèles de raisonnement insérés dans les discours.

Notre proposition associant flux, structure d'information et situation permet de disposer d'un autre avantage : il permettrait de lier des structures hétérogènes.

6.4.4. Description de documents et extraction d'information. Caractérisation des relations entre l'extraction et la représentation des connaissances.

Nous aimerions maintenant expliquer un aspect complexe du modèle, qui consiste à utiliser une représentation de connaissances (une ontologie plus précisément) pour guider une extraction d'information. Plus précisément, on utilise la structure lexicale, de bas niveau d'abstraction, de l'ontologie pour fournir un vocabulaire au modèle d'extraction d'information, fondé sur les primitives que l'on a précédemment dégagées.
Nous voulons ici seulement spécifier comment notre projet (globalement, mais surtout relativement aux questions d'extraction de structures d'information) s'articule aux

problématiques propres de l'extraction et du text mining.
La question de l'extraction d'information est relativement hétérogène par rapport à l'ensemble de notre problématique parce qu'elle implique d'entrer dans les documents. La perspective est celle de l'intertextualité puisqu'il est question de citations de documents dans d'autres documents.

L'extraction d'information peut sembler une problématique relativement annexe par rapport à notre projet. Or, considérant les structures d'information et leurs dimensions linguistiques, la perspective apparait beaucoup plus proche.
Nous voudrions évoquer maintenant certaines possibilités d'usage des flux en considérant d'autres traitements des documents que les classifications. Par exemple, on pourra considérer l'annotation de documents comme une problématique qui pourrait servir à spécifier des mises en relation de documents en considérant d'autres outils de représentation. En effet, l'annotation, parce qu'elle se place à l'intérieur du document, permet de considérer d'une autre façon le niveau des tokens. En effet, cette problématique a cela de particulier qu'elle permet de relier des parties de document à des marqueurs et ainsi structurer le document avant même qu'il ne soit classé. On trouvera dans la TEI les mêmes propriétés. Si ces perspectives peuvent sembler relativement éloignées des nôtres, elles permettent de montrer que le bas niveau d'abstraction ne peut être considéré comme seulement une masse signifiante imperméable.

6.4.3.1. Spécificité du problème de l'extraction d'information et de l'extraction de connaissances.

L'approche par la seule mise en relation des publications et des descriptions de leurs ressources n'est pas suffisante parce qu'elle ne prend pas en compte les usages de ces ressources, et plus particulièrement les processus dans lesquels ils s'inscrivent. Nous parlons ici des méthodes de l'extraction d'information, qui constituent des outils ayant des contraintes particulières, liées à l'automatisation de la tâche. En effet, comme il n'est plus seulement question de publications, mais de l'ensemble des objets intervenant dans un processus de production de connaissances, on doit présenter l'ensemble des moyens pour construire la description des documents par leur usage. Cette caractérisation peut être externe au document (les processus auxquels le document participe), il peut également être interne. Dans ce dernier cas, on identifie les processus qui ont permis la production de cet objet et dont on trouve les traces dans la structuration interne du document.
Inversement par rapport aux problématiques de la linguistique formelle pour qui la justesse de la modélisation du phénomène linguistique prime, les méthodes d'extraction de connaissances constituent des outils finalisés, reposant sur des représentations formelles inscrites dans une fonctionnalité précise. Les opérations sur les unités linguistiques visent à proposer des représentations d'autres phénomènes que ceux du langage (notamment des connaissances, mais également des notions et des savoirs communautaires). On peut relier ces méthodes à celles historiquement bien plus anciennes de la terminologie, de lexicologie et lexicographie. Elles visent à structurer les entités lexicales en utilisant des règles (structurelles ou pas) de modélisation prédéfinies et orienter vers des usages sociaux définis (comme les terminologies et les lexiques) ou technologiques (comme les lexiques sémantiques). Ces modélisations lexicales, qui peuvent évoluer vers des structurations de structures sémantiques plus larges comme les événements (et plus largement encore la prédication), sont utilisées dans le cadre d'une extraction gouvernée par une ontologie particulière (et peut contribuer à son enrichissement).

Dans le cadre de notre projet, le problème pourrait être relativement simple, puisqu'il s'agirait d'extraire des structures d'information dans lesquelles serait présent le nom du corpus utilisé et le cadre dans lequel ce nom apparait, et qui permettrait de spécifier le contexte dans lequel ce nom est utilisé. De cette façon, on pourrait spécifier si ce corpus est bien utilisé comme corpus de référence ou s'il est simplement mentionné à titre d'exemple, de référence ou de contre-référence. Ainsi, l'extraction permettrait de spécifier l'usage du corpus dans la publication. Il suffirait alors de spécifier les structures d'informations dans lesquelles ces noms de corpus apparaissent et de les faire apprendre à la machine.
Le simple problème de l'identification d'un nom de corpus devient rapidement celui d'une structure d'information prédicative d'un usage de ce nom, avant d'être celui de la description de cet usage lui-même.

Les questions liées à la structure d'extraction ne sont pas exactement superposées à l'ontologie. Nous commencerons par aborder la difficulté spécifique associée à la représentation des connaissances et aux ontologies. En effet, les connaissances mobilisées pour caractériser l'usage d'un corpus ne sont pas directement celles manifestées dans le cadre de l'ontologie, puisqu'il ne s'agit pas du même contexte (description de la production d'un objet *VS* usage de cet objet dans le cadre d'une recherche distincte).
Le problème posé par la dimension localisée de l'attestation du nom de corpus est encore accentué lorsqu'il s'agit de caractériser la dimension des connaissances, à savoir l'intégration de la description de l'usage du corpus dans le cadre plus global de la menée d'une recherche. Dans ce cadre, les structures d'événements apparaissent comme les modèles candidats les plus adaptés.

L'extraction d'information est fondée sur des principes initiaux de partialité, à savoir que l'on extrait à partir de patterns identifiés. On définit donc préalablement ce que l'on veut extraire avant de caractériser comment cet objectif se traduit dans le cadre des textes, et enfin comment on peut élaborer un outil qui permettrait de rendre cette extraction conforme au modèle choisi.

Nous distinguons entre extraction de connaissances et extraction d'information. L'extraction de connaissances consiste à partir d'une structure de connaissances, à déterminer les connaissances contenues dans un texte quelconque. L'extraction d'information consiste à identifier dans des textes des structures linguistiques répondant à certaines caractéristiques prédéfinies. Ces structures contiennent certains faits nouveaux qui sont caractérisés à partir de structures ou patterns prédéfinis.

L'extraction d'information sert entre autre à alimenter et enrichir des ontologies : « These two tasks are combined in a cyclic process: ontologies are used for interpreting the text at the right level for IE to be efficient and IE extracts new knowledge from the text, to be integrated in the ontology. »[292] . En effet, on utilise des ontologies pour structurer des patterns qui à leur tour permettent d'enrichir les ontologies.

Cette limite nous oblige, sur le projet, à recourir à des méthodes d'extraction d'information dans les documents à partir de certaines indications fournies par les descriptions des ressources primaires. Cette articulation de l'extraction des structures d'information à l'ontologie (comme décrivant les ressources primaires) fait correspondre ce que le texte peut nous apprendre des données qu'il utilise (dans quel contexte il les utilise, quel rôle elles jouent dans l'argumentation du propos scientifique, quelles dimensions des données sont les plus

utilisées, etc.), et comment ces données identifiées et contextualisées renseignent sur les données primaires auxquelles elles réfèrent. En effet, le principe consiste à établir une structure à partir du nom de corpus et de certaines indications descriptives de ce nom à l'intérieur de l'ontologie, essentiellement au travers des métadonnées. Ces indications sont supposées attester de l'usage de ce corpus, mais ne constituent pas une structure phrastique entière. Une part du projet consiste à identifier des entités lexicales et grammaticales qui permettront de relier ces unités lexicales communes à la description du corpus (la ressource primaire) et à la publication. (Cette question est importante parce que ce n'est pas parce qu'un corpus est cité dans une publication qu'il constitue la donnée primaire utilisée pour mener l'étude en question).

On peut illustrer cette question en prenant quelques exemples :
« Switchboard (Godfrey et al., 1992), a large corpus of telephone conversations involving about 500 speakers who were given a pre-determined topic to talk about. »
Le nom de la ressource est suivi d'un concept de haut niveau d'abstraction de l'ontologie (`corpus, tool`).
Dès lors que ce concept est acquis, il est suivi d'un descripteur de plus bas niveau d'abstraction (`telephone conversations`).
Une troisième étape permettrait d'obtenir une spécification encore plus précise (`500 speakers`). Cette dernière spécification est insérée dans les valeurs décrites par les métadonnées de ce corpus de même que « pre-determined topic ».

The BioCreative corpus contains only one entity subsuming genes and gene products (proteins, RNA, etc.) labeled for 7500 training sentences and 2500 evaluation sentences.
« BioCreative » : nom de la ressource
« Corpus » : générique (haut niveau de l'ontologie).
« Training sentences », « evaluation sentence » : spécifiques de l'ontologie.

ABNER (A Biomedical Named Entity Recognizer) is an open source software tool for molecular biology text mining.
« ABNER (A Biomedical Named Entity Recognizer) » : nom de la ressource
« tool » : générique (haut niveau de l'ontologie).
«text mining» : spécifiques de l'ontologie.

Nous reviendrons plus loin sur la façon dont on peut spécifier cette extraction, et la compléter. Il s'agissait seulement ici d'illustrer notre hypothèse de départ.

Le problème qui apparaît immédiatement est celui de l'annotation des publications. En effet, pour pouvoir extraire les structures d'information qui contiendraient toutes les informations relatives à l'usage des données primaires, il faudrait que les documents soient annotés et que l'on ait établi quelque équivalence entre les marqueurs d'annotation utilisés dans un corpus ou un autre (par exemple WORDNET ou FRAMENET) et notre propre modélisation de la structure d'information.
Un travail sur corpus annoté (ou corpus d'apprentissage) précède donc l'utilisation de notre modèle de la structure d'information.

6.4.3.2. Extraction d'information et text mining.

L'extraction d'information est une part du text mining. Parce qu'il s'agit de connaître quelles sont les données qui seront proposées à l'analyse, l'extraction d'information constitue un ensemble de procédures visant à représenter certains contenus. « [Information extraction] requires deeper analysis than key word searches, but its aims fall short of the very hard and long-term problem of text understanding, where we seek to capture all the information in a text, along with the speaker's or writer's intention"[293].

L'extraction d'information à partir de documents numériques a connu une expansion récente, liée notamment au fait que le format le plus utilisé pour publier des documents en ligne (le PDF d'Adobe) peut dorénavant être exploré par un outil d'extraction d'information sans avoir à être préalablement converti en format texte[294].

Le problème de l'ensemble des approches de type « text mining » est qu'elles n'assurent à aucun moment de la pertinence de l'extraction de la connaissance contenue dans le texte. En effet, ces outils sont paramétrés en fonction d'un objectif prédéfini. Les outils statistiques sont définis par différentes possibilités de calcul, notamment de similarité. Il s'ensuit que les capacités de traitement des outils statistiques dépendent des paramètres qu'on leur fournit et des indications sur ce qu'ils doivent extraire.

Rappelons que trois modèles existent pour l'extraction d'entités linguistiques depuis des textes :

- des modèles linguistiques, qui peuvent être syntaxiques ou sémantiques, et qui requièrent l'annotation linguistique des textes[295].
- des modèles à fondement statistique et probabiliste. Ils peuvent être supervisés par des modèles linguistiques.
- des modèles conceptuels requérant des outils statistiques[296].

Les outils d'extraction peuvent impliquer des annotations des textes, à savoir une préparation du texte antérieure à l'extraction. Dans le cadre du web de données, un certain nombre de modèles d'annotation sémantique ont été élaborés. EARMARK[297] constitue un exemple particulièrement développé et inscrit dans la perspective de la mise en relation des documents. Cette préparation permet une extraction linguistique (fondée sur des critères syntaxiques ou sémantiques),

Nous avons présenté notre modèle de structure d'information en le positionnant par rapport aux travaux linguistiques. Dans le cadre d'une utilisation web, il convient de le mettre en relation à d'autres structures équivalentes, ce qui nous permettra aussi de la distinguer clairement des structures d'événement de type de celles proposées par SEM[267].

[267] (Pour rappel : http://semanticweb.cs.vu.nl/2009/04/eventExtended/http://semanticweb.cs.vu.nl/2009/04/eventExtended/).

Par ailleurs, quelles que soient leurs qualités, les méthodes d'extraction d'information fondées sur des analyses linguistiques présupposent que les textes soient préalablement annotés. Or, cette phase préparatoire constitue un frein à leur utilisation puisqu'il faut que le corpus d'apprentissage soit de composition similaire à celui sur lequel on souhaite une application. Des méthodes plus légères, empruntant notamment aux probabilités, permettent d'amoindrir les hétérogénéités entre le corpus d'apprentissage et celui sur lequel se réalise la tâche d'extraction. Notre caractérisation de la structure d'information repose certes sur des critères sémantiques, mais pas sur des principes grammaticaux. Ainsi, pour son application, elle ne requiert pas une analyse grammaticale.

Ces différents modèles fonctionnent grâce à des méthodes différentes : on distingue alors les modèles supervisés, semi-supervisés ou non supervisés. La distinction entre ces différentes méthodes réside dans la plus ou moins forte contrainte exercée par les modèles a priori. Plus un modèle (syntaxique, de connaissances, de structure sémantique) sera contraignant, plus le modèle sera supervisé.

L'extraction d'information s'est particulièrement développée autour de deux aspects : l'extraction d'entités nommées et les relations entre entités et événements. (Pour un tour d'horizon : Serrano & alii.[298] p.5). Les deux tâches sont bien évidemment liées. Les entités nommées « correspondent de façon générale aux noms de personne, organisation, lieu, mais aussi aux dates, unités monétaires, pourcentages, unités de mesure, etc. »

Nous verrons que les structures d'information ne montrent pas une régularité totale, comme toute réalisation du langage naturel. Les structures fondamentalement conceptuelles ou sémantiques, voire syntaxiques, sont donc relativement inadaptées à extraire seules de l'information. On doit donc les articuler à des outils probabilistes de façon à obtenir des extractions satisfaisantes ; on ne peut donc guère utiliser de modèles linguistiques ou conceptuels à grande échelle. Si effectivement, dans le domaine biomédical, il existe des outils d'extraction d'événement purement sémantiques (Miwa & alii, op.cit., p. 3), ceux-ci reposent sur un cadre lexical et conceptuel très limité et une expérimentation sur un corpus annoté de façon très fouillée (GENIA[299]). Nous proposerons donc l'utilisation d'approches mixant les structures conceptuelles et les modèles probabilistes.

6.4.3.3. Présentation des outils du text-mining.

Le text mining constitue un ensemble de techniques permettant d'extraire des structures textuelles préétablies, appelées « patterns » à l'intérieur de documents non structurés ou semi-structurés.

En ce sens, le text mining requiert à la fois des connaissances et des méthodes issues de domaines différents : le traitement automatique des langues, la linguistique de corpus, la recherche d'information et évidemment l'informatique.

Une autre idée fondamentale du text mining consiste à traiter des collections de documents. L'ensemble des documents comme les finalités du traitement peuvent être très différents.

Le text mining peut appréhender les unités symboliques à différents niveaux : la suite de caractère, le mot, le terme, le concept. Il s'agit donc d'une méthodologie qui peut être associée à des projets relativement différents, qu'il s'agisse de recherche d'information, de veille, d'analyse de discours, de catégorisation, ou encore d'enrichissement de thésaurus, d'ontologie…
Toutes ces problématiques impliquent l'analyse du contenu des documents, ce qui n'est pas le cas du data mining. Ce dernier explore les bases de données et les descriptions des documents, alors que le text mining extrait les informations pertinentes à partir du document lui-même.

Méthodologiquement, le text mining requiert quatre étapes fondamentales :

- Les tâches de prétraitement du document de façon à obtenir un format canonique
- Les opérations fondamentales
- La visualisation et plus globalement la représentation des résultats
- Le post-traitement.

Le text mining consiste ainsi à répondre à des questions qui sont extérieures à une problématique linguistique, mais relèvent de besoins liés à des systèmes d'information. On peut définir le text mining au travers de sa méthodologie. On peut également le définir de façon un peu plus restrictive à l'exemple de V. Gupta & G. Lehal[300] : « Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. »

Le text mining est particulièrement adapté à l'extraction de structures d'information à partir d'un modèle préétabli, mais avec l'étape de l'apprentissage de lexiques. Cette phase correspond au résultat de l'application des modèles sur des unités lexicales : grâce à la sémantique du modèle et son application sur des catégories lexicales, il devient possible de caractériser des clusters ou ensemble d'unités lexicales possibles dans le contexte de la structure d'information.
Cela dit, le text mining possède trop de liens avec d'autres méthodes et techniques pour être envisagé uniquement sous l'angle de l'extraction de structures d'informations. (La mise en relation thématique de documents, le résumé, la catégorisation, la clusterisation, la représentation visuelle de l'information et l'association de documents constituent des problématiques proches, impliquant certaines méthodes et techniques identiques).

Les difficultés sont relativement différentes en fonction de certains choix liés à par exemple la nature des structures que l'on souhaite extraire. Par exemple, des questions comme la distribution demandent à être résolues en utilisant des patterns syntaxiques et des rôles sémantiques[301]. Ce choix implique l'annotation des documents. Mais surtout, il permet de lier les extractions à l'obtention d'informations à propos de relations représentées à l'intérieur d'ontologies.

Ainsi, le text mining est orienté et spécifié par les domaines dans lesquels s'inscrivent les documents, tout simplement parce que ce qu'il permet d'obtenir, ce sont des connaissances nouvelles pertinentes dans un domaine de connaissances déjà construit. La finalité du text mining peut varier en fonction justement de la spécificité des relations entre connaissances

que l'on peut identifier à l'intérieur d'un domaine de connaissances. Ainsi, le lien aux ontologies constitue un phénomène marquant de la recherche sur le text mining.

L'idée centrale est l'extraction de patterns. Ils constituent la tâche centrale du text mining. Cette extraction prend généralement la forme de questions qui sont structurées par des algorithmes. Qu'il s'agisse de distribution, d'annotation ou encore d'ensembles récurrents de termes, ce sont des outils de calcul de proximité et de dépendance qui sont caractérisés.

Les patterns sont des structures qui posent les problèmes usuels liés au traitement des unités linguistiques : comment peut-on être sûr d'avoir extrait les unités pertinentes, surtout toutes les unités (sans avoir généré le moindre bruit) ?
Les méthodes se distinguent dans la prise en compte des aspects strictement linguistiques des textes et dans les dimensions statistiques lexicales.
Un exemple particulièrement pertinent de cette approche est celui présenté par M. Miwa & alii[302]. Il est certes fondé sur un domaine biomédical où la structuration des publications est beaucoup plus normée que dans le cadre des Sciences Humaines et Sociales. C'est donc dans un tel cadre que l'on a les avancées les plus marquantes.

6.4.3.4. Les modèles d'extraction d'information : évolution actuelle.

La structure d'information que nous avons précédemment définie constitue un outil qu'il convient de situer par rapport à d'autres, qui pourraient avoir un rôle similaire. Un premier modèle d'extraction de structure d'information fondé sur les événements est FRAMENET : il s'agit d'une structure sémantique fondée sur les principes de C. Fillmore[268]. Il s'agit d'une structure fondée sur un modèle de l'événement associé à des cas sémantiques.

Nous porterons une attention particulière au modèle d'extraction d'événement SEM[303]. Celui-ci a pour vocation de distinguer explicitement les instances de bas niveau d'abstraction des constituants de haut niveau d'abstraction, caractérisés au travers de types.
Dans ce cadre, le haut niveau d'abstraction se caractérise par un type de relations appelé événement, et qui comprend une classification. Dans le domaine biomédical, une approche similaire est proposée (M. Miwa & alii., pp.1-2 : "Events constitute structured representations of biomedical knowledge. They are usually organised around verbs (e.g., *activate, inhibit*) or nominalised verbs (e.g., *expression*), which we call *trigger expressions*. Events have arguments, which contribute towards the description of the event. These arguments, which can either be entities (e.g., *p53*) or other events, are often assigned semantic roles,
which characterise the contribution of the argument to the description of the event."

Cette structure d'événement s'applique à la fois à un cadre d'extraction d'information à l'intérieur de textes, à la représentation de la trajectoire de navires et évidemment la mise en relation de documents relatifs à des événements historiques précis.

Néanmoins, l'aspect le plus intéressant de SEM est l'héritage de la théorie des situations, à savoir que SEM propose une structure type d'événements, incluant pour chaque type un ensemble de rôles et des contraintes liées.

[268] https://framenet.icsi.berkeley.edu/fndrupal/about.

Les travaux de M. Miwa & alii., partent au contraire de structures d'événements considérées des structures prédicatives, fondées sur des concepts caractérisant des événements biomédicaux (« Gene expression , Transcription, Protein, Catabolism, Phosphorylation, Localization, Binding, Regulation, Positive regulation, Negative regulation. »). Ces événements sont associés à des thèmes ou des causes qui caractérisent des arguments primaires.

Ces événements requièrent une approche par apprentissage-machine, utilisant alors des outils qui permettent de réduire les incertitudes.

Si la définition linguistique de la structure d'information n'a guère évolué, par contre, la numérisation des documents a facilité l'extraction de structures de plus en plus complexes. En même temps, elle a permis de transformer la nature et la forme des données traitées. En effet, si l'on regarde du point de vue de la linguistique, on peut identifier une première grande période d'élaboration de corpus électroniques, et qui se concrétise à la fois par les méthodologies de la linguistique de corpus (voir notamment les travaux de D. Biber[304]), destinées en grande partie à fournir des données pour l'apprentissage des langues et par les grands corpus annotés du LDC[269] et de ELDA[270].
Une seconde période s'ouvre aujourd'hui avec la capacité à travailler sur des corpus encore plus volumineux avec le traitement du flux continu des courriers électroniques, twits, etc. Cette transformation des données s'accompagne d'enjeux d'utilisation importants : ce sont les entreprises, les cabinets d'études et autres sondeurs de l'opinion qui vont s'intéresser à cette masse discursive numérique. Comme nous le verrons à propos du DECOODA[271], les outils linguistiques tout autant que les hypothèses et méthodes mises en œuvre diffèrent totalement des travaux de la période précédente. Parmi celles-ci, l'une des plus avancées est le calcul de similarité entre textes, phrases, groupes de mots, ce qui requiert l'analyse parallèle de plusieurs textes. [305]

La méthode la plus classique consiste à travailler à partir de textes annotés. Cette phase implique l'annotation des corpus, ce qui empêche une étude à grande échelle. La méthode est la suivante : on dispose d'un corpus annoté à partir duquel la machine apprend un certain nombre de règles. Ces règles apprises peuvent alors être appliquée sur d'autres corpus, qui eux ne sont pas annotés préalablement. Le choix du corpus d'apprentissage se révèle crucial pour la mise en œuvre de ces méthodes puisqu'en fonction du type de corpus de départ le domaine d'application du modèle va pouvoir être défini.
Nous voudrions maintenant identifier les méthodes fondées sur des apprentissages de relations entre structures lexicales prédéfinies. Ces méthodes n'impliquent pas l'annotation. Elles requièrent par contre des outils probabilistes. L'extraction d'information à partir du web s'inscrit dans ce cadre. En effet, il n'est guère possible, dans le cadre du web, d'annoter les documents. Elles permettent d'utiliser des catégories lexicales qui ne seraient pas totalement liées à des places et des catégories syntaxiques.[306]
Nous ne traitons pas ici l'apprentissage de règles et de structures linguistiques à partir d'outils statistiques. L'apprentissage de la structure peut donner lieu à des règles d'apprentissage.
.

[269] https://www.ldc.upenn.edu/
[270] http://www.elda.org/
[271] http://decooda.com/

Ces méthodes, dont le DECOODA constitue une application, n'impliquent pas strictement des principes théoriques linguistiques. Il peut être opportun d'utiliser des outils d'analyse issus d'autres disciplines traitant du discours, comme notamment la psychologie cognitive.
La difficulté de ce type de modèle est qu'il faut trouver des méthodes et des outils qui se substituent aux catégories et modèles de structures syntaxiques.

Comme nous l'avons vu, certains modèles d'extraction d'information visent à représenter des événements. Si l'on suit Laurie Serrano & alii, (op. cit.), l'extraction d'événements est fondée sur une dimension temporelle associée à un fait apparaissant alors comme pertinent dans un cadre précis. Ainsi, l'ontologie permet de caractériser un horizon d'attente qu'une information vient combler. Les auteurs ajoutent à la dimension temporelle celle des acteurs et des lieux.

Les flux, et par conséquent notre modèle de la structure, structurent de façon plus précise la représentation de l'événement en explicitant les propriétés conditionnelles à l'événement, en instaurant une hétérogénéité entre le cadre de l'événement et ce qu'il advient dans le temps. En ce sens, la conversion d'un modèle de représentation de l'information dans un modèle d'extraction apparaît comme un objectif possible pour notre travail.

Dans la dernière partie de notre travail, nous verrons en quoi cette hétérogénéité est essentielle pour ne pas confondre une information et une connaissance.

On considère que chaque constituant de la structure d'information peut être considéré comme une entité nommée. On peut alors considérer la structure d'information comme une structure de relations qui constituerait un pattern. Rappelons que ce pattern, dans le cadre du projet, identifierait la situation d'usage d'une ressource (ou donnée primaire), telle que décrite dans une publication.

Conclusion.
Dans cette partie, nous n'avons guère approfondi les questions de l'extraction, qui dépassent le cadre de notre propos (mais s'intègrent pleinement à notre projet). La question de la structure d'information n'est pas seulement une question de linguistique mais de connaissances. Nous avons donc caractérisé cette structure comme un assemblage de primitives assurant le transfert d'information.

L'utilisation d'une telle structure en text mining répond exactement à la définition proposée par V. Gupta. Notre structure d'information constitue la forme minimale de la structure de connaissances.
En définitive, notre modèle de structure d'information se distingue de l'ensemble des autres en ce qu'il est fondé sur l'hétérogénéité. Il repose sur l'idée relativement simple selon laquelle on apporte une information nouvelle dans un contexte connu par l'adjonction à ce contexte d'une entité hétérogène, jusque-là extérieure.

6.5. Retour sur les données : portée des exemples, données primaires et utilisation du modèle de l'activité.

Toute donnée scientifique primaire dépend fondamentalement de la façon dont on la construit et la définit, du degré d'abstraction de cette définition et donc des propriétés que l'on retient pour la caractériser. La différence est que le web est un objet que l'on construit, alors que la

pharmacie hospitalière est un objet construit que l'on observe.
Les flux et les structures d'informations sont une modélisation descriptive pour expliciter des processus que l'on observe dans le monde. C'est la même méthode que la linguistique formelle. Par contre, de tels processus ne sont pas (encore) observables sur le web. C'est là où l'on vise à les faire exister.

Les problèmes sont relativement similaires tant que l'on s'intéresse à la façon dont l'information circule. C'est la raison pour laquelle notre modélisation est adaptée aux questions posées par les métadonnées et plus généralement la description documentaire.
Le présupposé de cette approche par microcosme est une définition de l'information comme fait culturel majeur.

Les flux sont dans notre approche caractérisés à la fois dans le cadre d'un modèle et dans celui d'observables empiriques et matériels (les structures d'information et les situations de la pratique hospitalière). Ils ne sont pas exclusifs d'une seule approche. En effet, nous avons veillé à explorer la question des flux en considérant qu'il s'agit d'abord d'un phénomène culturel, considérant la formalisation essentiellement comme un outil d'expression adéquat.

Nous avons cherché à montrer que les flux pouvaient être considérés autrement que comme un seul modèle formel, utilisable dans le cadre du web sémantique. Ses fondements sont un modèle de la signification. Par conséquent, il s'agit d'un outil caractérisant des phénomènes humains. Ils s'intègrent donc facilement dans l'esprit du web, par lequel les outils et les méthodes sont d'abord adaptés aux usagers avant d'être des outils informatiques.

Enfin, il s'agit d'un outil de mise en relation de données structurées, telles que celles qui sont utilisées en Sciences de l'Information et de la Communication : les flux permettent de donner un sens à des raisonnements dans le cadre des mises en relation (ou mapping) de données structurées. Il s'agit de raisonner en montrant qu'une connaissance entraîne dans un autre cadre une connaissance hétérogène. Nous avons ainsi montré que c'est justement de cette façon qu'il est possible de caractériser une dynamique au sein des structures conceptuelles, notamment de haut niveau d'abstraction.

Nous aimerions maintenant élargir le propos à notre démarche générale.

Notre travail articule deux dimensions : une description puis une modélisation de « pratiques culturelles » au travers de l'exemple de la pharmacie, et un effort de conception au travers du projet. Les mises en relation de concepts entre des perspectives d'analyse et de conception ne vont pas de soi. Les théories étant développées relativement à des finalités spécifiques, elles n'ont pas vocation à fournir le socle pour une mise en relation entre analyse et conception.
En d'autres termes, ce ne sont pas sur les concepts et les objets abstraits qu'une mise en relation est possible, mais bien sur ce qui est décrit du monde social d'une part, et les fonctionnalités d'une autre. Si l'on s'aperçoit qu'il y a un lien entre d'une part les circulations d'information dans certains protocoles d'analyse et de contrôle pharmaceutique, et d'autre part, la façon dont il est possible d'articuler des données hétérogènes de façon à former une information pertinente dans le cadre d'une ontologie, alors on pourra utiliser l'expérience acquise pour spécifier et organiser des dispositifs d'une autre ampleur (et dont nous parlerons en conclusion). En effet, en pharmacie, il avait été possible d'établir une inférence entre des données hétérogènes et distribuées, de façon à ce qu'elles forment une structure informationnelle susceptible d'être interprétée. Cette observation nous a permis d'établir un

lien entre flux et structure d'information, et surtout de montrer que le modèle de flux constitue une abstraction de phénomènes empiriques, observables. Notre hypothèse a été qu'il n'existe pas de rupture entre une activité scientifique d'observation et de description d'une part, et un travail de modélisation et de formalisation d'autre part. Nous avons essayé de rendre compte de cela dans la proposition de niveaux d'abstraction.

On pourra toujours objecter que les modèles (en l'occurrence ici les flux) que l'on utilise ne constituent pas une déduction de l'usage, et qu'ils préexistent historiquement à notre description. Or, d'une part, les flux constituent aussi une problématique (hétérogène par rapport à une théorie ou un modèle), à savoir que l'on peut observer des régularités de transmission des unités signifiantes depuis une localisation vers une autre sans augmentation d'ambiguïté ou de perte de précision. D'autre part, c'est avec un autre statut et des propriétés augmentées par l'analyse (voire explicitées) que le modèle peut maintenant être caractérisé. Nous pouvons, grâce à l'observation, préciser quelles sont les classes, ce que l'on peut y trouver comme entités et de quelle façon les flux contribuent à la formation des structures d'information. Par conséquent, on propose un enrichissement par l'interaction entre les différents niveaux d'abstraction.

Caractérisation du modèle de l'activité.
Le jeu que l'on propose n'est pas associé à une discipline quelconque et renvoie en grande partie aux concepts définis dans les ontologies d'événements. L'ontologie des processus informatiques proposée par G. Kassel, P. Lando et alii.[307] ne prend en compte que la dimension conceptuelle.
La notion d'activité que l'on développe se distingue des ontologies proposées pour représenter ce niveau par le fait que l'on ne se situe pas à un haut niveau d'abstraction. En règle générale, le cadre dans lequel se situent les ontologies de l'activité est celui du haut niveau d'abstraction. Or, justement, nous avons montré que l'activité se structure en premier lieu autour d'objets matériels et d'objets de référence. Ainsi, le cadre des ontologies de haut niveau n'apparait pas nécessairement pertinent pour rendre compte d'un phénomène comme l'activité.
On se dirige ainsi vers l'élaboration de représentations formelles de l'activité qui ne seraient pas systématiquement fondées sur des ontologies de haut niveau. Cela est également lié au fait que notre ontologie représenterait l'interprétation dans l'activité, et non, par exemple, la production d'un objet dans le cadre d'une activité.

Nous devons également lever une autre ambigüité : comment est-il possible d'articuler une sémantique interprétative et une représentation de l'activité, sachant que les modèles de l'activité caractérisent un processus ?
Les paires de D. Chalmers montrent qu'il est possible de conceptualiser cette dualité. Tout notre travail ultérieur consistera à traduire cela dans une représentation.

Conclusion pour la partie 6.

Nous avons dans cette partie donné des éléments permettant d'envisager comment les modèles que l'on avait présentés antérieurement pouvaient être intégrés dans un projet relevant du web de données, et donc situé à l'intérieur des problématiques et des champs de ce domaine. Nous avons vu également que les questions qui se posent autour des descriptions bibliographiques pouvaient être élargies dans celles, beaucoup plus générales, des outils et des domaines du web sémantique.
Nous avons également vu que les questions qui se posent aux bibliothèques numériques ne

pouvaient être dissociées de celles qui se posent globalement à ce domaine de travail.

Le projet collectif auquel réfère cette partie constitue une application très partielle du modèle général précédemment exposé. Les perspectives ouvertes sont effectivement beaucoup plus vastes et s'insèrent ans les questions les plus fondamentales de l'évolution du web, et notamment le rapport à la langue.

Enfin, une des spécificités du web sémantique réside dans le fait que la mise en relation de documents ou de données permet d'effectuer des tâches comme par exemple des fonctions pédagogiques (recherche sémantique, génération automatique d'inter-références, tables de contenus, et navigation dirigée par des ontologies). De cette façon, il est possible de connecter des représentations de documents avec des inférences en lien à une activité, ou tout du moins un service. C'est dans cette perspective qui consiste à doter le web de raisonnements susceptibles d'offrir une valeur ajoutée à l'information que se situe notre projet.

7. CONCLUSION.

Nous avons couvert un nombre de domaines important dans le cadre de la problématique de la description documentaire et bibliographique au sein du web de données. Néanmoins, ces outils ne sont explorés que de façon relativement sommaire par rapport aux possibilités qu'ils offrent. Par ailleurs, une telle amplitude de connaissances se traduit néanmoins par l'absence d'un état de l'art complet dans chacun des domaines traversés et chacune des pistes demande un développement spécifique approfondi. Mais en même temps, nous avons insisté sur la complémentarité des différentes perspectives.

Nous aimerions dans cette conclusion insister sur deux aspects :

- La nécessité à construire un domaine de compétences propres et articulé aux problématiques des équipes partenaires,
- La construction de consortiums de dimension adéquate pour la menée à bien de ces projets. En effet, tout notre propos relativement à l'e-science ne peut manquer d'avoir des conséquences sur la façon dont on doit aujourd'hui envisager le développement de notre recherche, lequel comprend évidemment ses financements.

7.1. Elaboration d'un domaine de compétences pour des projets relatifs aux langages documentaires et aux bibliothèques numériques.

A la fin de ce travail peut proposer une méthodologie de travail pluridisciplinaire dans le cadre de l'évolution des pratiques de recherche (notamment les recherches sur projet). En effet, comme on l'a montré, les spécialistes de métadonnées ou des ontologies redécouvrent les dimensions sémantiques et sémiotiques des entités qu'ils manipulent ou encore la caractérisation de la distribution (est-ce seulement une façon d'identifier des structures de connaissances distinctes ou au contraire de caractériser des ensembles de connaissances liées à des pratiques et des activités humaines spécifiques ?). Il y a donc une valeur ajoutée importante à structurer les projets de recherche pluridisciplinaires.

Par ailleurs, cette solidarité des questions linguistiques, cognitives, informatiques, d'Intelligence Artificielle et de documentation entraîne une façon particulière d'envisager les projets de recherche et surtout le lien que l'on peut établir entre les projets du web et ceux, antérieurs ou parallèles, des sciences cognitives et de la linguistique formelle.

Ainsi, la méthodologie que l'on a construite dans ce travail a comme vocation à élaborer des projets pluridisciplinaires conformes à la spécificité du web de données (par rapport à l'informatique traditionnelle) : l'immédiate adaptation à l'usager, la mise en relation de données structurées, la description.

Ainsi, on construit la place des Sciences de l'Information dans le cadre des projets de recherche mobilisant le web documentaire. Au-delà des connaissances liées à la dimension strictement documentaire, les compétences concernent les questions d'évolution du domaine et du cadre institutionnel et organisationnel tout autant que l'articulation entre les aspects sémantiques, logiques et les langages de représentation.

Univers professionnel et technologies.
Un premier problème est la place et le rôle de la recherche en Sciences de l'Information par rapport à celle menée en Informatique. En effet, on ne peut manquer d'être interpelé par le fait que la plupart des chercheurs qui inventent les ontologies bibliographiques soient des informaticiens, et non des chercheurs en Sciences de l'Information.
Pour nous, considérer que les sciences de l'information se préoccupent des applications impliquant la gestion et la mise en valeur de collections ou les conséquences de l'intégration des nouvelles technologies est réducteur de la portée du lien entre un univers professionnel et l'élaboration d'outils.

Le propos a été de ne pas partir de demandes immédiates ou de besoins formulés mais de remonter vers des questions plus générales d'usage de l'information dans un cadre professionnel afin de pouvoir appréhender les questions de circulation d'information et de contenu de l'information. Cette caractérisation de l'information permettra ensuite de proposer des projets pour les bibliothèques.

Notre travail possède une double dimension : l'élaboration d'un outil par la modélisation d'un phénomène culturel et l'adaptation de cet outil à un cadre documentaire particulier. En Sciences de l'Information, les approches du langage (quelles qu'elles soient) constituent le plus souvent des nécessités pour la construction d'outils documentaires. Par exemple, la morphosyntaxe pour l'extraction d'unités lexicales de façon à aider à la tâche d'indexation. Néanmoins, ce n'est pas ce qui nous intéresse pour le moment : nous nous intéressons à des approches descriptives concernant à la fois les phénomènes linguistiques et la façon dont ces analyses supportent l'élaboration de modèles destinés à la construction d'inférences automatisées.

Un second élément fondamental de notre projet est l'articulation entre l'étude d'un univers professionnel et une orientation projet, qui évidemment implique des connaissances dépassant le seul cadre disciplinaire fixé par l'étude de l'univers professionnel. Dans ce travail, nous avons voulu contribuer à éclairer sur ce que sont les objets des sciences de l'Information et les méthodes que l'on pouvait mettre en œuvre pour élaborer des modèles de ces objets scientifiques de façon à élaborer des propositions pour les bibliothèques.
Ces projets ne peuvent être mono-disciplinaires. Ils impliquent entre autre une dimension informatique importante. Par ailleurs, considérant que l'étude de l'évolution des bibliothèques numériques implique une excellente connaissance des dimensions informatiques, on ne pouvait manquer d'opérer un rapprochement important. Cela dit, un positionnement en Sciences de l'Information et de la Communication permet d'intégrer différentes approches, du fait à la fois de l'ancrage dans les Sciences Humaines et Sociales et le lien à un secteur professionnel.

Notre travail contient parmi différentes trajectoires un mouvement qui part d'un modèle existant, logique et mathématique, et qui s'emploie à comprendre la portée de ce modèle dans le monde de l'activité quotidienne. Le premier enjeu, et qui constitue un point essentiel pour les Sciences de l'Information, réside dans le lien qu'il peut y avoir entre les dimensions humaines, et notamment les usages, et la conception des outils. Plutôt que de partir d'usages pour améliorer un outil existant, nous avons préféré prendre un modèle, comprendre quels phénomènes de l'activité humaine il permettait la représentation, et enfin quel projet scientifique on peut formuler à l'aide de la reformulation de ce modèle.
Le second enjeu est celui de la validation des modèles. Pour les modèles mathématiques, on

se sert de preuves. Les preuves, qui constituent des sujets de débat, sont des mécanismes internes au monde de l'expression mathématique, et qui ne caractérisent pas la portée des formalismes construits dans le monde de l'activité. Nous avons au contraire considéré ce modèle comme un outil descriptif et analytique.

Nous avons largement parlé de modélisation, mais pas dans un sens informatique. On considère essentiellement des modèles conceptuels fondés sur des pratiques ordinaires. Cette spécificité explique que nous n'avons pas présenté de système d'information, au sens entendu par les informaticiens. On caractérisera également de la même façon la limite entre une recherche en SHS et en informatique. Le lien entre les phénomènes humains et les représentations modélisées constitue pour nous un enjeu essentiel des Sciences de l'Information. Il permet aussi de commencer à construire un vocabulaire commun entre des approches scientifiques fondamentalement différentes.

Le point commun entre les activités professionnelles que nous avons étudiées (le contrôle thérapeutique essentiellement) et l'univers des bibliothécaires réside dans le fait que l'on considère l'information comme un outil nécessaire à l'action (la recherche dans le cas de l'e-science). Le nouveau rôle des bibliothécaires transforme en grande partie la façon dont peut être appréhendée d'information et lui permet d'évoluer vers les conceptions que nous avons mises en avant. Ainsi, nous fournissons un support théorique au tournant professionnel de l'activité de bibliothécaire comme fournisseur de services à la recherche.

Usages sociaux et élaboration d'outils.

Les approches cognitives que nous avons utilisées sont distinctes de celles qui ont trait à l'analyse des usages en Sciences de l'Information parce qu'elles ne spécifient pas ce que constitue une demande, un besoin ou une adéquation de la réponse par rapport à la demande. En d'autres termes, les travaux sur l'usage ont une approche fonctionnelle des phénomènes sociaux, centrés sur le comportement de l'utilisateur vis-à-vis d'un dispositif. Elles sont également à postériori, à savoir qu'elles interviennent après la diffusion de l'outil. En ce sens, elles constituent essentiellement des analyses du rapport entre un outil existant et des usagers[308].
Ce type de travail ne permet pas de positionner le chercheur dans une position de novateur, mais d'observateur ou d'accompagnateur des mutations. Mais ce rôle et ces outils ne permettent pas de proposer une innovation.

Pour nous, cela n'a pas été le cas : nous avons considéré que l'information constituait un phénomène culturel à part entière, et donc la dichotomie entre usage et outil ne se posait pas en ces termes dans notre travail. Si effectivement le raisonnement correspond à un phénomène culturel relativement universel, la question de l'usage renvoie à des phénomènes qu'il convient de distinguer.
En premier lieu, l'e-science connait une expansion différente en fonction des pays (nous avons vu que la Grande Bretagne est particulièrement novatrice), mais également des disciplines et de leurs ressources propres et enfin des pratiques des communautés plus restreintes. Par ailleurs, comme il s'agit d'un phénomène innovant, il n'est pas impulsé par les chercheurs (des disciplines concernées s'entend), mais bien, comme on l'a vu, par le biais de politiques publiques associées à des instituions. Parmi ces dernières, les bibliothèques jouent un rôle actif important.
L'usage ne peut alors recouvrir que l'étude de l'ensemble des possibilités d'utilisation d'un

outil. En effet, un outil possède des fonctionnalités qui peuvent être adaptées à des contextes très différents d'usage. En proposant d'implémenter dans un outil une connaissance aussi générale que celle que nous avons proposé, on rend possible un éventail très large d'usages. Lorsque l'on propose un schéma RDFS, on se situe dans un cadre où les usages restent ouverts, bien plus que si l'on propose un outil de navigation. En effet, un schéma propose de réaliser un certain nombre d'opérations, et peut donc être adopté par différentes communautés qui pourront l'utiliser pour construire des dispositifs relativement différents.
Ainsi, les schémas publiés sont complets à partir du moment où ils déposent une connaissance de manière explicite et donc permettent la réalisation d'opérations. Ensuite, un second travail est celui effectué par les bibliothèques, centres de documentation et autre utilisateur qui vont adopter ces schémas et les inscrire dans leur propre conception de dispositif. Ainsi, les bibliothèques mettent en place des solutions originales associées à leurs besoins, et utilisant les outils à disposition, publiés dans les Linked Data. Ainsi, la publication d'un schéma n'est ni la garantie d'un usage ni d'une caractérisation des dispositifs qui vont l'utiliser.

Dans une seconde acception, les usages correspondent aux façons dont un dispositif est utilisé. On considère alors les questions en aval. Les usages concernent les dispositifs.
Les travaux sur les usages renvoient aussi en partie à la notion de communauté et aux questions d'appropriation communautaire. Il s'agit alors d'approches sociologiques. Elles permettent notamment de rendre compte de la façon dont des communautés se structurent et construisent des zones d'influence par l'appropriation ou non de la technologie. Mais pour cela, il est nécessaire que les outils soient diffusés, ce qui est encore loin d'être le cas de façon de façon significative.

Néanmoins, on peut encore tirer d'autres enseignements de notre étude et de notre projet. Au-delà des questions documentaires, l'approche pour laquelle les flux pourraient apporter le plus d'éléments est celle de la navigation. En effet, les flux explicitent les contenus portés depuis une structure vers une autre ; cette dimension est donc fondamentalement proche de la navigation, notamment dans les hypertextes. A un niveau de précision très fin, comme celui de l'interaction homme-machine, c'est dans le cadre de la construction des scénarios et la compréhension de la connaissance qui se construit dans une trajectoire que l'on pourrait au mieux utiliser les flux. Cet usage terminal, en lien avec d'usager, constitue une autre application possible du schéma.

C'est cette dichotomie entre des sciences fondées sur la conception et d'autres sur l'observation et l'analyse que l'on a essayé de dépasser en proposant de considérer que les outils élaborés pouvaient avoir une dimension culturelle fondamentale. L'opposition entre l'univers informatique et celui de l'usage (qui constitue effectivement la pratique documentaire) peut ainsi être dépassée. Ce qui nous amène à une troisième dimension de l'usage qui est l'adoption par les concepteurs de dispositifs, de ces schémas et autres outils à disposition.
Une telle structuration implique que le travail de recherche distingue ce qui est du domaine de la cognition, et qui renvoie à des connaissances structurantes et générales, et des applications contextualisées inscrivant le schéma dans une panoplie d'outils.

La méthodologie que l'on a choisie a une conséquence importante : les flux d'information ne peuvent plus être considérés comme un simple formalisme permettant de caractériser des transferts de données structurées hétérogènes, mais comme un phénomène culturel pertinent dans des contextes qui sont ceux d'activités humaines.
En justifiant les flux dans un cadre empirique d'observation et de modélisation, on passe

d'une appartenance disciplinaire à plusieurs autres. Egalement, on modifie l'usage que l'on peut faire des flux puisqu'ils ne constituent plus seulement un modèle pour la conception mais bien un outil d'analyse. En les considérant ainsi dans leur plus grande généralité, on rend possible des usages pour lesquels fondamentalement le modèle n'était pas prévu.
Ainsi, la référence à un raisonnement commun permet la compréhension du modèle et son utilisation, à savoir son passage de modèle formel au statut d'outil. Ensuite, cet outil requiert une certaine configuration pour être utilisable : l'ontologie, l'extraction d'information caractérisent la maniabilité de l'outil. Enfin, l'insertion dans un ou plusieurs dispositifs constitue la dernière étape de l'élaboration d'un usage pour un outil.

Cette structuration en trois étages (un fondement caractérisé par les flux d'information, les configurations de l'outil pour le rendre utilisable et enfin son intégration dans un dispositif opérationnel) permet de développer des projets pouvant s'inscrire dans des niveaux différents. Jusqu'à présent, nous avons développé le projet sans distinguer ces différents niveaux.
Il nous apparait néanmoins qu'ils font chacun intervenir des compétences en Sciences de l'Information et dans des disciplines proches relativement distinctes : psychologie sociale, ergonomie, analyse des usages pour les dispositifs, informatique (représentation des connaissances et TAL) pour le second, et enfin anthropologie cognitive et logique pour le premier. Ce sont aussi des compétences spécifiques de Sciences de l'Information qui sont requises à chaque étage : analyse des usages (et lien à la profession), métadonnées et description bibliographique et documentaire pour le second, classifications et organisation des connaissances pour le premier.

En cherchant à fonder un projet du web de données sur des données et des analyses de Sciences Humaines et Sociales, nous nous nous sommes rapproché du programme des Sciences Cognitives, voir du TAL, pour lesquels l'objectif consiste à faire faire par des machines des opérations supposées être celles (ou proches de celles) de l'être humain. Or, nous avons montré qu'à partir d'une certaine complexité, l'élaboration d'outils pour le web de données requiert l'intégration de la culture (entendue au sens de l'anthropologie cognitive). Cette intégration ne concerne pas l'usage (non pas que nous ne nous préoccupions pas de l'usage, mais ce n'est pas notre propos maintenant) mais la formulation même de l'identité de l'outil, de ses fonctionnalités et de ses résultats. Même si aujourd'hui la majeure partie des innovations ne reposent pas encore sur la représentation de phénomènes culturels, les questions sont posées, entre autre par A. Gangemi. Surtout, à partir du moment où le web va accompagner l'activité humaine (par exemple dans le cas de l'e-science), la question de la représentation des objets, des opérations et des raisonnements de cette activité ne pourra manquer de se poser.
Enfin, de façon très générale, nous avons voulu montrer la capacité heuristique des travaux en Sciences Humaines et Sociales, à savoir leur capacité à faire émerger des modèles qui pourront par ailleurs être exploités dans le cadre du développement du web de données.

Positionnement « méta » des sciences de l'information.
Nous venons de présenter trois niveaux distincts pour caractériser la mise en place d'une utilisation d'un modèle au départ formel à l'intérieur d'un dispositif. Nous aimerions replacer maintenant ces dimensions dans le cadre du web de données.
L'expression « méta » renvoie à la proposition de L. Floridi de considérer les Sciences de l'Information au même titre que le Logique, mais avec comme finalité la structuration des connaissances. Cette définition n'est pas totalement satisfaisante dans la mesure où elle ne couvre pas la dimension la plus proche des usagers, celle des dispositifs.

Nous avons choisi un cadre de travail, à la fois empirique et théorique, nettement plus large que celui qui généralement est utilisé en Sciences de l'Information. Cela se justifie par le fait que les bibliothèques numériques, qui sont inscrites dans le cadre du web de données, ont fondamentalement des dimensions qui vont au-delà des problématiques et du cadre théorique des seules bibliothèques. Il nous a fallu trouver et adopter des cadres théoriques et méthodologiques de portée adéquate par rapport au web de données. Ces cadres ne sont pas très courants dans les Sciences de l'Information. Un de nos objectifs aura été entre autre de les y intégrer.
Plus globalement, notre travail suit un autre mouvement initié par le web de données. Les définitions de l'information comme ensemble structuré et signifiant de données sont fondées dans le cadre des bases de données, des systèmes objet et des outils statiques d'enregistrement et de mise à disposition des données. Or, les outils sur lesquels se fonde le web de données, comme les schémas, sont d'abord des constructions dynamiques permettant la mise en relation et l'enrichissement de contenus par leur relation à d'autres. Ainsi, la construction d'un cadre théorique dynamique pour caractériser l'information constitue un enjeu crucial pour le développement de la recherche relativement au web de données.

Classification et catégorisation constituent des problématiques importantes que l'on n'a pas directement abordées dans ce travail. On a postulé l'hétérogénéité des classifications sans ne jamais véritablement avoir abordé la façon dont les classifications et catégorisations documentaires étaient structurées. L'enjeu d'un travail ultérieur pourrait être l'application de tels modèles à des classifications et des catégorisations couramment utilisées en documentation et en bibliothèques. Nous avons évoqué cette question à propos des fondements sociaux des classifications, mais sans véritablement nous y attarder. Nous n'avons pas non plus caractérisé quel pourrait être le rôle des situations pour l'analyse et la structuration des classifications. C'est un champ de recherches ouvert par ce travail.

Enfin, un des enjeux de notre travail est également de faire entrer dans le champ des Sciences de l'Information et de la Communication tout un corpus d'outils et de théories qui jusque-là ont été très peu utilisés.

Retour sur la menée de notre travail. Projets novateurs et généralisation d'une innovation.
L'ensemble des travaux auxquels nous faisons référence comme cadre théorique ont été élaborés dans un période antérieure à l'explosion des technologiques du web. La théorie des situations, comme la théorie des flux sont antérieurs à l'an 2000. Ce phénomène est commun à l'ensemble des perspectives appliquées, notamment à propos des ontologies, avec les travaux fondateurs de T.R. Gruber[309]. On pourrait encore citer la logique des prédicats qui fonde le travail de Tim Berners-Lee ou la logique aristotélicienne et les propositions de B. Smith. La maîtrise des théories précède la mise en œuvre de celles-ci dans le cadre de projets.

L'impression de multiplicité, de contradictions et de divergences d'évolution constitue une spécificité du domaine du web sémantique. Cette impression est due en partie à certaines particularités de la recherche dans le domaine et sa diffusion. En effet, nous avons affaire à une absence de distance entre le moment où la recherche aboutit et celle où ses produits (les outils sous forme de schémas) sont mis à disposition le web. Un schéma RDFS est beaucoup plus rapide à élaborer et à publier que le développement d'un analyseur, par exemple.
Par ailleurs, cette recherche est peu couteuse et ne demande qu'un nombre relativement peu

élevé d'intervenants compte tenu de l'impact qu'elle peut avoir.

Un dernier phénomène essentiel est le poids des choix des réseaux, des plateformes et des bibliothèques pour le succès ou l'échec d'une proposition. La validation et la « vérification » sont liées au degré de généralisation de l'outil sur le web. Le succès d'une entreprise de recherche n'est pas seulement lié à la qualité des outils proposés mais également à la capacité des auteurs à convaincre les acteurs (infrastructures, plateformes, moteurs de recherche) d'adopter leurs propositions.

Nous avons envisagé de présenter notre travail comme une étape dans une dynamique à la fois d'acquisition, d'organisation et d'usage d'une connaissance. Le projet est loin d'être achevé, et il nous apparait particulièrement intéressant de développer maintenant ses différentes facettes. Si aucun étage ne peut être indépendant, il n'en reste pas moins que le premier est le plus aisé à spécifier.

Caractère évolutif du web et données et fondements théoriques.

Avant de nous poser quelques questions sur notre propre démarche, il nous apparait primordial de comprendre que le contexte scientifique et sectoriel est en transformation constante. Il ne suit pas nécessairement non plus une ligne de développement qui puisse être anticipée. Cette situation ne favorise pas le développement de propositions théoriques.

Simplement, ce que l'on sait est que l'on a une nouvelle forme d'organisation scientifique. On a d'un côté des assemblages d'outils en lien à l'utilisateur final et qui sont le fait des promoteurs des infrastructures et développeurs des plateformes. Ces acteurs sont par ailleurs des producteurs d'outils autant que des utilisateurs d'outils disponibles. Le cas de CLARIN est exemplaire de cela. D'un autre côté, d'autres acteurs ne sont que producteurs d'outils et existent en dehors de toute plateforme (en dehors de la plateforme d'édition du schéma). Une part essentielle de leur travail consiste à s'allier aux plateformes et infrastructures de façon à populariser leur production. BIBO et le DC sont dans cette situation et voient leur développement dans l'adoption de leur proposition par le plus grand nombre de plateformes, portails et infrastructures.

Cette situation entraîne une transformation complète du rôle du chercheur et des propositions qu'il peut rendre publiques. Dans le cadre du web de données, une proposition est toujours une proposition d'outil, quel que soit l'étage dans lequel on se situe. Cet outil est disponible pour tous et constitue le produit d'une recherche. Economiquement, on a affaire à une logique d'offre.

Le domaine du web de données ne possède guère de régulation hormis la régulation technique associée par exemple aux critères de satisfaction des LINKED DATA. Ainsi, un outil peut perdurer sous forme de schéma sans ne jamais se réaliser dans un dispositif précis. Il peut également devenir un outil sans par ailleurs ne jamais rencontrer un usage.

Enfin, un dernier aspect réside dans la façon dont se construisent les communautés de recherche. Néanmoins, notre point de vue est celui de l'interprétation de l'information. Cette position est issue du cadre de la sémantique formelle, par laquelle on cherche à comprendre comment les expressions peuvent être interprétées, considérant la plus grande généralité possible. Ainsi, on ne s'intéresse pas à tel usager ou catégorie d'utilisateur, mais bien à la façon dont n'importe quel utilisateur peut être amené à interpréter l'information qu'il perçoit.

Au vu des disparités disciplinaires (entre les SHS et la médecine par exemple), géographiques (en vertu de politiques nationales différentes) comme également des spécificités de travail collaboratif des différents laboratoires, il apparaît malaisé de prendre le point de vue des utilisateurs comme êtres sociaux.
La généralité du point de vue de l'interprétation comme activité générique permet d'aborder l'ensemble des domaines de recherche que l'on vient de présenter. Cette position a d'abord un intérêt analytique, parce qu'elle permet d'observer et de relier des phénomènes hétérogènes comme notamment les expressions linguistiques, le cours de l'activité dans lequel ces expressions sont interprétées et enfin le raisonnement mis en œuvre dans le cadre de cette activité.
Ce point de vue permet de développer une analyse mais son principal intérêt réside dans le fait qu'il permet de fédérer différentes perspectives pour l'élaboration de projets de recherche pluridisciplinaires. La seule dimension de l'étude qui ne soit pas couverte par l'interprétation concerne le cadre des bibliothèques numériques et la structuration de leur offre.

Sémantique, information et communication.
Nous aimerions revenir dans cette conclusion sur les points sur lesquels la réflexion est loin d'être achevée :

1. En premier lieu la relation entre sémantique formelle et ontologies (et leurs corrélats lexicaux et terminologiques).
2. En second lieu, l'articulation entre cognition située et distribuée et information.

A l'heure où, comme on l'a évoqué, les environnements dans lesquels l'information est accessible changent totalement (portabilité et miniaturisation des appareils et des écrans), on ne peut plus penser l'information dans le cadre unique de l'interaction entre un utilisateur et une organisation des connaissances. Egalement, la mise en relation généralisée des données (illustrée par exemple par les structures de données des grandes bibliothèques) et les nouvelles structurations émergentes entraînent d'autres formes de consultation et de navigation. Enfin, l'apparition d'un web de travail (ou e-science) par lequel des documents hétérogènes et des outils numériques peuvent être utilisés, réemployés et où les procédures de vérification et d'exploitation des données sont communes, accélère encore cette mutation.
Si l'on prend le point de vue de l'interprétation de l'information, à savoir ce qu'il est donné à utiliser à un usager, le changement opère à la fois sur le cadre d'utilisation, les modalités d'accès et de représentation de l'information mais également l'usage même de cette information.
Notre réflexion s'appuie sur le fait que les documents et les expressions qui peuplent le web de données n'utilisent pas la seule langue naturelle, mais bien des langages normés. Ainsi, les problèmes posés par le traitement automatique de la langue sont en grande partie contournés dans le cadre de notre approche. Néanmoins, ces langages normés possèdent des propriétés sémantiques identiques à la langue naturelle et sont interprétés comme des expressions du langage naturel, au moins dans le cadre de l'activité que nous avons défini.

On ne peut manquer de conclure sur une mise en relation que nous avons opéré entre les modèles logiques et les Sciences Humaines et Sociales. Elle apparait rare en dehors de la linguistique formelle bien évidemment. Néanmoins, au vu de la puissance des modèles proposés par la logique, on ne peut manquer d'insister sur leur capacité à représenter des phénomènes réguliers observables dans le cadre activités humaines. Dans cette direction (depuis les Sciences Humaines et Sociales vers les modélisations), la puissance des modèles ne domine pas ce qu'elle prétend décrire. Ainsi, les flux et les structures d'information constituent des représentations abstraites de phénomènes observables dans le monde. La

représentation formelle permet non seulement la généralisation et l'exploitation de ces observations, mais également leur caractérisation comme une connaissance collective opérationnelle.

Au sein de la pharmacie hospitalière, on aurait pu modéliser d'autres phénomènes, comme par exemple la temporalité et les révisions de croyances. Nous aurions pu également explorer plus précisément le lien entre la singularité des patients et les populations de la base de données USC*PACK. Nous étions plus intéressé par les questions d'information.
Par ailleurs, le web s'est jusqu'à présent développé en considérant que le monde des données n'était pas soumis aux contraintes et aux contingences du monde ordinaire. De ce fait, il n'a guère eu à s'intéresser aux activités menées dans le monde. Or, comme on l'a vu avec l'e-science, la situation pourrait bien changer dès lors que le sujet, voire même des équipes, partagent leur travail avec le web.
Nous avons largement utilisé des travaux de sémantique formelle, pour deux raisons : d'une part, il s'agit de la seule interface développée entre les phénomènes humains et les représentations formelles, d'autre part, les questions d'informations sont imbriquées à celles de l'interprétation du discours, comme nous l'avons vu tout au long de ce travail.
Si l'on en revient encore un instant aux liens entre activité, structure d'information et flux, c'est pour montrer que la solidarité entre ces dimensions permet de réévaluer le rôle des ontologies et la structuration des outils du web dans le cadre de la prise en charge de dimensions procédurales au sein du web.

7.2. Consortiums, équipes et organisation de la recherche.

Nous avons présenté les transformations de la recherche liés notamment à l'e-science. Il convient maintenant de proposer un regard réflexif, inspiré notamment par notre présentation de projets auprès de l'ANR.

Comme il est de règle dès lors que l'on propose une innovation (M. Akrich, M., Callon, M., & Latour, B. *Sociologie de la traduction: textes fondateurs,* op.cit), il faut que celle-ci soit adoptée, à savoir reconnue par le corps social et qu'elle puisse grandir dans un cadre qui assure sa promotion.
Or, ce cadre possède plusieurs dimensions : une dimension économique et une dimension politique.
En premier lieu, comme on l'a vu dans la première partie, le développement du web de données dépend largement d'un contexte politique, qui définit des priorités de recherche et de développement, et des moyens mis à disposition des infrastructures et des bibliothèques pour un développement des dispositifs.

Le développement de cette recherche dépend également du coût de la diffusion de l'innovation. Avec le web de données, on constate une évolution des moyens et des coûts de la diffusion d'une innovation. La configuration nouvelle de la publication de la recherche développée par les LINKED DATA reconfigure complètement la façon dont on peut penser un résultat scientifique sous forme d'outil. La publication d'un travail dans le cadre des LINKED DATA consiste à éditer et rendre disponible pour différents usages un schéma accomplissant un certain nombre d'opérations.

De ce fait, on amoindrit considérablement la distance entre un produit de recherche et un produit industriel. Par ailleurs, un outil n'est pas élaboré en fonction d'un usage unique mais

peut être utilisé pour des usages différents par des acteurs variés. Il suffit alors qu'un usager quelconque, cherchant à mettre en place un dispositif, utilise cet outil.

Ainsi, le dépôt d'un schéma apparaît comme une première étape dans le cycle de vie d'un outil. L'usage permettra de redéfinir ses fonctionnalités et dirigera son évolution.

Cette question se traduit également par un positionnement par rapport à l'e-science. En effet, ce vaste projet sert à donner des réponses numériques et automatiques à des questions relatives à la menée et à l'évaluation de la recherche. L'e-science s'intéressera par exemple à concevoir un outil numérique permettant la découverte de molécules nouvelles[310]. Elle s'intéresse, comme on l'a vu, assez peu à la dimension documentaire, au moins en termes de recherche (comme on l'a vu, elle intéresse particulièrement les professionnels de l'information). La plupart des travaux concernant le document sont relatifs à son exploration, sa compréhension et sa contextualisation. Ils ne renvoient pas à un projet de structuration de l'information. En ce sens, notre projet devrait effectivement permettre de réintroduire plus spécifiquement la dimension documentaire dans le cadre de l'e-science.
Ce positionnement appelle une nouvelle question. On a évoqué l'importance de l'analyse des activités dans le cadre de l'e-science, notamment à propos des protocoles et tests médicaux.

Comme on l'a également vu, le web de données hérite de traditions scientifiques diverses qui chacune apporte certaines propositions à l'édifice. Le domaine des bibliothèques s'intéresse plus spécifiquement à la mise en relation des données et des structurations de connaissances, et commence seulement à articuler ces questions d'offre documentaire avec le cours des activités professionnelles des utilisateurs. Il serait effectivement intéressant de poursuivre l'idée d'une connexion constante entre l'offre documentaire et l'activité. C'est en cela que l'on pourra effectivement traduire la volonté des bibliothécaires scientifiques de s'inscrire entièrement à l'intérieur des protocoles d'e-science : on proposerait alors un accompagnement informationnel à chaque étape de la formulation, de la réalisation et de la valorisation d'un projet de recherche.
En définitive, si l'e-science transforme considérablement la pratique scientifique (entre autre en ne définissant plus la publication comme le seul objet public relatif à une recherche), on peut également considérer que l'accès à l'information et son usage dans le cours d'une activité de recherche sont en cours de modification. Les conséquences que cela peut avoir sur l'organisation et la représentation de l'information sont encore largement à définir.

En définitive, le rôle imparti au concept d'information dans l'élaboration et la mise en œuvre de recherches et de projets liés aux bibliothèques numériques dans le cadre du web de données est essentiel parce qu'il permet de fédérer des travaux de Sciences Humaines et Sociales (notamment les analyses d'activité et les analyses linguistiques), des domaines de recherche fondés sur la représentation formelle de phénomènes (comme la logique), l'informatique et l'Intelligence Artificielle et enfin les Sciences de l'Information.

La disponibilité de l'information scientifique et son accès aisé ont constitué depuis plus d'une dizaine d'années une situation exceptionnelle permettant d'envisager des rapprochements inédits, d'exploiter des ressources jusque-là peu explorées, notamment dans le cadre des Sciences de l'Information. Nous avons pleinement exploité cette opportunité qui a permis de constituer un programme de recherches à partir de ressources hétérogènes mais liées par des fondations et des problématiques similaires.

On aura remarqué, notamment à propos de l'e-science, mais également des ontologies, que le domaine biomédical est celui dans lequel les avancées en matière de propositions d'outils web tout autant que de structuration de l'information sont les plus visibles. Cette avancée s'explique en partie par une forte culture de la structuration de l'information, notamment relativement à la planification des publications, à la publicité faite aux ressources primaires et au financement de la recherche. Par ailleurs, ces sciences ont une dimension pluridisciplinaire importante : on y trouve de l'informatique, des sciences sociales et humaines et des Sciences de l'Information. Par conséquent, elles jouent un rôle pionnier dans la construction des ressources numériques, de leur accès et de leur usage.

Nous avons essayé tout au long de ce travail de montrer les liens entre des perspectives relativement différentes, de façon à rendre compte d'un phénomène multiforme comme peut l'être l'information. Dans ce cadre, les aspects méthodologiques et de réalisation de ce genre de positionnement entraînent un temps de travail et des moyens que nous n'avons pas encore réunis. De plus, nous avons insisté sur l'intérêt de la dimension pluridisciplinaire de ce type de travail : la conséquence en est un temps beaucoup plus long de mise en place et d'effort d'intercompréhension. Enfin, notre travail a largement consisté à élaborer et étayer des hypothèses. Il demande néanmoins une validation soit sur des quantités de données, soit par l'usage d'une théorie de la preuve.

ANNEXE 1

La validation empirique de notre application des flux sera déclinée en trois sortes de cadres :

- une situation de vie quotidienne, qui constitue un exemple didactique
- une situation où l'exemple précédent est mis en série (où la même information que celle présentée précédemment est insérée dans un cadre de gestion de document)
- Enfin, une situation de travail où l'exemple s'insère dans une activité et donne lieu à une action.

a. Exemple de scène de la vie quotidienne : circulation d'informations en gare.

J'attends sur le quai de gare le train 1048 en provenance de Guéret, dans lequel se trouve Gégé. Le train devait arriver à 4H55.
A 4H50, une voix annonce : « Le train 1048 en provenance de Guéret aura 30mns de retard ».
Je prends le téléphone et j'annonce à ma famille : « le train de Gégé aura une ½ heure de retard ».

Cet exemple représente une circulation ordinaire d'information.

Pour intégrer la dimension temporelle, il faut ajouter un nouveau contexte, celui de l'archivage des productions informationnelles. On distingue ainsi plus facilement les composants de l'information (les expressions, leur production, transfert et interprétation) et ceux de la description de cette information (les outils permettant de les classer et de les archiver). Evidemment, on montrera que ces deux structures sont articulées par une relation instance/type.

b. Mise en série de l'exemple.

Nous reprenons l'exemple de la circulation d'informations en gare et envisageons la précédente information dans une série ; cette série permet l'archivage des informations de gestion du trafic.
Imaginons maintenant que les informations soient enregistrées et stockées de façon à donner lieu à des études statistiques concernant le retard des trains ou la communication en gare. De façon à ce que le document soit conservé, il faudra alors le classer dans [annonce de retard], la série dans lequel il se situe, son objet de référence [le train 1048] et enfin le moment et le lieu concernés. La classification que l'on vient de présenter préserve l'information parce qu'elle respecte la forme syntaxique et l'intégralité du message ; elle utilise les marqueurs lexicaux en inscrivant les objets et les propriétés dans des types.
Mais parallèlement, cette information peut être diffusée dans un autre contexte, où elle aura la même signification (l'annonce d'un retard) mais permettra d'autres inférences (sur la fréquence des incidents de trains par exemple). En ce sens, cette information n'aura plus une durée de vie limitée à la première situation : elle sera en quelque sorte archivée. Néanmoins, elle reste la même information parce qu'elle porte syntaxiquement et sémantiquement la mémoire des situations dans lesquelles elle s'est constituée et a été diffusée. Ce qui lui permet d'intégrer un autre contexte et un autre flux d'information (celui qui caractérise des collections d'informations sur le trafic), c'est le fait qu'elle a été décrite. La description de l'information permet sa

classification, sa mémorisation et sa transmission au sein d'une collection. Cette hypothèse sous-tend l'ensemble du travail.

c. Conditions pour une généralisation de l'exemple.

Enfin, notre exemple peut sembler relatif à une situation sociale limitée ; nous examinons maintenant les conditions de sa généralisation. Précédemment, la mise en série avait permis de mémoriser l'information décrite. L'opération consiste maintenant à éliminer toutes les traces de contexte pour obtenir une représentation aisément transposable du phénomène que l'on vient d'évoquer.

Comment passer de la circulation d'informations en gare aux marchés financiers ou à la pharmacie hospitalière ?
Dans les trois cas, il s'agit de permettre l'interprétation d'états de fait localisés distinguant radicalement le lieu d'émission de celui de réception, mais avec une action à distance sur le lieu de production de cette information. (En effet, un avis pharmaceutique s'applique sur le traitement du patient, tout comme un ordre de bourse s'applique sur les marchés d'origine de l'information et non dans le bureau du trader. Nous y reviendrons).

Lorsque l'on considère ce type de message, on doit postuler un contexte interprétatif invariable quelle que soit l'instance : il existe un contexte initial passé d'information partagée T0, un contexte prévisible futur T2, et un contexte T1 correspondant au moment où l'information est diffusée.
« Il existe un train x parti de y à T0 et devant arriver à z à T2 ». « Ce train est dans un lieu w à T1 ».
De la même façon un résultat d'analyse est obtenu à propos d'un patient dans un certain état à T1, après un certain état et un traitement à T0, en vue d'une éradication à T2.
Enfin, une information à propos d'un événement sur un marché en T1 est relative à un état T0 et à une action dans cet événement en T2.
(Ce même contexte est nécessaire pour l'interprétation du second message ; il peut donc être décrit par une certaine structure récurrente : *TRAIN(x), PROVENANCE (LOC. Y, T0), ARRIVEE (LOC Z :* T2*), SITUATION (LOC W, T1)).* La SITUATION est dans cet exemple le retard.

Cette situation sert de cadre à l'information, à savoir que l'état nouveau sera transmis parce que l'on a une situation historique et une fonction in fine. Ce cadre constant permet alors le déploiement de résultats ponctuels, décrivant une situation limitée dans le temps.
La formulation précédente peut être généralisée et adaptée à des contextes différents :

PATIENT(x), PROPRIETE (x, T0, T1, T2), RESULTAT (x, T2), SITUATION (T1)).
MARCHE (x), INDICE Y (x, T0, T1, T2), RESULTAT (T2), SITUATION (T1)).

On remarquera que cette représentation contient une durée qui correspond au déroulement de l'activité. T2 représente le fait que le RESULTAT modifie l'état des connaissances du récepteur dans un temps postérieur à l'obtention (par analyse et mesure) de ce résultat. (Par exemple, entre le moment où l'estimation est réalisée et

celui où elle sert à envisager une stratégie par celui qui attend en gare, la situation initiale a pu changer).

Cette formulation du contexte peut apparaître très insuffisante ; elle sera spécifiée plus loin, lorsqu'il s'agira effectivement de modéliser le contexte d'interprétation. Néanmoins, elle permet de cerner un objet d'étude par une schématisation des principaux paramètres de l'information.

BIBLIOGRAPHIE

[1] Schöpfel, J., and H. Prost. *Développement et Usage des Archives Ouvertes en France. Rapport. 1e partie: Développement*'. Tech. rep., Université Charles-de-Gaulle Lille 3, 2010.

[2] Manon E., Janik J., Feltin G E-Science, « perspectives et opportunités pour de nouvelles pratiques de la recherche en informatique et mathématiques appliquées », *i-expo 2011*, Paris : France (2011) - http://hal.archives-ouvertes.fr/hal-00611166

[3] http://ec.europa.eu/research/infrastructures/pdf/esfri_evaluation_report_2011.pdf

[4] Eccles, K., Schroeder, R., Meyer, E. T., Kertcher, Z., Barjak, F., Huesing, T., & Robinson, S. (2009). "The Future of e-Research Infrastructures". In *Proceedings of NCeSS International Conference on e-Social Science, Cologne, June* (pp. 24-26).

[5] Abney S., Bird S., "The human language project: building a Universal Corpus of the world's language"s, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp.88-97, 2010, Association for Computational Linguistics

[6] Ding, Y., & Fensel, D. (2001, July). Ontology Library Systems: The key to successful Ontology Reuse. In *SWWS* (pp. 93-112).

[7] http://www.ontobee.org/

[8] d'Aquin, M. and Noy, N. F. (2012). Where to publish and find ontologies? A survey of ontology libraries, in*Web Semantics: Science, Services and Agents on the World Wide Web*, 11 pp. 96–111.

[9] Romano, J. V.; Lopez, Allen; and Phi, Maianh. (2012). "Understanding eScience: Reflections on a Houston Symposium." *Journal of eScience Librarianship* 1(2): Article 6. http://dx.doi.org/10.7191/jeslib.2012.101

[10] Hamasu, Claire; Jones, Barb; and Kelly, Betsy. (2012). "Discussing "eScience and the Evolution of Library Services"." *Journal of eScience Librarianship* 1(2): Article 5. http://dx.doi.org/10.7191/jeslib.2012.1015

[11] S. Chambers S., Schallier W. , *Bringing Research Libraries into Europeana: Establishing a Library-Domain Aggregator*, Liber Quarterly 20 (1), September 2010 – ISSN: 1435-5205. P105–118, http://liber.library.uu.nl/, Igitur, Utrecht Publishing & Archiving Services

[12] Lossau, N. *An Overview of Research Infrastructures in Europe - and Recommendations to LIBER*. LIBER Quarterly**,** North America, 21, apr. 2012. Available at: <http://liber.library.uu.nl/index.php/lq/article/view/URN%3ANBN%3ANL%3AUI%3A10-1-113632>. Date accessed: 13 Jun. 2013.

[13] Collins, E., Jubb, M.. *Information handling in collaborative research*. LIBER Quarterly, North America, 22, feb. 2013. Available at: <http://liber.library.uu.nl/index.php/lq/article/view/URN%3ANBN%3ANL%3AUI%3A10-1-114291>. Date accessed: 13 Jun. 2013.

Code de champ modifié

Code de champ modifié

Code de champ modifié

Code de champ modifié

Code de champ modifié

[1414] Romary L.,*Stabilizing knowledge through standards - A perspective for the humanities*, CoRR,abs/1011.0519,2010.

[15] Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., & Trippel, T. (2012).*CMDI: a Component Metadata Infrastructur*. Talk presented at LREC. Istanbul. 2012-05-22. (Autre version: http://www.balisage.net/Proceedings/vol7/html/Broeder01/BalisageVol7-Broeder01.html et http://www.clarin.eu/node/3219)

Code de champ modifié

Code de champ modifié

[16] Antoine Isaac, « Web sémantique, web de données – Pôle 2 - Entre thésaurus et ontologies : une affaire d'interopérabilité et d'alignement » *Documentaliste*, Volume 48, N° 4, paru le 17 janvier 2012, page(s) 48-49 - - See more at: http://www.adbs.fr/web-semantique-web-de-donnees-pole-2-entre-thesaurus-et-ontologies-une-affaire-d-interoperabilite-et-d-alignement-112547.htm?RH=REVUE#sthash.TIbQ1C7O.dpuf

[17] Antoine Isaac, Raphaël Troncy, Designing and using an audio-visual description core ontology in Workshop on Core Ontologies in Ontology Engineering, 2004/10, http://www.eurecom.fr/~troncy/Publications/Troncy_Isaac-coront04.pdf

Code de champ modifié

[18] Tom Heath and Christian Bizer (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool

[19] Nikolov, Andriy and d'Aquin, Mathieu (2011). Identifying relevant sources for data linking using a semantic web index. In: WWW2011 Workshop: Linked Data on the Web (LDOW 2011) at 20th International World Wide Web Conference (WWW 2011), 29 March 2011, Hyderabad, India.

[20] Steiner, Raphael Troncy Thomas, Hausenblas, Michael, Auer, Sören, Decker, Stefan, Hauswirth, Manfred, « How Google is using Linked Data Today and Vision For Tomorrow », 2010, http://CEUR-WS.org/Vol-700/Paper5.pdf

Code de champ modifié

[21] Ines Bannour and Haïfa Zargayouna, "Une plate-forme open-source de recherche d'information sémantique, in "CORIA'12",, pp.167-178, 2012.

[22] Djouadi, Yassine,Généralisation des opérateurs de dérivation de Galois en recherche d'information basée sur l'analyse formelle de concepts, in CORIA 2012, pp. 373-386, 2012

[23] d'Aquin, Mathieu; Gridinoc, Laurian; Angeletou, Sofia; Sabou, Marta and Motta, Enrico (2007). Watson: a gateway for next generation semantic web applications. In: *The 6th International Semantic Web Conference (ISWC 2007)*, 11-15 Nov 2007, Busan, Korea.

[24] Giovanni Tummarello and Szymon Danielczyk and Richard Cyganiak and Renaud Delbru and Michele Catasta and Ecole Polytechnique Federale and Stefan Decker},Sig.ma: Live views on the Web of data, inProc. WWW-2010}, 2010},pp.1301—1304, ACM Press

[25] Rose Dieng-Kuntz and Olivier Corby. Conceptual Graphs for Semantic Web Applications. In F. Dau, M.-L. Mugnier, and G. Stumme, editors, *Conceptual Structures: Common Semantics for Sharing Knowledge, Proc. of the 13th Int. Conference on Conceptual Structures (ICCS'2005)*, pages 19–50, Kassel, Germany, July 17-23 2005. Springer-Verlag, LNAI 3596.

[26] Stelios Piperidis, The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions, in LREC 2012 proceedings, 2012,pp. 36-42.

[27] Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Harris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz and Valérie Mapelli},The META-SHARE Metadata Schema for the Description of Language Resources}, in LREC proceeding, 2012, pp.1090-1097.

[28] Antoine Zimmermann, Markus Krötzsch, Jérôme Euzenat, and Pascal Hitzler. 2006. Formalizing Ontology Alignment and its Operations with Category Theory. In *Proceedings of the 2006 conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006)*, Brandon Bennett and Christiane Fellbaum (Eds.). IOS Press, Amsterdam, The Netherlands, The Netherlands, 277-288.

[29] Emmanuel Nauer and Yannick Toussaint, Classification dynamique par treillis de concepts pour la recherche d'information sur le web, in CORIA, 2008}, pp.71-86.

[30] Auer, S., & Hellmann, S. (2012). The web of data: Decentralized, collaborative, interlinked and interoperable. In *8th International Conference on Language Resources and Evaluation*.

[31] Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., & Windhouwer, M. (2012). Standardizing a component metadata infrastructure. In N. Calzolari (Ed.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, May 23rd-25th, 2012* (pp. 1387-1390). European Language Resources Association (ELRA).

[32] Gregor Thurmair and Vera Aleksic and Christoph Schwarz, Large Scale Lexical Analysis, In N. Calzolari (Ed.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, May 23rd-25th, 2012* (pp. 1387-1390). European Language Resources Association (ELRA).

[33] McCrae J, Montiel-Ponsoda E, Cimiano P. Collaborative semantic editing of linked data lexica. Presented at the The 2012 International Conference on Language Resource and Evaluation (LREC)

[34] Jérôme Dinet (2009). Pour une conception centrée-utilisateurs des bibliothèques numériques. Communication et langages, 2009, pp 59-74 doi:10.4074/S0336150009003068

[35] Smith, B., & Ceusters, W. (2010). Ontological realism: A methodology for coordinated evolution of scientific ontologies. *Applied ontology*, *5*(3), 139-188.

[36] Van Hooland, S., Verborgh, R., De Wilde, M., Hercher, J., Mannens, E., & Van de Walle, R. (2011). Free your metadata: Integrating cultural heritage collections through Google Refine reconciliation. *Pre-submission paper available on. http://freeyourmetadata. org/publications/freeyourmetadata. pdf.*

[37] Collier, J. (2011). Information, causation and computation. *Information and Computation: Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation*, *2*, 89.

[38] R. G. D'Andrade, (1989), "Cultural Cognition", in R. Posner, *Foundations of Cognitive Sciences*, MIT Press

[39] Theureau, J. (2011). L'observatoire des cours d'action, des cours de vie relatifs à une pratique et de leurs articulations collectives. *Approches pour l'analyse des activités*, 23

[40] G Antoniou, F van Harmelen, *A Semantic Web Primer* MIT Press, 2004

[41] Wulf, J., Jorm, D., Casperson, M., & Newson, L. (2012, April). Automated Assembly of Custom Narratives from Modular Content using Semantic Representations of Real-world Domains and Audiences. In *Workshop on the Semantic Publishing (SePublica 2012) 9 th Extended Semantic Web Conference Hersonissos, Crete, Greece, May 28, 2012* (p. 60).

[42] Domingue, John and Fensel, Dieter (2009). Toward a service web: integrating the Semantic Web and service orientation. IEEE Intelligent Systems, 23(1), pp. 86–88.

[43] http://www.w3.org/TR/ws-arch/#whatis

[44] Dumitru Roman, Uwe Keller, Holger Lausen, Jos de Bruijn, Rubén Lara, Michael Stollberg, Axel Polleres, Cristina Feier, Cristoph Bussler, Dieter Fensel, *Web Service Modeling Ontology* in Applied Ontology, Volume 1, Number 1/2005, IOS Press, pp. 77-106.

[45] Pedrinaci, C., Domingue, J., & Sheth, A. (2010). Semantic web services. *Handbook of Semantic Web Technologies*, *2*, 977-1035

[46] Pour une présentation des principaux courants, voir notamment :
Barsalou, L.W. (2005). Abstraction as dynamic interpretation in perceptual symbol systems. In L. Gershkoff-Stowe & D. Rakison (Eds.), *Building object categories* (389-431). Carnegie Symposium Series. Majwah, NJ: Erlbaum.
Smith, B., "Beyond Concepts, or: Ontology as Reality Representation", Achille Varzi and Laure Vieu (eds.), *Formal Ontology and Information Systems. Proceedings of the Third International Conference (FOIS 2004)*, Amsterdam: IOS Press, 2004, 73–84
Guarino, N. and Welty, C. **An Overview of OntoClean** in S. Staab, R. Studer (eds.), Handbook on Ontologies, Springer Verlag 2004, pp. 151-172
Partee, B., 1999. Semantics. In *The MIT Encyclopedia of the Cognitive Sciences*, ed. R.A. Wilson and F.C. Keil, 739-742. Cambridge, MA: The MIT Press.

[47] Ceusters, W., Smith, B., & Flanagan, J. (2003). Ontology and medical terminology: Why description logics are not enough. *Towards an Electronic Patient Record (TEPR 2003), Boston, MA*.

[48] Merrill, G. H. (2010). Ontological realism: Methodology or misdirection?. *Applied Ontology*, *5*(2), 79-108.

[49] Andy Clark **&** David J. Chalmers (1998). The Extended Mind. *Analysis* 58 (1):7-19.

[50] Barsalou, L.W., & HALE, C.R., (1993). Components of conceptual representation : From feature lists to recursive frames. In I. Van Mechelin, J. Hampton, R. Michalski, & P. Theuns (Eds.), *Categories and Concepts: Theoretical views and inductive data analysis* (97-144). San Diego, CA: Academic Press.

[51] Barsalou, L.W., Situated simulation in the human conceptual system, in *Language and Cognitive Processes*, 2003, 18 (5/6), 513-562.

[52] Pour son ouvrage fondateur :
Hutchins, E. (1995). *Cognition in the Wild*. Cambridge: Massachusetts Institute of Technology Press.

[53] Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, Freeman.

[54] Pour une synthèse des travaux dans ce domaine :
J. A. Van Der Lubbe (1997), *Information Theory*, Cambridge : Cambridge University Press.

Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié

[55] Kokar, M. M., Matheus, C. J., & Baclawski, K. (2009). Ontology-based situation awareness. *Information fusion*, *10*(1), 83-98.

[56] Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies*, *43*(5), 625-640.

[57] Kratzer, Angelika, "Situations in Natural Language Semantics", *The Stanford Encyclopedia of Philosophy* (Fall 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2011/entries/situations-semantics/>.

[58] Michel CHAMBREUIL (sous la direction de) : *Sémantiques*. Hermès, 1998

[59] Partee, B. H. (1996). The development of formal semantics in linguistic theory. *The Handbook of Contemporary Semantic Theory*, 11-38.

[60] Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and philosophy*, *4*(2), 159-219.

[61] Stalnaker, Robert, *Inquiry* Cambridge, MA: MIT Press, 1984.

[62] Vendler, Z. (1967). *Linguistics in Philosophy*. Cornell : Cornell University Press.
Vendler, Z. (1972). *Res Cogitans: An Essay in Rational Psychology*. Cornell : Cornell University Press.

[63] http://jjayez.pagesperso-orange.fr/

Code de champ modifié

[64] Fauconnier, G., & Sweetser, E. (Eds.), (1996) ***Spaces, Worlds, and Grammar***. Chicago University Press. Fauconnier, G. (1985). ***Mental spaces: Aspects of meaning construction in natural language***. Cambridge: MIT Press.

[65] Jayez, Jacques & Mari, Alda (2005) «Togetherness» *Proceedings of SuB9* Emar Maier, Corien Bary, and Janneke Huitink (éds), 155-169
Jayez, Jacques (1999) Underspecification, context selection and generativity *The Language of Word Meaning* P. Bouillon et F. Busa (éds), Cambridge, Cambridge University Press, 124-148

[66] Laura Kallmeyer and Aravind K. Joshi (2003) Factoring Predicate Argument and Scope Semantics: Underspecified Semantics with LTAG. *Research on Language and Computation*, 2003, 1:1-2, 3-58.

Code de champ modifié

[6767] Johan Bos. 2004. Computational semantics in discourse: Underspecification, resolution, and inference. *Journal of Logic, Language and Information*, 12(2).

[68] Pinkal, M. (1985). *Logic and Lexicon*. Oxford : Oxford University Press.

[69] Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge MA : MIT Press.

[70] Copestake, A. & Briscoe, T. (1995). Semi-productive Polysemy and Sense Extension. *Journal of Semantics 12*, pp. 15-67.

[71] Massimo Poesio and David Traum, "Conversational Actions and Discourse Situations," *Computational Intelligence*,v. 13, n.3, 1997.

[72] König E. and U. Reyle. 1997. A general reasoning scheme for underspecified representations. In H. J. Ohlbach and U. Reyle, eds, Logic and its Applications. Festschrift for Dov Gabbay. Part I. Kluwer, Dordrecht, Holland.

[73] Akrich, M., Callon, M., & Latour, B. (2006). *Sociologie de la traduction: textes fondateurs*. Presses des Mines.

[74] Vinck, D. (2009). De l'objet intermédiaire à l'objet-frontière. *Revue d'anthropologie des connaissances*, *3*(1), 51-72.

[7575] Floridi, L. (2004). LIS as Applied Philosophy of Information: A Reappraisal.*Library Trends*, *52*(3), 658-665.

[76] Goldman, A. (2010). Why social epistemology is real epistemology. *Social epistemology*, 1-28

[77] Tarciso Zandonade, Social Epistemology from Jesse Shera to Steve Fuller LIBRARY TRENDS, Vol. 52, No. 4, Spring 2004, pp. 810–832

[78] Newell, A. & Simon, H. (1972). Human Problem Solving. Englewood Cliffs, NJ: Prentice-Hall.
Ericsson, K. & Simon, H. (1984). Protocol Analysis. Cambridge, MA: MIT Press.

[79] Pacherie, E. 2001. Peut-on penser l'objectivité sans l'espace? In F. Wolff (éd.), *Philosophes en liberté - Positions et Arguments 1*, Paris : Ellipses, pp. 46-66.

[80] Luciano Floridi & Jeff Sanders, The Method of Abstraction, invited chapter for the *Yearbook of the Artificial* (Issue II, 2004, Peter Lang) dedicated to "Models in contemporary sciences", pp. 177-220.

[81] Floridi, Luciano, "The Method of Levels of Abstraction", in *Minds and Machines*, Volume 18, Number 3 / September 2008, Springer Netherlands

Code de champ modifié

[82] Cardelli, L., & Wegner, P. (1985). On understanding types, data abstraction, and polymorphism. *ACM Computing Surveys (CSUR)*, *17*(4), 471-523.

[83] Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends in cognitive sciences*, *4*(5), 197-207.

[84] Madsen, B. N., & Thomsen, H. E. (2009). Ontologies vs. classification systems. *NEALT PROCEEDINGS SERIES VOL. 7*, 27.
Madsen, B. N., & Erdman Thomsen, H. Ontologies vs. Classification Systems. In *NODALIDA 2009. The 17th Nordic Conference of Computational Linguistics* (pp. 27-32).

[85] Blackburn, Patrick, Rijke, Maarten and Venema, Yde, *Relational methods in logic, language and information* (1995), technical report, CWI (Centre for Mathematics and Computer Science, Amsterdam, The Netherlands.

[86] Kripke, S. (1965). Semantical analysis of intuitionistic logic I. *Formal systems and recursive functions*, 92-130.

[87] Steedman, M. (2000). The productions of time. *Draft. Available at http://www. cogsci. ed. ac. uk/steedman/papers. html*.

[88] Steedman, M. (2004). Where Does Compositionality Come From?. In *Compositional Connectionism in Cognitive Science, Papers from the 2004 AAAI Fall Symposium, Technical report FS-04-03* (pp. 59-62).

[89] Harel, D., Tiuryn, J., & Kozen, D. (2000). *Dynamic logic*. MIT press.

[90] R. Muskens, J. van Benthem and A. Visser, "Dynamics", in van Benthem, J. F. A. K., & Ter Meulen, A. (Eds.). (1996). *Handbook of logic and language*. Access Online via Elsevier.

[91] Van Benthem, J. (2003). Logic and the Dynamics of Information. *Minds and Machines*, *13*(4), 503-519.

[92] Blackburn, P. (2000). Representation, reasoning, and relational structures: a hybrid logic manifesto. In *Logic Journal of IGPL*.

[93] Gärdenfors, P. (2005). *How logic emerges from the dynamics of information* (pp. 83-108). Springer Netherlands.

[94] Couzinet, V. (2001). Jean Meyriat, théoricien et praticien de l'information-documentation. *ADBS éditions, Paris*.
Fondin, H. (2005). La science de l'information ou le poids de l'histoire. *Les Enjeux de l'information et de la communication*.
Pédauque, R. T. (2003). Document: forme, signe et médium, les re-formulations du numérique.

[95] Bar-Hillel Y. 1964, *Language and Information* (Reading, Mass.; London: Addison-Wesley).
Bar-Hillel, Y., & Carnap, R., (1952). *A outline of a Theory of Semantic Information*, Technical Report n° 247, Cambridge, MIT.

[96] Laporte, É. (2005). In memoriam Maurice Gross. *Archives of Control Sciences*, *15*(3), 257-278.

[97] Dalrymple, M., Lamping, J., Pereira, F., & Saraswat, V. (1994). A deductive account of quantification in LFG. *arXiv preprint cmp-lg/9404009*.

[98] Halliday, M. A. K. Mathiessen, Christian M. I. M. *An Introduction to Functional Grammar*, 3rd edition Arnold 2003

[99] L. Floridi, On Defining Library and Information Science as Applied Philosophy of Information, *Social Epistemology* 2002 (16.11), 37-49.

[100] Manuel E. Bremer, Do Logical Truths Carry Information? Minds and Machines, Volume 13, Number 4 / novembre 2003, Springer Netherlands, pp. 567-575

[101] L. Floridi, Semantic Conceptions of Information, The Stanford Encyclopedia of Philosophy (Winter 2005 Edition).

[102] Barwise, J., Perry, J, 1983, Situation and attitudes, Cambridge, Mass. MIT Press

[103] L. Floridi, Is Information Meaningful Data?, *Philosophy and Phenomenological Research*, 2005, 70.2, 351-370.

[104] Dodig-Crnkovic, G. (2005, April). System modeling and information semantics. In *Proceedings of the fifth promote IT conference*. Janis Bubenko, Owen Eriksson, Hans Fernlund, and Mikael Lind (Studentlitteratur: Lund).

Code de champ modifié

Code de champ modifié

Code de champ modifié

Code de champ modifié

Code de champ modifié

Code de champ modifié

[105] Klein, G. O., & Smith, B. (2010). Concept Systems and Ontologies: Recommendations for Basic Terminology. *Transactions of the Japanese Society for Artificial Intelligence= Jinko Chino Gakkai ronbunshi*, *25*(3), 433.

[106] Gregory Chaitin, The Halting Probability Omega: Irreducible Complexity in Pure Mathematics, *Milan Journal of Mathematics*, Vol. 75, 2007, pp. 291-304

[107] Chaitin, Gregory. (2011). How real are real numbers?. *Manuscrito*, *34*(1), 115-141. Retrieved December 17, 2013, from http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-60452011000100006&lng=en&tlng=en. 10.1590/S0100-60452011000100006.

[108] Devlin, K. J. (1995). *Logic and information*. Cambridge University Press.

[109] Dretske, F. (1981). *Information and the Flow of the Information*, Stanford, CSLI Publications.

[110] Bawden, D. (2007, July). Organised complexity, meaning and understanding: an approach to a unified view of information for information science. In *Aslib Proceedings* (Vol. 59, No. 4/5, pp. 307-327). Emerald Group Publishing Limited.

[111] Burgin, M. (2005). Is information some kind of data. *Petitjean, M.(Ed.)*.

[112] Johan Van Benthem, Logic and the Dynamics of Information, *Minds and Machines* Volume 13 , Issue 4 (November 2003) pp.503 - 519

[113] Recanati, F. (2004). *Literal meaning*. Cambridge University Press.

[114] GÃĪrdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. The MIT press.

[115] Dretske, F. (1981). *Information and the Flow of the Information*, Stanford, CSLI Publications.

[116] Grice, H. P. (1957). Meaning. *The philosophical review*, *66*(3), 377-388.

[117] Laurier, Daniel, « Fonction d'indication et sélection naturelle », in *Intellectica*, n° 21, 1995/2, pp. 135-158

[118] Luciano Floridi, "Information", invited contribution to the *Encyclopedia of Science, Technology, and Ethics*, (ESTE) edited by Carl Mitcham (Macmillan) 2004. http://www.philosophyofinformation.net/publications/pdf/este.pdf.

Code de champ modifié

[119] Perry, J. (2001). The problem of the essential indexical. *ADVANCES IN CONSCIOUSNESS RESEARCH*, *30*, 143-162.

[120] Visetti, Y. M. (1995). Fonctionnalismes 1996. *Intellectica*, *21*, 282-311.

[121] Guizzardi, Giancarlo, and Terry Halpin. "Ontological foundations for conceptual modelling." *Applied Ontology* 3.1 (2008): 1-12.

[122] Barwise, J. & Seligman, J., (1997). *Information flow: the logic of Distributed Systems*, Cambridge University Press.

[123] Kalfoglou, Y., & Schorlemmer, M. (2005). Using Formal Concept Analysis and Information Flow for modelling and sharing common semantics: lessons learnt and emergent issues. In *Conceptual Structures: Common Semantics for Sharing Knowledge* (pp. 107-118). Springer Berlin Heidelberg.

[124] Wybraniec-Skardowska, U. (2007). Meaning and Interpretation. I. *Studia Logica*, *85*(1), 105-132.

[125] Review by J. Van Benthem & D. Israel, in *Journal of Logic, Language and Information*, July 1999, Volume 8, Issue 3, pp 390-397

Code de champ modifié

[126] Kent, Robert E. "Semantic integration in the information flow framework." *Semantic Interoperability and Integration* 4391 (2005).

[127] Desclés, Jean-Pierre, and Ismail Biskri. "Logique combinatoire et linguistique: grammaire catégorielle combinatoire applicative." *Mathematiques Informatique et Sciences Humaines* 132 (1995): 39-68.

[128] Moortgat, Michael. *Categorial grammar and formal semantics*. John Wiley & Sons, Ltd, 2002.

[129] Fernando, Tim. "Three processes in natural language interpretation." *Reflections on the Foundations of Mathematics: Essays in Honor of Solomon Feferman. Natick, Mass.: Association for Symbolic Logic* (2002): 208-227.

[130] Jayez, Jacques, and Daničle Godard. "True to fact (s)." In *Proceedings of the 12th Amsterdam Colloquium*, pp. 151-156. 1999.

[131] Jeff Speaks Truth theories, translation manuals, and theories of meaning in Linguistics and Philosophy, Volume 29, Number 4 / août 2006, Springer Netherlands pp.487-505

Code de champ modifié

Code de champ modifié

[132] von Heusinger, Klaus 1999. *Intonation and Information Structure. The Representation of Focus in Phonology and Semantics*. Habilitationsschrift. Universität Konstanz

[133] Vallduví, Enric (1993) *Information packaging: A survey*. Report of the Word Order, Prosody, and Information Structure Initiative, University of Edinburgh.

[134] Kruijff-Korbayová, Ivana, and Mark Steedman. "Discourse and information structure." *Journal of Logic, Language and Information* 12.3 (2003): 249-259.

[135] Harris, Z. (1988). *Language and Information*, New York : Columbia University Press.

[136] Pawlak, Z. (1981). Information systems theoretical foundations. *Information systems*, *6*(3), 205-218

[137] Collier, J. 2010. *Information, causation and Computation*. World Scientific. chapter Information and Computation: Es-says on Scientific and Philosophical Understanding of Foun-dations of Information and Computation.

[138] Ingwersen, Peter, and Kalervo Järvelin. *The turn: Integration of information seeking and retrieval in context*. Vol. 18. Springer, 2005.

[139] Pour un état de l'art relatif au contexte, voir la présentation de P. Brézillon : http://www-ftp.lip6.fr/lip6/reports/2002/lip6.2002.010.pdf

[140][140] Abiteboul, S., Bienvenu, M., Galland, A., & Antoine, É. (2011, June). A rule-based language for web data management. In *Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 293-304). ACM.

[141] Paolo Bouquet, Jérôme Euzenat, Chiara Ghidini, Deborah L. McGuinness, Valeria de Paiva, Luciano Serafini, Pavel Shvaiko, and Holger Wache, *Proceedings of the 2007 workshop on Contexts and Ontologies Representation and Reasoning(C&O:RR-2007),* 2007, Roskilde University Computer Science Research Report 115

[142] Natalya F. Noy Semantic Integration: A Survey Of Ontology-Based Approaches, in *SIGMOD Record*, Vol. 33, No. 4, December 2004 pp. 65 - 70

[143] Krötzsch, M., P. Hitzler, et al. (2005). Category Theory in Ontology Research: Concrete Gain from an Abstract Approach. Karlsruhe, Germany, Institut AIFB, Universität Karlsruhe: 6.

[144] http://www.loa-cnr.it/DOLCE.html

[145]http://nl.ijs.si/et/talks/esslli02/, http://nl.ijs.si/et/talks/esslli02/esslli02-4.html

[146] Agre Philip, Computation and Human Experience, Cambridge University Press, 1997.

[147] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi et P.F. Patel-Schneider. The description logic handbook. Cambridge university press,2007.

[148] Kent, R. E. (2011). The information flow framework: A descriptive category metatheory. *arXiv preprint arXiv:1108.4133*.

[149] Kalfoglou, Y., & Schorlemmer, M. (2003). IF-Map: An ontology-mapping method based on information-flow theory. In *Journal on data semantics I* (pp. 98-127). Springer Berlin Heidelberg.

[150] Schorlemmer, M., & Kalfoglou, Y. (2008). Institutionalising ontology-based semantic integration. *Applied Ontology*, *3*(3), 131-150.

[151] Dapoigny, R., & Barlatier, P. (2007). Goal reasoning with context record types. In *Modeling and Using Context* (pp. 164-177). Springer Berlin Heidelberg.

[152] Zimmermann, A., Krötzsch, M., Euzenat, J., & Hitzler, P. (2006). Formalizing ontology alignment and its operations with category theory. In *Proc. 4th International conference on Formal ontology in information systems (FOIS)* (pp. 277-288).

[153] Beaver, D., van Benthem, J., & Scotto di Luzio, P. (2002). *Words, proofs, and diagrams*. CSLI Publications.

[154] RDF (http://www.w3.org/RDF/) et OWL (http://www.w3.org/2004/OWL/).

[155] http://www.w3.org/2004/02/skos/

[156] Kutz, O., Mossakowski, T., & Lücke, D. (2010). Carnap, Goguen, and the Hyperontologies: Logical Pluralism and Heterogeneous Structuring in Ontology Design. *Logica Universalis*, *4*(2), 255-333.

Code de champ modifié

Code de champ modifié

Code de champ modifié

[157] Joseph A. Goguen and Rod M. Burstall. 1992. Institutions: abstract model theory for specification and programming. *J. ACM* 39, 1 (January 1992), 95-146. DOI=10.1145/147508.147524 http://doi.acm.org/10.1145/147508.147524

[158] Ken Herold An Information Continuum Conjecture, Minds and Machines, Volume 13, Number 4 / novembre 2003, pp. 553-566

[159] Jon Barwise & Gerard Allwein (ed.), Logical reasoning with diagrams, Oxford University Press , Oxford, 1996,

[160] François Recanati, "Relativized Propositions", in *Situating Semantics, Essays on the Philosophy of John Perry*, Edited by Michael O'Rourke and Corey Washington, MIT Presse, 2007, pp. 119-154

[161] Gibson, J. J. The ecological approach to visual perception, Laurence Erlbaum, 1979.

[162] Recanati, F. (1997). The dynamics of situations. *European Review of Philosophy*, *2*, 41-75.

[163] Jerry Seligman and Larry Moss. *Situation theory*. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*. North Holland and MIT Press, 1997

[164] Barwise, J., & Cooper, R. (1993). Extended Kamp Notation: a graphical notation for situation theory. *Situation Theory and its Applications*, *3*, 29-53.

[165] Cooper, R. (2012). Type theory and semantics in flux. *Handbook of the Philosophy of Science*, *14*, 271-323.

[166] Israel, D. J., & Perry, J. (1991). *What is information?*. CSLI. From *Information, Language and Cognition*, edited by Philip Hanson, Vancouver: University of British Columbia Press, 1990: 1-19.]

[167] Cooper, Robin (2000) Information States, Attitudes and Dependent Record Types, in *Logic, Language and Computation, Volume 3*, ed. by Lawrence Cavedon, Patrick Blackburn, Nick Braisby and Atsushi Shimojima, CSLI Publications.

[168] Recanati, F. (2007). *Perspectival Thought: A Plea for (Moderate) Relativism: A Plea for (Moderate) Relativism*. Oxford University Press.

[169] Recanati, F. (2006). Relativized propositions. *Situation semantics: Essays on the philosophy of John Perry. Cambridge, MA: MIT Press/Bradbord Books*.

[170] Recanati, F., 1996: Domains of Discourse. *Linguistics and Philosophy* 19: 445-475.

[171] Allwein, G., & Barwise, J. (1996). *Logical reasoning with diagrams*. Oxford University Press, Inc..

[172] Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3), 243-276.

[173] Jackendoff, R. (1983). *Semantics and Congnition* (Vol. 8). MIT press.

[174] Vallduv, Enric. "Information packaging: A survey." *Report prepared for Word Order, Prosody, and Information Structure. Centre for Cognitive Science and Human Communication Research Centre, University of Edinburgh* (1993).

Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié

[175] Stalnaker, R. (1998). On the representation of context. *Journal of Logic, Language and Information*, *7*(1), 3-19.

[176] Peregrin, Jaroslav. "Topic and focus in a formal framework." *Discourse and Meaning* (1996): 235-254.

[177] Sgall, P., Hajicová, E., & Panevová, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Springer.

[178] Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, 91-136.

[179] I.Kruijff-Korbayov, Information Structure, MPI, November 2007, http://www.coli.uni-saarland.de/~korbay/Courses/MPI07/mpi3-064092.pdf

Code de champ modifié

[180] Halliday, Michael AK. "Notes on transitivity and theme in English: Part 2." *Journal of linguistics* 3.02 (1967): 199-244.

[181] *coral.lili.uni-bielefeld.de/Courses/.../Halliday.ppt*

[182] Chafe,W. L. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C.N. Li (ed.) *Subject and Topic*, 25-55. New York, San Francisco and London: Academic Press.

[183] Vallduví, E. (1993). The informational component. *IRCS Technical Reports Series*, 188.

[184] Choi, H. W. (1997). Information structure, phrase structure, and their interface. In *Proceedings of the LFG97 Conference. Stanford: CSLI Publications*.

[185] Heim, Irene. "File change semantics and the familiarity theory of definiteness." (1983): 164-189.

[186] Webber, Bonnie, Aravid Joshi, Matthew Stone, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29:545–587

[187] Paper presented at the ESSLLI 2001 Workshop on *Information Structure, Discourse Structure and Discourse Semantics*, Edited by Ivana Kruijff-Korbayov´a and Mark Steedman. Voir : http://www.coli.uni-saarland.de/~korbay/esslli01-wsh/Proceedings/proc-for-cdrom.pdf

[188] Cook, P., & Bildhauer, F. (2011). Annotating information structure: The case of topic. *Beyond Semantics. Corpus-based Investigations of Pragmatic and Discourse Phenomena. Bochumer Lin--guistische Arbeitsberichte*, *3*, 45-56.

[189] Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998, August). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 86-90). Association for Computational Linguistics.

[190] Jackendoff, R. (1992). *Semantic structures* (Vol. 18). The MIT Press.

[191] Moldovan, D., Badulescu, A., Tatu, M., Antohe, D., & Girju, R. (2004, May). Models for the semantic classification of noun phrases. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics* (pp. 60-67). Association for Computational Linguistics.

[192] Godard, D., & Jayez, J. (1999). Quels sont les faits?. *M. Plénat, et al*, 117-136.

[193] Moortgat, M. (2002). *Categorial grammar and formal semantics*. John Wiley & Sons, Ltd.

[194] Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation?. *AI magazine*, *14*(1), 17.

[195] Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine learning*, *34*(1-3), 233-272.

[196]Freitag, D. (1998, August). Toward general-purpose learning for information extraction. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 404-408). Association for Computational Linguistics.

[197] Kushmerick, N. (1997). *Wrapper induction for information extraction* (Doctoral dissertation, University of Washington).

[198] Kosala, R., Van den Bussche, J., Bruynooghe, M., & Blockeel, H. (2002). Information extraction in structured documents using tree automata induction. In *Principles of Data Mining and Knowledge Discovery* (pp. 299-311). Springer Berlin Heidelberg.

[199] Grishman, R. (2012). Information Extraction: Capabilities and Challenges. *Lecture Notes]. Retrieved from http://cs. nyu. edu/grishman/tarragona. pdf*.

[200] **Jayez**, Jacques ; Mari, Alda (2005). Togetherness. In Maier, E.; Bary, C. & Huitink, J. (eds.) Proceedings Sinn und Bedeutung 9: 155-169

[201] Ingwersen, P. (1995), "Information and information science", in Kent, A. (Ed.), Encyclopedia of Library and Information Science, Vol. 56, Marcel Dekker Inc., New York, NY, pp. 137-74.

[202] Popper, K.R. (1977), "The worlds 1, 2, and 3", in Popper, K.R. and Eccles, J.C. (Eds), Self and its Brain, Springer, Berlin.

[203] Mark Burgin: Foundations of Information Theory. CoRR abs/0808.0768 (2008)

Code de champ modifié

[204] Clark, A., & Chalmers, D. (1998). The extended mind. *analysis*, *58*(1), 7-19.

[205] Salembier, P. (2002). Cadres conceptuels et méthodologiques pour l'analyse, la modélisation et l'instrumentation des activités coopératives situées. *Systèmes d'information et Management (SIM)*, (2), 37-56.

[206] Hutchins, E. (2000). Distributed cognition. *Internacional Enciclopedia of the Social and Behavioral Sciences*.

[207] Clark, A., & Chalmers, D. (1998). The extended mind. *analysis*, *58*(1), 7-19.

[208] Wilson, R. A., & Clark, A. (2009). How to situate cognition: Letting nature take its course. *The Cambridge handbook of situated cognition*, 55-77.

[209] Havelange, V., Lenay, C., & Stewart, J. (2002). Les représentations: mémoire externe et objets techniques. *Intellectica*, *35*(2), 115-129.

[210] Roy G. D'Andrade. 1989. Cultural cognition. In *Foundations of cognitive science*, Michael I. Posner (Ed.). MIT Press, Cambridge, MA, USA 795-830.

[211] Hutchins, E. (2005). Material anchors for conceptual blends. *Journal of pragmatics*, *37*(10), 1555-1577.

[212] Keller, C. M., & Keller, J. D. (1996). *Cognition and tool use: The blacksmith at work*. Cambridge University Press.

[213] Goodwin, C. (1994). Professional vision. *American anthropologist*, *96*(3), 606-633.

[214] CICOUREL, A. V. (1994). La connaissance distribuée dans le diagnostic médical. *Sociologie du travail*, *36*(4), 427-449.

[215] Vygotsky L. S. (1985a) *Pensée et langage.* Trd. F. Sève, Paris, Éditions Sociales.
[216] De Terssac, G. (2012). AUTONOMIE ET TRAVAIL. *Dictionnaire du travail*, 47-53.

[217] Goodwin, C., & Goodwin, M. H. (1996). Seeing as situated activity: Formulating planes. *Cognition and communication at work*, 61-95.

[218] Dubucs, J. (1995). Les états mentaux sur la place publique. *Intellectica*, *2*(21), 115-134.

[219] Burge, T. (1986). Intellectual norms and foundations of mind. *The Journal of Philosophy*, *83*(12), 697-720.

[220] Burge, T. (2009). Perceptual objectivity. *Philosophical Review*, *118*(3), 285-324.

[221][221] Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, *104*(3), 192.

[222] Suchman, L. (1987). *Plans and Situated Actions: the problem of human-machine communication.* Cambridge: Cambridge University Press.

Lave, J. (1988). *Arithmetics in Practice*. Cambridge, MA: Harvard University Press.

[223] Dana H. Ballard, Mary M. Hayhoe, Polly K. Pook and Rajesh P. N. Rao (1997). Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences, 20, pp 723-742.

[224] Sutton, J. (2004). Representation, levels, and context in integrational linguistics and distributed cognition. *Language Sciences*, *26*(6), 503-524.

[225] Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends in cognitive sciences*, *4*(5), 197-207.

[226] Kripke, S. A. (1973). *Naming and necessity* (pp. 253-355). Springer Netherlands.

[227] Kaplan, D. (1979). On the logic of demonstratives. *Journal of philosophical logic*, *8*(1), 81-98.

[228] Quine, W. V. (2013). *Word and object*. MIT press.

[229] Chalmers, D. (2004). The representational character of experience. *The future for philosophy*, 153-81.

[230] Aydede, M. (2000). On the type/token relation of mental representations. *Facta Philosophica: International Journal of Contemporary Philosophy*, *2*(1), 23-49.

[231] Chalmers, D. J. (2004). Imagination, indexicality, and intensions. *Philosophy and Phenomenological Research*, *68*(1), 182-190.

[232] Chalmers, D. J. (2002). On sense and intension. *Nous*, *36*(s16), 135-182.

[233] Chalmers, D. J. (1994). The components of content. (http://cogprints.org/321/1/content.html)

Code de champ modifié

[234] Chalmers, D. J. (2004). 13 Phenomenal Concepts and the Knowledge Argument. *There's Something About Mary: Essays on phenomenal consciousness and Frank Jackson's knowledge argument*, 269.

[235] Chalmers, D. (2006). The foundations of two-dimensional semantics. *Two-dimensional semantics*, 55-140.

[236] FRANÇOIS, J. (2003). La Faculté de langage: travaux récents d'inspiration fonctionnaliste, sur son architecture, ses universaux, son émergence et sa transmission. *CoReLa*, *1*(1).

[237] Agre, P. (1997). *Computation and human experience*. Cambridge University Press.

[238] Clancey, W. J. 1993. The Knowledge Level Reinterpreted: Modelling Socio-Technical Systems. *International Journal of Intelligent Systems*, **8**: 33-49.

[239] Devlin, K and Rosenberg, D. *Language at Work: Analyzing Communication Breakdown in the Workplace to Inform System Design*, Stanford, CA: CSLI Publications (1996).

[240] de Carvalho, E. C. A., Jayanti, M. K., Batilana, A. P., Kozan, A. M., Rodrigues, M. J., Shah, J., ... & Pietrobon, R. (2010). Standardizing clinical trials workflow representation in UML for international site comparison. *PloS one*, *5*(11), e13893.

[241] Addison, J., Whitcombe, J. and William Glover, S. (2013), How doctors make use of online, point-of-care clinical decision support systems: a case study of *UpToDate*©. Health Information & Libraries Journal, 30: 13–22. doi: 10.1111/hir.12002

[242] Ketchum AM, Saleh AA, Jeong K (2011) Type of evidence behind point-of-care clinical information products: a bibliometric analysis. J Med Internet Res 13(1): e21.

[243] Alper BS, White DS, Ge B. Physicians answer more clinical questions and change clinical decisions more often with synthesized evidence: a randomized trial in primary care. Ann Fam Med. 2005;3(6):507-513

[244] Coppola, B., Gangemi, A., Gliozzo, A., Picca, D., & Presutti, V. (2009). Frame detection over the semantic web. In *The Semantic Web: Research and Applications* (pp. 126-142). Springer Berlin Heidelberg.

[245] http://www.vivoweb.org/files/presentations/12Thu/2012_08_22_VIVO_Conference_Karma_v01.pdf

Code de champ modifié

[246] W. Marco Schorlemmer: Ontology Modularity, Information Flow, and Interaction-Situated Semantics - Extended Abstract. WoMO 2010: 5-10

[247] J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer, 2007.

[248] Kuchinke, W., Karakoyun, T., & Ohmann, C. (2012). Deliverable 6.2 Clinical Research Information Model.

[249] Parra, C., Villegas, M., & Bel, N. (2010). The basic metadata description (bamdes) and theharvestingday. eu: Towards sustainability and visibility of lrt. In *Proceedings of workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management at LREC* (pp. 49-53).

[250] Ball, A. (2009). Scientific data application profile scoping study report.

[251] Žumer, M., Zeng, M. L., & Hlava, M. M. (2012, June). A Domain Model for Describing and Accessing KOS Resources: Report of Processes in Developing a KOS Description Metadata Application Profile. In *DC-2012, Kuching, Sarawak, Malaysia*.

[252] Van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., & Gardellini, M. (2010). Virtual Language Observatory: The Portal to the Language Resources and Technology Universe. In *LREC*.

[253] Qin, J., Ball, A., & Greenberg, J. (2012). Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data. In *Twelfth International Conference on Dublin Core and Metadata Applications*. University of Bath.

[254] http://www.skipcr.cz/akce-a-projekty/dokumenty/akm-2011/Slavic.pdf

Code de champ modifié

[255] Romary, L., & Armbruster, C. (2010). Beyond Institutional Repositories. *International Journal of Digital Library Systems (IJDLS)*, *1*(1), 44-61.

[256] Murillo, A., & Greenberg, J. (2012). Data-at-Risk, Metadata Registration for Data, and Dryad.

[257] Claus Zinn and Peter Wittenburg and Jacquelijn Ringersma, An Evolving eScience Environment for Research Data in Linguistics, in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010, Valletta, Malta, Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis and Mike Rosner and Daniel Tapias ed., European Language Resources Association (ELRA).

[258] Reid, R., Pignotti, E., Edwards, P., & Laing, A. (2010, April). ourSpaces: linking provenance and social data in a virtual research environment. In *Proceedings of the 19th international conference on World wide web* (pp. 1285-1288). ACM.

[259] Dey, A.K.: Understanding and using context, In *Personal and Ubiquitous Computing Journal* 1(5), 4-7

[260] Goble, C., & De Roure, D. (2002). The grid: an application of the semantic web. *ACM SIGMOD Record*, *31*(4), 65-70.

[261] Gamble, M., & Goble, C. (2010). Standing on the shoulders of the trusted web: trust, scholarship and linked data. In *Proceedings of the Web Science Conference*.

[262] Schorlemmer, M., & Kalfoglou, Y. (2008). Institutionalising ontology-based semantic integration. *Applied Ontology*, *3*(3), 131-150.

[263] Pettifer, S., McDermott, P., Marsh, J., Thorne, D., Villeger, A., & Attwood, T. K. (2011). Ceci n'est pas un hamburger: modelling and representing the scholarly article. *Learned Publishing*, *24*(3), 207-220.

[264] Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*.

[265] Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., ... & Goble, C. (2011). Why linked data is not enough for scientists. *Future Generation Computer Systems*.

[266] A. van der Aalst, A. Hofstede, and M. Weske. *Business process management: A survey.* In M. Weske, editor, Business Process Management, volume 2678 of Lecture Notes in Computer Science, pages 1019{1019. Springer Berlin / Heidelberg, 2003. 10.1007/3-540-44895-0 1.

[267] Belhajjame, K., Corcho, O., Garijo, D., Zhao, J., Missier, P., Newman, D., ... & Goble, C. (2012). Workflow-centric research objects: First class citizens in scholarly discourse. In *Proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web*.

[268] http://www.w3.org/TR/2004/REC-rdf-mt-20040210/

[269] http://www.essepuntato.it/lode/http://purl.org/spar/fabio

[270] Chiarcos, C., Hellmann, S., & Nordhoff, S. (2011). Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group. *TAL*, *52*(3), 245-275.

[271] http://ontologydesignpatterns.org/wiki/Submissions:ConceptTerms

[272] http://lexinfo.net/, Paul Buitelaar, Philipp Cimiano, Peter Haase and Michael Sintek, Towards Linguistically Grounded Ontologies in L. Aroyo & al., *The Semantic Web: Research and Applications* Lecture Notes in Computer Science, 2009, Volume 5554/2009, 111-125, DOI: 10.1007/978-3-642-02121-3_12

[ii] http://www.monnet-project.eu/Monnet/Monnet/English?init=true, Montiel-Ponsoda E., G. Aguado de Cea, A. Gomez-Perez, W. Peters, 2011, Enriching ontologies with multilingual information, Natural Language Engineering, 17, pp 283-309

[iii] McCrae John, Aguado de Cea G., 2010, Lemon: Linked Data, Lexicons and Data Category Registrie, in Terminology and Knowledge Engineering -l Workshop Standardizing Data Categories in ISOcat: Implementing Group Work for Thematic Domains, TKE 2010, Dublín, Ireland.

[273] Harris, Z. (1988), *Language and Information*, New York : Columbia University Press.

[274] Picca, D., Gliozzo, A. M., & Gangemi, A. (2008, May). LMM: an OWL-DL MetaModel to Represent Heterogeneous Lexical Knowledge. In *LREC*.

[275] Ginzburg, J., & Sag, I. A. (2000). *Interrogative investigations* (pp. 247-254). Stanford: CSLI publications.

[276] Peroni, S., Gangemi, A., & Vitali, F. (2011, September). Dealing with markup semantics. In *Proceedings of the 7th International Conference on Semantic Systems* (pp. 111-118). ACM.

[277] Boulos, M. N., Roudsari, A. V., & Carson, E. R. (2002). Towards a semantic medical Web: HealthCyberMap's tool for building an RDF metadata base of health information resources based on the Qualified Dublin Core Metadata Set. *Medical science monitor: international medical journal of experimental and clinical research*, *8*(7), MT124.

[278] Baker, T. (2012). Libraries, languages of description, and linked data: a Dublin Core perspective. *Library Hi Tech*, *30*(1), 116-133.

[279]Sivaram Arabandi, Helen Moran , ClinicalKey: Terminology driven Semantic Search. Proceedings of the 3rd International Conference on Biomedical Ontology (ICBO 2012), KR-MED Series, Graz, Austria, July 21-25, 2012. http://ceur-ws.org/Vol-897/demo_9.pdf

[280] Speaks, J. (2006). Is mental content prior to linguistic meaning?. *Nous*, *40*(3), 428-467.

[281] Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., ... & Rosse, C. (2005). Relations in biomedical ontologies. *Genome biology*, *6*(5), R46.

Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié
Code de champ modifié

[282] Ferrario, R., & Guarino, N. (2009). Towards an ontological foundation for services science. In *Future Internet–FIS 2008* (pp. 152-169). Springer Berlin Heidelberg.

[283] http://www.meta-net.eu/meta-share/META-SHARE%20%20documentationUserManual.pdf

Code de champ modifié

[284] Picca, D., Gliozzo, A. M., & Gangemi, A. (2008, May). LMM: an OWL-DL MetaModel to Represent Heterogeneous Lexical Knowledge. In *LREC*.

[285] http://neurocommons.org/w/images/2/24/Ruttenberg_First_IAO_Workshop_Slides.pdf

Code de champ modifié

[286] http://icbo.buffalo.edu/Presentations/Ruttenberg.pdf

Code de champ modifié

[287] Gangemi, A., & ISTC-CNR, R. A semiotic metamodel for bridging lexical and formal semantics.

[288] Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*.

[289] Gangemi, A. (2012). Hybridizing formal and linguistic semantics for the Multilingual Semantic Web.. In P. Buitelaar, P. Cimiano, D. Lewis, J. Pustejovsky & F. Sasaki (eds.), *MSW*, : CEUR-WS.org.

[290] Davidson, D. (1967). The Logical Form of Action Sentences. In *The Logic of Decision and Action* (2nd ed.). Pittsburgh: University of Pittsburgh Press.

[291] Ovchinnikova, E., Vieu, L., Oltramari, A., Borgo, S., & Alexandrov, T. (2010, May). Data-Driven and Ontological Analysis of FrameNet for Natural Language Reasoning. In *LREC*.

[292] Nédellec, C., & Nazarenko, A. (2006). Ontologies and information extraction. *arXiv preprint cs/0609137*.

[293] Hobbs, J. R., & Riloff, E. (2010). Information extraction. *Handbook of natural language processing*, *2*.

[294] Øyvind Raddum Berg, Stephan Oepen, and Jonathon Read, Towards High-Quality Text Stream Extraction from PDF, Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, pages 98–103, Jeju, Republic of Korea, 10 July 2012.

[295] Saint-Dizier, P. (2012). DISLOG: A logic-based language for processing discourse structures. In *LREC* (pp. 2770-2777).

[296] Moraes, S., & Lima, V. (2012). Combining Formal Concept Analysis and semantic information for building ontological structures from texts: an exploratory study. In *LREC* (pp. 3653-3660).

[297] http://www.essepuntato.it/lode/imported/http://www.essepuntato.it/2008/12/earmark

Code de champ modifié

[298] Serrano, L., Bouzid, M., Charnois, T., & GRILHERES, B. (2012). Vers un système de capitalisation des connaissances: extraction d'événements par combinaison de plusieurs approches. *SOS-DLWD'2012 at EGC*.

[299] Kim JD, Ohta T, Tsujii J: **Corpus annotation formining biomedical events from literature.** *BMC Bioinf* 2008, **9:**10.

[300] Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60-76.

[301] Zweigenbaum, P., Demner-Fushman, D., Yu, H., & Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. *Briefings in bioinformatics*, *8*(5), 358-375.

[302] Miwa, M., Thompson, P., McNaught, J., Kell, D. B., & Ananiadou, S. (2012). Extracting semantically enriched events from biomedical literature. *BMC bioinformatics*, *13*(1), 108.

[303] Van Hage, W. R., Malaisé, V., de Vries, G., Schreiber, G., & van Someren, M. (2009, October). Combining ship trajectories and semantics with the simple event model (sem). In *Proceedings of the 1st ACM international workshop on Events in multimedia* (pp. 73-80). ACM.

[304][304] Biber, D. 1995. On the role of computational, statistical, and interpretive techniques in multi-dimensional analyses of register variation: A reply to Watson (1994). *Text* 15.341-370

Code de champ modifié

[305] Aggarwal, N., Asooja, K., & Buitelaar, P. (2012, June). DERI&UPM: Pushing corpus based relatedness to similarity: Shared task system description. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (pp. 643-647). Association for Computational Linguistics.

[306] Etzioni, O., Fader, A., Christensen, J., Soderland, S., & Mausam, M. (2011, July). Open information extraction: The second generation. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume One* (pp. 3-10). AAAI Press.
Piskorski, J., Belayeva, J., & Atkinson, M. (2011, September). Exploring the Usefulness of Cross-lingual Information Fusion for Refining Real-time News Event Extraction: A Preliminary Study. In *RANLP* (pp. 210-217).

[307] LAPUJADE, Gilles Kassel—Pascal Lando—Anne et FÜRST, Frédéric. Des Artefacts aux Programmes.
Kassel, G. (2009). Vers une ontologie formelle des artefacts. *Actes des 20es Journées Francophones en Ingénierie des Connaissances*.

[308] Sawyer, S., & Chen, T. T. (2002, November). Conceptualizing Information Technology in the Study of Information Systems: Trends and Issues. In *Global and organizational discourse about information technology* (pp. 109-131).

[309] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199-220.

[310] A.J.G. Gray, S. Askjaer, C.Y.A. Brenninkmeijer, K. Burger, C. Chichester, J. Eales, C.T.A. Evelo, C.A. Goble, P.T. Groth, L. Harland, A. Loizou, S. Pettifer, R. Ramgolam, M. Thompson, A. Waagmeester, and A.J. Williams, The Pharmacology Workspace: A Platform for Drug Discovery. ;In *Proceedings of ICBO*. 2012.