A travers les ontologies : vers de nouveaux agencements pour l'organisation et l'accès à l'information scientifique

Monique Commandré, Pierre-Michel Riccio

A travers les ontologies : vers de nouveaux agencements pour l'organisation et l'accès à l'information scientifique

Monique COMMANDRÉ* et Pierre-Michel RICCIO**
Monique.Commandre@univ-perp.fr
* Laboratoire VECT de l'Université de Perpignan Via Domitia
Avenue Foch, F-48000 Mende

Pierre-Michel.Riccio@mines-ales.fr
** Centre de recherche LGI2P de l'Ecole des Mines d'Alès
Parc Scientifique Georges Besse, F-30035 Nîmes cedex 1

Résumé : La prolifération d'informations accessibles via les réseaux numériques pose la question du niveau de performativité des ontologies comme dispositif de médiation entre des données préexistantes et un projet d'apprentissage et / ou de veille scientifique.
Cette contribution qui se situe au niveau de l'analyse des interrelations entre des outils d'indexation, des modèles de structuration induits par les différentes générations d'ontologies et les modalités de construction des connaissances par différentes communautés de pratique a pour objet d'esquisser une ingénierie de la médiation intégrée pour l'organisation et l'accès à l'information scientifique.

Mots clés : sciences de l'information et de la communication, organisation des connaissances, ontologies, réseaux sémantiques, documents numériques

Introduction

La prolifération d'informations accessibles via les réseaux numériques pose la question du niveau de performativité des ontologies (au sens de modèles de connaissances) comme dispositif de médiation entre des données préexistantes et un projet d'apprentissage et / ou de veille scientifique. L'usage intensif des réseaux numériques conduit les utilisateurs à laisser un grand nombre de traces, de données, qui constituent un réservoir d'information potentiellement intéressant pour les personnes en quête d'informations. Mais ces données préexistantes sont en général peu qualifiées et faiblement datées. Les acteurs en quête d'informations sont tentés de collecter via des moteurs de recherche généralistes ou spécialisés ces données pour construire un parcours adapté. Mais, si l'information est abondante, l'absence de qualification rend l'exercice difficile et dans un grand nombre de cas périlleux. Pour remédier à ce problème la tendance ces dernières années consiste à développer des systèmes de médiation de type ontologies ou métadonnées qui ont pour objet de faciliter la tâche de recueil d'une information de qualité exploitable. Ces outils intermédiaires sont toutefois réservés à des groupes d'initiés et généralement d'usage difficile.

En s'appuyant sur une étude de cas : construction d'une ontologie dans le domaine médical et élaboration d'un dispositif de recherche fondé sur les réseaux sémantiques étendus, nous proposons dans cet article d'éclairer les modes possibles de construction d'une architecture à trois niveaux (parcours, métadonnées, données) qui facilitera l'accès d'un large public à une information qualifiée. Cette contribution se situe au niveau de l'analyse des interrelations entre des outils d'indexation, des modèles de structuration induits par les différentes

générations d'ontologies et les modalités de construction des connaissances par des communautés de pratique. Elle a pour finalité d'esquisser une ingénierie de la médiation intégrée pour l'organisation et l'accès à l'information scientifique.

Considérants

Notre contribution au débat sur l'organisation et l'accès à l'information scientifique portera sur les problématiques d'extraction et d'appropriation de l'information dans des communautés de pratiques. Plus précisément nous interrogerons les dispositifs de médiation entre l'information accessible et les acteurs porteurs d'une intention d'apprentissage et /ou de veille scientifique. Le dispositif sera ici considéré dans la lignée des travaux de la sociologie de l'innovation, elle-même héritière de la pensée foucaldienne. Ainsi, le terme de dispositif sera utilisé pour désigner tous les assemblages sociotechniques d'humains et de non-humains auxquels s'intéressent ces sociologues, qu'il s'agisse de décrire les « programmes d'action » (Latour, 1996) ou les « scripts » (Akrich, 1992) inscrits dans des objets, ou encore « des assemblages d'éléments hétérogènes d'énoncés, d'agencements techniques, de compétences incorporées » (Callon, 1995).

Notre propos se place à l'intersection entre des objets technologiques complexes et des usages motivés par un projet d'apprentissage et / ou de veille scientifique. Pour ce faire nous convoquons une approche croisée témoignant d'une collaboration interdisciplinaire entre les sciences de l'ingénieur et les sciences de l'information et de la communication. Dans ce cadre nous limitons notre contribution à une étude praxéologique des significations.

Dés lors, nous nous intéresserons moins à l'appropriation des connaissances qu'à la construction de significations construites par les acteurs dans leur environnement professionnel. Dans ce contexte, la connaissance est lisible, visible, mobilisable et se distingue des savoirs « savants » de types encyclopédiques. Chez Jean Piaget la connaissance relève d'un processus d'intégration-organisation de savoirs à des représentations déjà existantes chez le sujet. De ce processus résulte une connaissance dynamique traduisant la mobilisation de ces savoirs, à titre déclaratif ou procédural. On retrouve par ailleurs dans la recherche en psychologie cognitive le concept de « modèle de situation ». Kintsch par son analyse de l'activité de lecture établit un modèle cognitif basé sur les processus de « construction / intégration » (Chevalier et Tricot, 2007). Ce modèle offre une vision en trois temps de l'activité de lecture de documents numériques : la structure de surface (la forme linguistique du texte comme les informations syntaxiques ou lexicales), la représentation sémantique (soit la décomposition propositionnelle du texte ainsi que les relations entre ses propositions témoignant de l'intention de l'auteur) et le modèle de situation qui attesterait d'une véritable compréhension et de l'intégration à des significations déjà établies. Ce modèle pourrait certainement être rapproché du concept de « connaissances instituées » chez Berger et Luckman (Berger et Luckman, 1986).

Notre étude peut dés lors considérer le caractère incarné des connaissances, ou « énacté » en référence aux travaux de Francisco Varéla. Leur construction s'inscrit dans un environnement complexe porteur de propositions de significations. L'acteur utilise l'environnement pour intégrer de nouvelles connaissances à des significations préexistantes.
Nous nous intéressons aux « connaissances actionnables ». Dans le champ de l'intelligence artificielle, et notamment chez Herbert Simon (Simon, 1957), les connaissances actionnables sont des informations identifiées dans un contexte opératoire. Elles sont produites par l'action et produisent de l'action. La définition donnée par Jean-Louis Le Moigne nous donne une

vision contemporaine de ce néologisme « depuis son introduction dans la littérature organisationnelle par D. Schön en 1983 (actionable knowledge), il semble accepté par l'usage, plus aisément peut-être que connaissance-processus proposé en 1967 par J. Piaget : il a certes l'inconvénient de privilégier de façon apparemment exclusive l'usage (voire l'utilitarisme) de telle connaissance (le savoir-faire ?), aux dépens de sa genèse et de sa production (le savoir pur, qui serait pure spéculation ?) » (Le Moigne, 2000). Cette notion de connaissance actionnable nous libère de l'exclusivité d'une analyse cognitive des phénomènes d'appropriation des connaissances.
Celle-ci est ici ramenée à une dynamique de mutualisation des connaissances réinvesties dans l'action par le collectif. La situation est délimitée dans un champ d'action : un projet, un programme, une action finalisée.

Nous nous intéressons aux « connaissances ordinaires » construites en situation, incarnées dans un environnement (physique, symbolique ou technologique), amplifiées et affinées par l'interaction. Nous illustrerons notre approche par des cas relevant du secteur scientifique.
Nous interrogeons les connaissances actionnables au sein de communautés de pratique, dans leur rapport aux données organisées au sein d'interfaces d'édition numérique.
Les interfaces d'édition numérique rendent disponibles des stocks de données indexées dans une logique de référencement permettant de les situer ou de les localiser. La question n'est pas nouvelle puisque la problématique de la recherche d'informations est née avec l'écriture et l'idée d'utiliser les calculateurs pour trouver plus facilement un texte, un document, un ouvrage, est apparue avec les premiers ordinateurs. Les premières applications ont consisté à construire un référentiel par exemple sous la forme d'un thésaurus basé sur la classification proposée par Melvil Dewey en 1876 (Dewey, 1876), à décrire chaque document ou ouvrage à l'aide de descripteurs ou mots-clés en s'appuyant sur cette classification (ce qui présentait l'avantage de limiter les problèmes d'orthographie ou de dénomination), puis à effectuer à partir d'une requête exprimée sous forme booléenne (termes séparés par une combinaison d'opérateurs logiques de type : et, ou) un appariement automatique produisant in fine une liste d'objets, de documents.

Les techniques ont évolué. La classification décimale DEWEY reflet d'une vision du monde propre aux Etats-Unis d'Amérique à la fin du 19ème siècle a été progressivement complétée / enrichie par la classification décimale universelle (CDU) créée sous l'impulsion de Paul Otlet et Henri La Fontaine (Otlet et La Fontaine, 2001). La description des documents a été élargie à de nouveaux éléments comme le titre, le créateur, le sujet, l'éditeur, le contributeur, la date, etc. (ensemble de métadonnées appelé aussi Dublin Core : norme internationale ISO 15836). Dans un certain nombre de cas, le document fait aussi l'objet d'une véritable indexation par analyse automatisée du texte intégral. Enfin, les requêtes ont considérablement évoluées avec l'apparition de questions en langage naturel et d'ontologies « spécification explicite (formelle) d'une conceptualisation (partagée) » (Gruber, 1993) qui peuvent servir de guide sémantique à l'expression de la requête et de support actionnable pour la recherche d'informations (Ranwez, 2010).

Le développement de langages structurés, et notamment du SGML (Standard Generalized Markup Language) pour le traitement de documentation technique dans l'industrie, a permis de rapprocher les intentions de lecture et les intentions opératoires. Le SGML a été ensuite assez rapidement remplacé par un langage associant expressivité et complexité : XML (eXtensible Markup Langage). Le balisage proposé par XML est porteur de sens, il est centré sur les contenus et participe à l'émergence au début des années 2000 du « Web Sémantique ».

Le Web Sémantique est une extension du Web qui a pour objet de faciliter la coopération homme-machine : « The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation » (Berners-Lee, 2001).
Le Web Sémantique repose sur un travail important de l'homme sur les documents à paraître. Il ne s'agit pas simplement de remplacer les balises HTML par des balises XML, mais d'organiser les mots-clés ou concepts constitutifs du document. Ceci doit permettre à terme de faciliter l'indexation de l'ensemble des ressources du Web pour une exploitation optimale.
La structuration des réseaux numériques, auxquels l'édition scientifique n'échappe pas, va tenter de progresser dans le défi d'une sémantisation des données mises à disposition. Cela va donner lieu à la création de normes et standards visant à harmoniser les données. Ainsi, des modèles sont institués comme garantie d'interopérabilité définissant « le bon niveau » de structuration des données. Les normes LOM (Learning Object Metadata) ou SCORM (Sharable Content Object Reference Model) apparaissent comme des spécifications permettant de créer des données structurées dans une granularité de contenus de plus en plus affinée.

Nous nous intéressons à des interfaces d'édition numérique en tant qu'organisateur de données. Les modèles d'organisation des données qui prévalent apparaissent plus ou moins fermés, autour de « grains de connaissances » identifiés par leur unité sémantique et leur portée didactique, ou au contraire ouverts sur des données mises à disposition dans une dynamique évolutive.

L'édition numérique bouleverse les codes de la publication, notamment scientifique, comme en atteste la multiplicité des initiatives de création d'espaces documentaires ou de bases de données spécialisées. La richesse de ces initiatives est sans nul doute un moyen de pallier à la crise de l'édition papier et une solution pour favoriser l'accès à l'information scientifique. L'édition numérique propose une alternative aux modes traditionnels de publication, faisant valoir les droits des auteurs. Le consortium ERUDIT (composé de l'Université de Montréal, de l'Université Laval et de l'Université du Québec à Montréal) propose un modèle novateur de promotion et de diffusion des résultats de la recherche (http://www.erudit.org/). Dans le même souci de valorisation de la recherche scientifique, le CNRS créé en 2000 le Centre de Communication Scientifique Directe (CCSD). Cet espace documentaire est dédié à la réalisation d'archives ouvertes et regroupe trois services : le Hyper Article en Ligne (HAL), Cours en Ligne (CEL), Thèse en Ligne (TEL). Que ce soit sous forme de portails ou d'espaces documentaires, l'édition numérique organise la communication d'informations scientifiques en structurant élaboration, diffusion et modalités de réception.

Nous nous intéressons à ces interfaces au sens où les définit Thierry Bardini comme un « entre deux » (Proulx et Bardini, 1998). A l'origine, la notion d'interface vient de la physique. Dans son acception actuelle, elle est indissociable de l'informatique et étroitement liée à la notion d'interactivité. Une interface est un « point de contact » entre l'objet et le « monde » dans lequel il se situe. Les « objets » peuvent être des systèmes informatiques et / ou des êtres humains. L'interface joue alors un rôle prépondérant dans l'échange et la communication. Elle prend une forme opératoire le plus souvent réfléchie dans le respect des principes d'ergonomie. Mais au-delà l'interface catalyse les intentions dans l'objet. L'approche interprétative développée par Umberto Eco dans une analyse sémio pragmatique rend compte de trois types d'intentions : l'intention lectoris, l'intention autoris et l'intention operis. La mise en cohérence des ces trois intentions facilite la construction du sens en contexte. L'édition numérique dans son projet « d'œuvre ouverte » se situe à l'interface des ces trois

intentions inscrites dans le processus de sémantisation. Nous pouvons légitimement nous poser la question de la détermination de ces interfaces de mise à disposition de données scientifiques sur l'intention de l'auteur et sur l'intention opératoire.

Problématique et méthode

Considérant que le projet d'apprentissage et / ou de veille scientifique est entendu comme un processus de construction de connaissances actionnables et que l'édition numérique est vue comme un stock de données indexées et référencées dans une logique plus ou moins ouverte de mise à disposition, nous pouvons poser une base de questionnement problématique. Qu'elle soit à vocation de service public ou à vocation commerciale, l'édition scientifique numérique organise l'accès à des stocks de données. Ces « réservoirs » rendent disponibles des données indexées, classées et référencées. La profusion d'informations scientifiques éditorialisées tient-elle ses promesses ? Ces données disponibles rencontrent-elles les projets des acteurs ? Les données rendues disponibles par l'édition numérique garantissent-elles toujours un critère de pertinence pour des connaissances actionnables au niveau du projet d'un acteur ou d'un collectif ?

Les réservoirs de données sont riches mais leur utilité et utilisabilité (Chevalier et Tricot, 2007) restent très faibles du point de vue de la construction de connaissances actionnables. Dans un grand nombre de cas, l'accès aux données est réalisé par des requêtes basées sur des descripteurs ou à travers une taxinomie aristotélicienne. Cette classification structure des données potentiellement intéressantes mais s'avère peu opérante du point de vue de la construction de connaissances actionnables.

A partir de ces constats, nous émettons une hypothèse concernant le potentiel des ontologies. La structuration et l'organisation des informations sous formes de documents numériques, de réseaux sémantiques et d'ontologies permettraient d'utiliser de manière profitable les données disponibles. Les données présentent un caractère informe, un potentiel plus ou moins activé, a contrario du terme d'information, du latin « informare » ou « informatio », qui signifie « donner une forme, une signification ». La donnée deviendrait donc information dans sa rencontre avec une intention (Schutz, 1987). La première intention à laquelle l'information va être soumise est celle du concepteur de l'ontologie. Comment les ontologies et leurs différentes générations induisent-elles la structuration de données en informations ? Quelles sont les médiations sociocognitives, relationnelles et technologiques observables ? Les ontologies en tant qu'agencements numériques se situent à l'intermédiaire entre des données préexistantes disponibles, et une intention d'apprentissage et / ou de veille scientifique. Nous verrons à ce propos quels sont les niveaux de performances des différentes générations d'ontologies (ontologies élaborées de manière manuelle ou ontologies quasi automatique) dans l'organisation et la structuration des données en informations.

C'est ainsi que nous pouvons définir la problématique centrale de notre contribution : **en quoi et comment les ontologies peuvent-elles constituer des dispositifs de médiation plus ou moins performants entre des données préexistantes et un projet d'apprentissage et / ou de veille scientifique ?**

Pour répondre à cette problématique nous rendrons compte de deux cas ; l'expérience de construction d'une ontologie de la pneumologie initiée par l'INSERM et l'utilisation d'une plateforme ToxNuc mise en place dans le cadre du programme Toxicologie Nucléaire Environnementale (Ménager, 2004). A travers ces projets de recherche nous tenterons

d'appréhender le développement d'ontologies et la position des acteurs dans leurs processus (individuels et collectifs) d'apprentissage et de veille scientifique. La posture méthodologique qui est la notre dans le cadre de ces recherches prend ses fondements dans l'ethnographie des usages. Selon Serge Proulx, l'ethnographie des usages permet « d'observer le plus finement possible l'action effective de la technique dans la société » (Proulx, 1999). Concrètement nous privilégions une approche par immersion, plus ou moins approfondie sur chacun des deux cas explicités ici, et une analyse des formes signifiantes par la mise en contexte. L'approche par immersion est organisée autour de techniques de recueil privilégiant l'observation participante, des entretiens semi-directifs, l'analyse des documents recueillis et l'analyse des documents exploités. Nous retrouvons les formes identifiées par l'anthropologie des communications : « L'anthropologie utilise généralement quatre formes de production de données : l'observation participante (l'insertion prolongée de l'enquêteur dans le milieu de vie des enquêtés), l'entretien (les interactions discursives délibérément suscitées par le chercheur d'une manière plus ou moins directive), les procédés de recension (le recours à des dispositifs d'investigation systématique), la collecte de sources écrites » (Héas et Poutrain, 2003).

Nous tenons à préciser que notre approche se réclame du paradigme interprétatif étant donné que notre objet d'étude s'inscrit à la croisée des technologies et des usages en situation. Notre démarche d'analyse est guidée par la mise en contexte des phénomènes signifiants laissant apparaître des médiations complexes au sein des dispositifs.

Étude de cas

Nous nous sommes appuyés sur la méthode des études de cas multiples (Yin, 1984) dans une approche qualitative et en compréhension (observation, échange de messages, de documents, entretiens téléphoniques et / ou de vive voix). Les informations collectées ont fait l'objet d'une triangulation (Zamanou et Glaser, 1989) et les résultats soumis à une assez large palette d'acteurs impliqués dans l'action pour en corriger les erreurs et enrichir les aspects négligés (Huberman et Miles, 1991).

1 - Construction d'une ontologie de la pneumologie

Nous considérerons que les ontologies sont des représentations de modèles de connaissances situées à mi-chemin entre information disponible et intention d'apprentissage. Pour mieux appréhender la notion d'ontologie nous allons présenter les travaux de doctorat réalisés par Audrey Baneyx sous la direction de Jean Charlet dans le laboratoire Santé Publique et Informatique Médicale de l'INSERM au Centre de recherche des Cordeliers (Baneyx, 2007).
Depuis le 31 juillet 1991 les établissements de santé doivent procéder à l'évaluation et l'analyse de leur activité en s'appuyant sur le Programme de Médicalisation des Systèmes d'Information (PMSI). L'action poursuit un double objectif une meilleure répartition des ressources et une amélioration à terme de l'efficacité des pratiques. Dans ce contexte, il est nécessaire d'harmoniser assez rapidement les langages médicaux, pour aider les médecins qui renseignent le dispositif à partir des comptes rendus d'hospitalisation sans avoir nécessairement vu les patients, comme pour favoriser l'échange entre partenaires hospitaliers : médecins, soignants et administratifs.
Dans ce contexte, l'effort d'Audrey Baneyx a porté plus particulièrement sur la construction d'une ontologie de la pneumologie. Après une immersion dans la spécialité qui lui a permis d'appréhender les besoins des praticiens et réaliser un état de l'art, la jeune femme - en s'appuyant sur la méthode ARCHONTE proposée par Bruno Bachimont (Bachimont, 2002) -

a mis au point une méthodologie complète pour élaborer une ontologie en partant des documents disponibles.

Sur le terrain elle a rencontré de nombreuses difficultés. Les projets de standardisation du contenu des dossiers médicaux et de codage des pathologies. La complexité de la démarche de soin. La difficulté pour obtenir un consensus chez les médecins codeurs. La difficulté méthodologique pour obtenir un recueil de données homogène pour un même état pathologique. La nécessité de respecter la structure naturelle de l'information pour la dénaturer le moins possible. La faiblesse des outils d'aide au codage fondés sur les thésaurus professionnels. Les difficultés pour rendre compte des richesses d'un langage de spécialité. Enfin, la crainte que le codage des actes et des pathologies soit plus utilisé pour surveiller l'activité que pour améliorer les pratiques. S'appuyant sur les travaux de T. Gruber « une ontologie est une spécification partagée d'une conceptualisation » (Gruber, 1993) puis de B. Bachimont « définir une ontologie pour la représentation des connaissances, c'est définir un domaine et un problème donnés, la signature fonctionnelle et relationnelle d'un langage formel de représentation et la sémantique associée » (Bachimont, 2002), l'état de l'art dresse un panorama : des différents types d'ontologies selon le degré de formalisme (hautement informelles, semi-informelles, semi-formelles et formelles), les objets modélisés (pour la représentation des connaissances, de domaine, de haut niveau, génériques, de tâches, d'applications) ou le niveau de granularité (fine, large), et des formalismes pour la représentation des connaissances (graphes conceptuels et logiques de description). La jeune femme présente ensuite un assez grand nombre de méthodologies de construction d'ontologies et des outils pour finalement privilégier la méthode ARCHONTE (choisir les termes pertinents et les positionner, formaliser les connaissances en précisant les propriétés, et opérationnaliser l'ensemble dans un langage de représentation des connaissances) et l'outil PROTEGE (développé et mis à disposition par le Stanford Medical Informatics).

Dans une démarche ascendante, le travail de construction de l'ontologie de la pneumologie est alors conduit en cinq étapes : analyse terminologique fondée sur l'analyse des corpus (en utilisant des outils pour repérer les syntagmes nominaux), identification sélection et extraction des termes pertinents, normalisation, formalisation et opérationnalisation. La construction est conduite avec un souci de validation des contenus en s'appuyant sur un ensemble de critères : clarté et objectivité, perfection, cohérence et extensibilité, biais d'encodage minimal, engagements ontologiques minimaux, distinction ontologique, minimisation de la distance sémantique et normalisation des termes. Un ensemble d'erreurs potentielles est aussi recherché : circularité, partition, redondance, sémantique, incomplétude. L'ensemble devant être naturellement validé par les praticiens.

2 - Construction de réseaux sémantiques étendus pour la toxicologie nucléaire

Les ontologies construites dans une démarche ascendante semblent être performantes mais aussi coûteuses à élaborer et entretenir. Or, si la qualification d'actes médicaux requiert un niveau élevé de précision, d'autres usages pourraient se contenter de modèles intermédiaires de connaissances, moins aboutis, mais beaucoup plus économiques à construire et faire évoluer. Pour appréhender d'autres formes ou démarches de construction de modèle de connaissances, nous allons présenter les travaux de doctorat de Reena T.N. Shetty réalisés sous la direction de Joël Quinqueton et Pierre-Michel Riccio au Centre de Recherche LGI2P de l'Ecole des Mines d'Alès (Shetty, 2008).

Le programme de recherche Toxicologie Nucléaire Environnementale a pour objectif d'identifier les effets toxiques d'éléments chimiques, radioactifs ou non, utilisés dans la recherche et l'industrie nucléaires. Ces travaux visent à déterminer les mécanismes de toxicité de ces éléments pour l'homme et son environnement et de proposer des procédés de dépollution et de traitement d'éventuelles contaminations. Ce programme inter-organismes, organisé en 15 projets de recherche, est piloté par le CEA et soutenu par le Ministère de la Recherche. Il fédère, depuis 2004, une communauté scientifique pluridisciplinaire d'environ 200 chercheurs/an et 100 doctorants et post-doctorants (CEA, CNRS, Inra, Inserm).
Dans ce contexte l'action du centre de recherche LGI2P de l'Ecole des Mines d'Alès a été d'accompagner le développement de la communauté scientifique notamment à travers la mise en place d'un système d'information collectif. Les travaux de Reena Shetty ont concerné plus particulièrement l'élaboration d'une méthodologie de construction de modèles de connaissances pour faciliter l'identification de documents et le travail collaboratif (Shetty, 2007).

La recherche d'informations sur les systèmes ouverts (internet) ou fermés (plateformes de travail collaboratif) devient vite exaspérante car les données accessibles sont difficilement exploitables. Les requêtes simples produisent un très grand nombre de résultats, et les utilisateurs passent beaucoup de temps à analyser ces résultats souvent peu pertinents par rapport à leur projet, leur besoin. A partir de ce constat l'idée est de mieux utiliser les capacités de calcul de la machine pour retreindre la taille de l'ensemble des résultats ou identifier avec une bonne précision l'information recherchée. Les travaux sur le « web sémantique » et les « ontologies » sont prometteurs, car ils permettent une meilleure « compréhension » des documents et facilitent à l'aide d'outils appropriés la qualité des recherches. Mais la construction d'ontologies et un travail long et très coûteux.

Aussi, l'idée de ce travail de recherche a été d'imaginer une approche semi-automatique de construction de modèles de connaissances pour la recherche d'information dans de larges corpus. Le principe : trouver une forme d'ontologie utilisable avec les même outils, mais qui puisse être construite rapidement et mise à jour facilement, en faisant intervenir les spécialistes du domaine sur une partie strictement indispensable et laissant la machine calculer le reste. Nous l'avons appelée réseaux sémantiques étendus ou ESN pour : Extended Semantic Network.

La première étape consiste à partir d'un ensemble de documents soigneusement sélectionnés à créer par calculs mathématiques issus de l'analyse de données et de la classification automatique (fondés sur la proximité des mots dans l'espace) un ou plusieurs réseaux de proximités (un réseau de proximité par thème). Dans le même temps des spécialistes élaborent, en s'appuyant sur un guide de construction, de petits réseaux sémantiques (environ cinquante nœuds ou concepts par réseau) qui représentent la vision du monde de l'acteur en situation sur un thème.

La deuxième étape consiste à fusionner de façon automatique réseaux sémantiques et réseaux de proximité pour générer des quasi-ontologies qui pourront être ensuite utilisées à la place de véritables ontologies pour restreindre de façon efficace l'ensemble des résultats d'une requête. Pour certains spécialistes la démarche est discutable puisqu'elle consiste à rapprocher des concepts et des termes, mais elle respecte le point de vue de l'acteur qui construit de façon manuelle le ou les réseaux sémantiques, les termes n'étant finalement utilisés que pour étendre et préciser les différents concepts. Les premiers résultats sur le programme ToxNuc-E

sont encourageants, mais il reste à conduire une véritable campagne de validation des travaux dans différentes situations.

Pour un modèle à trois niveaux

L'étude de cas montre qu'il est possible de construire rapidement des modèles de données ontologiques permettant pour un usage précis (la recherche de documents dans des corpus scientifiques) d'obtenir des résultats d'un bon niveau de qualité tout en conciliant primat de l'intervention humaine (expert en situation) et calculs automatiques (sans hésiter à repousser les blocages habituels).

En 2004, Jean-Michel Salaün et Jean Charlet ont dressé - dans l'introduction d'un numéro thématique de la revue I3 consacré au document numérique (Salaün et al., 2004) - un état des lieux du travail collectif de réflexion pluridisciplinaire sur la notion de document. Leur hypothèse était que le concept de document omniprésent dans notre quotidien pourrait être plus approprié pour comprendre et décrire les situations que celui d'information. S'appuyant sur un travail de rédaction collective d'un document signé du pseudonyme Roger T. Pédauque (Pédauque, 2003), ils ont identifié : les avancées, des interrogations, et un approfondissement nécessaire pour poursuivre le processus. Le point de vue de vue de ces chercheurs, et à travers eux celui des membres du réseau I3, nous semble constituer un bon point de départ pour appréhender l'organisation et l'accès à l'information scientifique éditorialisée et au-delà les nouveaux agencements numériques.

Jean-Michel Salaün et Jean Charlet ont regroupé les travaux par rapport à leur angle principal d'approche : signe ou forme, texte et contenu, medium ou relation. Dans le signe ou la forme, leur point de vue est que le numérique a déplacé la question du support, qui en assurait la stabilité en fixant l'inscription, vers la problématique de la structure, les travaux convergeant autour de la compréhension des structures logiques du document - facilitée par la popularisation du langage XML - et dans les interrogations sur les formes perceptibles. Dans le texte et le contenu ils ont identifié une problématique principale : comment construire des modèles pour traiter les contenus afin de les réorganiser pour produire de nouveaux documents ? Basés en grande partie sur les métadonnées ces travaux ont pour ambition de produire de nouveaux documents adaptés à la demande du lecteur. Enfin, dans le medium ils ont identifié deux terrains : la communication organisationnelle et les medias de masse. Le document enrichit un patrimoine de connaissances partagées dans une dynamique retreinte à des groupes de travail ou dans une démarche de diffusion élargie.

Pour ces chercheurs, le document décomposé de facto dans un substrat intermédiaire – bases de données centralisées ou réseaux hypertextuels – n'a de véritable existence qu'à deux moments : lors de sa création par l'auteur, et lors de sa reconstruction par un lecteur. Ceci pose des questions de validation, de hiérarchisation et de responsabilité éditoriale. D'autres chercheurs abondent dans cette direction. Pour Sylvie Lainé-Cruzel, le document est un objet signifiant et fini, le terme de ressource devant être privilégié pour un document électronique immergé dans le substrat. Pour Sylvie Leleu-Merviel le document est avant tout une image qui fait sens et peut être décomposée en différentes couches techniques ou sémiotiques. Enfin des chercheurs comme Marie-Anne Chabin, Dominique Cotte, Marie Després-Lonnet, Dominique Boullier ou Franck Gitella considèrent le document comme le résultat d'un processus et mettent en exergue le travail de construction éditoriale qui pour Sandra Bringay, Catherine Barry et Jean Charlet devient une construction collective. Appréhender l'organisation et l'accès à l'information numérique scientifique pose de notre point de vue

trois questions importantes : celle de la tendance ou nécessité à séparer fonds et forme, celle des acteurs, auteurs ou lecteurs mais aussi gestionnaires de contenus ou formateurs, dotés d'une intentionnalité, d'une vision du monde en situation (Schutz, 1987), et enfin celle de la qualité des documents à travers la validation, la hiérarchisation et la responsabilité éditoriale.

Un modèle à trois niveaux

Pour appréhender l'information scientifique éditorialisée nous proposons un modèle général en trois niveaux : une couche basse ou substrat dans laquelle serait stocké l'ensemble des données de base accessibles constituant un réservoir d'informations, une couche haute dédiée aux intentions dans l'usage (concepteurs, apprenants...) et une couche intermédiaire dans laquelle serait concentré l'ensemble des supports de médiation (index, thésaurus, ontologies, réseaux sémantiques,...).

Le modèle en trois niveaux est bien le reflet d'une décomposition de la connaissance en construction : données, informations et connaissances actionnables. Mais, il convient à notre avis de prendre un peu de recul quand à la composition des différentes couches.

Pour la couche basse ou substrat, que l'on se situe dans un système ouvert (web étendu), semi-ouvert (plateformes professionnelles), ou fermé (plateformes de travail collaboratif), l'expérience montre qu'il existe un processus de sélection des données. Les législations internationales, les pratiques culturelles, l'éthique guident de facto le processus de publication des données. Pour ne pas être poursuivi en justice (par exemple pour diffamation ou plagiat), parce que le dépôt n'est pas complètement anonyme (ce qui pose la question de la responsabilité des acteurs et de leur image par rapport aux autres acteurs), ou plus simplement parce que les contributeurs semblent être dans d'assez larges proportions dotés d'une volonté de bien faire, les données accessibles sont plutôt d'assez bonne qualité. Un point important est que les données n'existent pas seules : pour qu'elles puissent être repérées, le contributeur va associer des mots-clés, les moteurs de recherche vont créer et utiliser des index (différents selon le type de moteur). L'ensemble fait qu'il est aujourd'hui difficile de considérer que l'information numérique accessible est une accumulation de données sans forme et sans cohérence au sens intrinsèque. Il existe toutefois de nombreux problèmes : coexistence de versions différentes d'un même ensemble de données, repérage des auteurs et / ou contributeurs, positionnement dans le temps, insertion dans un champ, cohérences locales, qui font que les données publiées peuvent être complètement inexploitables dans un projet d'apprentissage ou de veille scientifique.

La couche haute doit de notre point de vue faciliter l'expression des intentions de l'auteur en charge d'une action de publication comme celles de l'apprenant ou d'un acteur inscrit dans un projet de veille. Méthodes et outils devraient faciliter dans ce cas le travail de publication de l'auteur comme le parcours in situ de l'apprenant. Il nous semble difficile, et dans un grand nombre de cas impossible, de dissocier la contribution directe de ces acteurs (qu'il s'agisse d'un scénario d'apprentissage rédigé par un formateur ou de la découverte d'un parcours d'apprentissage ou de veille) d'éléments qui vont apporter une valeur ajoutée. De nombreux outils technologiques peuvent être utilisés : modèles utilisateurs mémorisant les pratiques (sur la machine de l'utilisateur ou sur une machine distante) ou traces d'usages pour in fine donner plus de souplesse ou trouver une meilleure adéquation dans la démarche (adaptation de la restitution aux profils des utilisateurs pour améliorer la construction des connaissances). Ces outils peuvent aussi être mis à contribution pour mieux situer les éléments

d'apprentissage ou de veille en contexte (par exemple en accédant à la biographie des auteurs de référence).

En ce qui concerne la couche intermédiaire de nombreux travaux ont été initiés, comme ceux que nous avons présenté pour la construction d'une ontologie de la pneumologie ou la mise au point de réseaux sémantiques étendus pour la toxicologie nucléaire environnementale. Aujourd'hui, deux points sont à approfondir : en termes d'accés la mise en cohérence d'un projet éditorial ou d'apprentissage par rapport aux données accessibles via les représentations collectives (élaborées sous forme d'ontologies ou de réseaux sémantiques étendus), et en termes d'organisation la mise en cohérence, la validité, l'homogénéité, la constance et la fiabilité des informations (qui vont qualifier les données accessibles dans la couche basse).

Conclusion

La réponse à la problématique que nous posions sur les niveaux de performativité des différentes générations d'ontologies doit être nuancée. Nous constatons tout d'abord que les ontologies sont relatives à la responsabilité éditoriale qui prévaut. Cette dernière reposerait sur un niveau d'expertise suffisant pour légitimer les savoirs organisés. Nous interrogeons la place des experts dans leurs intentions éditoriales. Les objectifs et les finalités qui prévalent à la constitution d'ontologies déterminent l'organisation de « monde de connaissances ». La responsabilité éditoriale est donc primordiale et doit être interrogée dans ses intentions. Le projet éditorial s'inscrit dans une intention qui au mieux recouvre une dimension heuristique mais qui peut aussi être marquée par des intentions politiques ou économiques. La construction d'ontologies ne peut pas faire l'impasse d'une justification des schèmes épistémologiques qui prévalent à son existence. Ce qui nous invite à revenir à la question initiale de l'explicitation des paradigmes épistémologiques de référence : ceux par lesquels nous tenons pour « valables » les connaissances constituées au sein d'ontologies. Nous prônons une responsabilité éditoriale qui non seulement organise les connaissances mais surtout rend lisible l'épistémè et les « schèmes de liaison » à l'œuvre dans la construction de « mondes de connaissances ». Nous rejoignons les préoccupations éthiques d'une « science avec conscience » et préconisons une mise en contexte de la construction d'ontologies.

Ces dernières apparaissent par ailleurs comme des spécifications partagées par une communauté de pratiques. En ce sens, l'ontologie n'est valable que pour une communauté et reproduit des formes paradigmatiques reconnues au sein de cette communauté. En ce sens, les ontologies apparaissent comme des systèmes de reproduction des connaissances constitutives de paradigmes de référence. Les ontologies et les nouveaux agencements numériques qui leur donnent formes, s'ils facilitent l'organisation et l'accès à l'information scientifique éditorialisée, ne favorisent pas l'ouverture créative à travers des apprentissages à la marge des disciplines constituées ou des paradigmes de référence. La médiation n'étant pas neutre elle oriente forcément l'accès aux informations dans une vision paradigmatique. De manière plus conséquente et comme nous avons pu l'observer dans le cas de construction d'une ontologie de la pneumologie, les ontologies orientent parfois la pratique. La dimension pragmatique instaurée par l'utilisation des ontologies dans les activités scientifiques (sous formes d'évaluation notamment) n'échappe pas à une orientation paradigmatique. Un des paradoxes dans la construction d'ontologies repose sur les récurrences de la démarche éditoriale au détriment de la pertinence d'un monde de connaissances ouvert sur la pratique.

Nous préconisons donc une ingénierie de la médiation dans les dispositifs d'information scientifique éditorialisée. Cette ingénierie de la médiation va garantir la qualité des documents

numériques à travers la validation, la hiérarchisation et la responsabilité éditoriale et, située dans un dialogue entre pratiques et épistémè, va renforcer le lien entre les acteurs. Une ingénierie de la médiation compléterait l'apport de l'ingénierie des connaissances, dont la principale contribution est d'élaborer des représentations collectives, en resituant les projets de conception ou d'usage au centre de la démarche.

Bibliographie

Akrich M., The De-Scription of Technical Objects, Shaping Technology / Building Society: Studies in Sociotechnical change, MIT Press., Cambridge, 1992.

Bachimont B., Isaac A. and Troncy R., Semantic commitment for designing ontologies: A proposal, 13th International Conference on Knowledge Engineering and Knowledge Management EKAW'2002, in Lecture Note in Computer Science, Springer, vol. 2473, 2002, p. 114-121.

Baneyx A., Construire une ontologie de la pneumologie, Thèse de doctorat en informatique, Université Paris 6, 2007.

Berger P. et Luckmann Th., La construction sociale de la réalité, Paris, Méridiens-Klincksieck 1986.

Berners-Lee T., Hendler J., Lassila O., The Semantic Web, Scientific American, May 2001.

Callon M., Four Models for the Dynamics of Science, Handbook of Science and Technology Studies, Thousand Oaks, Sage, 1995, p. 29-63.

Chevalier A. et Tricot A., Ergonomie des documents électroniques, Paris, Presses Universitaires de France, 2007.

Dewey M., A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library, Gutenberg Project, 1876, http://www.gutenberg.org/etext/12513

Gruber T., « A translation approach to portable ontology specifications », in Knowledge Acquisition, vol. 5, n° 2, 1993, p. 199-220.

Héas S. et Poutrin V., Les méthodes d'enquête qualitative sur Internet, ethnographiques.org, vol. 4, novembre 2003.

Huberman A.M. et Miles M.B., Analyse des données qualitatives, Edition du renouveau pédagogique (traduction Catherine De Backer et Vivian Lamongie), Bruxelles, Belgique, 1991.

Latour B., Petites leçons de sociologie des sciences, Paris, Seuil, 1996.

Le Moigne J.L., Quand savoir devient comprendre, Ingénierie des pratiques collectives - La Cordée et le Quatuor, L'Harmattan, 2000.

Ménager M.T., Programme Toxicologie Nucléaire Environnementale : comme fédérer et créer une communauté scientifique autour d'un enjeu de société, Colloque Intelligence Collective : partage et redistribution des savoirs, Site de Nîmes de l'Ecole des Mines d'Alès, 29-30 septembre 2004.

Otlet P. et La Fontaine H., Classification décimale universelle : édition abrégée (7ème éd), Editions du Céfal, Liège, 2001, 292 p.

Pédauque R.T., Document : forme, signe et médium, les re-formulations du numérique, working paper, version 3 du 08 juillet 2003.

Proulx S. et Bardini T., Entre publics et usagers: la construction sociale d'un nouveau sujet communicant, Médiations sociales, systèmes d'information et réseaux de communication, Actes du Congrès de la Société française des sciences de l'information et de la communication, Metz, 1998, p. 267-274.

Proulx S., La construction sociale des objets informationnels : matériaux pour une ethnographie des usages, Actes du Colloque Comprendre les usages d'Internet, Paris, 3 et 4 décembre 1999.

Ranwez S., Ranwez V., Sy M.F., Montmain J. et Crampes M., Utilisation de proximités sémantiques pour améliorer la recherche et le rendu d'information, Actes des 21èmes Journées Francophones d'Ingénierie des Connaissances, Nîmes, Presses des Mines, Collection Mathématiques et Informatique, 9-11 juin 2010, p. 247-258.

Salaün J.M. et Charlet J., Introduction : un dialogue pluridisciplinaire pour penser le « document numérique », Information-Interaction-Intelligence, vol. 4, n° 1, 2004, p. 7-17.

Schutz A., Le chercheur et le quotidien : phénoménologie des sciences sociales, Paris, Méridiens Klincksieck, 1987.

Shetty R.T.N., Quinqueton J., Riccio P.M., Penalva J.M. and Villerd J., Collaborative Platform Using Knowledge Cartography – ToxNuc-E, Proceedings of the 2007 International Symposium on Collaborative Technologies and Systems CTS'2007, Orlando, Florida, USA, 21-25 May 2007, New-York, IEEE, p. 312-320.

Shetty R.T.N., Enrichissement de réseaux sémantiques par la proximité de concepts, Thèse de doctorat, Ecole des Mines de Paris, 2008.

Simon H.A., Models of man, John Wiley & Sons Inc., 1957.

Yin R.K., Case Study Research, Design and Methods, Sage, London, 1984.

Zamanou S. and Glaser S.R., Communication intervention in an organization: measuring the results through a triangulation approach, Annual Meeting of the Speech Communication Association, San Francisco, November 1989, 37 p.