UFR-SHA
- - -
Mention Information-Communication
Spécialité Documentation

Année universitaire 2017-2018

Construction d'ontologies pour aider à la recherche d'information : entre formalisme, pratiques professionnelles et prise en compte des usages

Mémoire pour l'obtention du Master esDOC

Présenté par Audrey Gris

Le 19 septembre 2018

Sous la direction de Monsieur David Guillemin

Université de Poitiers

UFR-SHA
- - -
Mention Information-Communication
Spécialité Documentation

Année universitaire 2017-2018

Construction d'ontologies pour aider à la recherche d'information : entre formalisme, pratiques professionnelles et prise en compte des usages

Mémoire pour l'obtention du Master esDOC

Présenté par Audrey Gris

Le 19 septembre 2018

Sous la direction de Monsieur David Guillemin

Université de Poitiers

« Les chercheurs veulent un monde parfait, bien modélisé,
et les industriels veulent juste quelque chose qui marche. »

Florence Amardeilh

Remerciements

Je tiens tout d'abord à remercier mon directeur de mémoire, David Guillemin. Grâce à sa réactivité, ses recentrages, ses conseils avisés et ses corrections, j'ai rédigé ce mémoire plus sereinement.

J'exprime ma sincère reconnaissance à Marie-Paule Cochet, Nadia Fafi et Henri-Maxime Suchier, mes trois tuteurs de stage au sein du service Banque de Contenus à Ouest-France. J'ai vraiment apprécié leur accueil, leur gentillesse, et le travail que nous avons pu effectuer en collaboration. Je les remercie aussi d'avoir accepté de m'accorder un entretien dans le cadre de ce mémoire. Mention spéciale à Nadia qui a pris le temps de répondre à mes questions avant le stage, m'a accueillie au sein du cercle des documentalistes et m'a encouragée lors de l'élaboration de ce document.

J'adresse aussi mes remerciements à Florence Amardeilh, Jean Charlet et Michel Chein, que j'ai également interviewés. Ils m'ont consacré une partie de leur temps et m'ont suggéré quelques lectures ; je les en remercie chaleureusement. Je tiens aussi à remercier Thomas Francart qui m'a donné de précieuses pistes et a accepté de me mettre en relation avec Florence Amardeilh.

Je n'oublie pas mes camarades de promotion, notamment Tanguy Germain et Noémie de la Cotte, avec qui j'ai aimé travailler et partager des cookies lors du dernier projet du M2. Un grand merci à Timothée Béguier pour ses conversations passionnantes sur le cinéma et pour tous les films que j'ai pu découvrir grâce à lui.

Mille mercis à ma famille et à mes amis qui me soutiennent. Merci aux personnes (aussi bien « les anciens » que « les nouveaux de Rennes ») avec qui j'ai passé de supers moments cet été : Amandine, Aurélie, Élodie, Fabrice, Fanny, Kathleen, Lauren, Léna, Lucas, Mamie Josette, Marinette, Marylou, Nathalie, Nicolas, Ophélie, Perrine, Polyn, Thaïs...Des remerciements tout particuliers à Nathalie Gris qui a accepté de relire patiemment ce mémoire (et qui a également la gentillesse de relire mes rapports universitaires depuis cinq ans déjà!)

Table des abréviations

ABES	Agence Bibliographique de l'Enseignement Supérieur
ADBS	Association des professionnels de l'information et de la documentation
AFNOR	Association française de normalisation
AFP	Agence France-Presse
BDC	Banque de Contenus (service de Ouest-France) aussi appelé Service Informatique Banque de Contenus (SIB)
CIDOC-CRM	*International Committee for Documentation Conceptual Reference Model*
ECM	*Enterprise Content Management*
FRBR	*Functional Requirements for Bibliographic Records*
GED	Gestion Électronique des Document
HTML	*HyperText Markup Language*
HTTP	*HyperText Transfer Protocol*
IC	Ingénierie des Connaissances
IA	Intelligence Artificielle
INRIA	Institut national de recherche en informatique et en automatique
IRISA	Institut de Recherche en Informatique et Systèmes Aléatoires
LERUDI	LEcture Rapide en Urgence du Dossier Informatisé du patient (projet)
LIRMM	Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
OFS	Open Food System (projet)
OWL	*Web Ontology Language*
RDF	*Ressource Description Framework*
RDFS	*RDF Schema*

RIF	*Rule Interchange Format*
RTO	Ressource Termino-Ontologique
SBC	Système à Base de Connaissances
SHACL	*Shapes Constraint Language*
SI	Système d'Information
SOC	Système d'Organisation des Connaissances
SparQL	*Protocol and RDF Query Language*
SRI	Système de Recherche d'Information
TAL	Traitement Automatique du Langage
URI	*Universal Resource Identifier*
URL	*Uniform Resource Locator*
W3C	World Wide Web Consortium
XML	*Extensible Markup Language*

Sommaire

Introduction

La bonne circulation de l'information dans une entreprise tient aujourd'hui un rôle primordial. Il est important que tous les employés aient le même niveau d'information et puissent disposer facilement des documents dont ils ont besoin dans l'exercice de leurs tâches. Par ailleurs, les décisions au sein d'une organisation ne peuvent être prises qu'après un important travail de synthèse et de communication de l'information. Effectivement, avant de faire un choix, des informations relatives à ces décisions, au contexte professionnel et aux conséquences que celles-ci peuvent avoir, doivent être ordonnées, classées et communiquées. Or, avant de traiter ces informations, il faut les chercher, ce qui peut constituer une opération plus ou moins longue et fastidieuse, en fonction de la gestion de l'information mise en place.

En outre, l'innovation est directement liée au traitement, à la circulation, au partage, à l'exploitation et à l'enrichissement des connaissances : pour faire de la veille technologique et concurrentielle, prendre des décisions concernant les axes de recherche à suivre ou développer de nouveaux produits, une gestion optimale des connaissances de l'organisation est indispensable. Pour qu'une entreprise soit « intelligente » - et donc fonctionne et se développe de manière convenable - la collecte, l'analyse, la validation, la valorisation, le stockage et la recherche d'information sont primordiaux. Ainsi, la valeur ajoutée d'une organisation reposera sur la capacité des employés à savoir comprendre, interpréter, réutiliser l'information pour être en mesure de créer des connaissances et d'innover pour contribuer à son bon développement[1].

Néanmoins, les sources d'informations structurées, comme les bases de données, et non-structurées, comme les documents d'activité ou les *e-mails*, se sont multipliées. Les volumes d'informations et de données ont fortement augmenté et leur rythme d'échange s'est intensifié : l'appropriation des outils informatiques par une grande majorité des professions a conduit à une multiplication des supports de diffusion et de stockage. Dans les organisations, l'information est alors surabondante - en 1996, David Shenk a parlé pour la première fois d' « infobésité » - elle est devenue beaucoup plus coûteuse à gérer et à conserver. Mettre en place une gestion intelligente des informations et limiter le coût de leur conservation est alors devenu une des préoccupations majeures pour les entreprises.

Par exemple, dans un objectif de gestion améliorée des informations et des connaissances, de nombreuses entreprises ont investi dans des outils de Gestion Électronique des Documents (GED)

[1] VUILLEQUEZ Jean-Yves. *Le moteur de recherche d'entreprise : quels enjeux organisationnels et technologiques ?* [en ligne]. Mémoire pour obtenir le Titre professionnel « Chef de projet en ingénierie documentaire » INTD - CNAM, 2013.125 p. [Consulté le 25/11/2017]. Disponible à l'adresse : https://memsic.ccsd.cnrs.fr/mem_00945629/document

ou dans un *Enterprise Content Management* (ECM), ou Gestion de contenu d'entreprise qui permet de mutualiser les outils et sources existantes de l'organisation. Dans ces systèmes d'information, des moteurs de recherche d'entreprise ont été implantés avec la promesse de trouver rapidement de l'information depuis un point d'accès unique, sans que l'utilisateur sache par avance dans quelle base ou application se trouve l'information.

Pourtant, la recherche d'information sur les moteurs semble souvent être une activité peu efficace et mal exploitée. Une enquête de Delphi Group menée en 2009 dans quinze grandes entreprises américaines a montré que les employés passaient plus d'un tiers de leur temps de travail à chercher de l'information, ce qui est considérable. Mscdermott a également montré que 38 % des employés des grandes entreprises employaient la majorité de leur temps à rechercher des informations. Pourtant, selon un rapport interne de Google de 2008, seul un quart des recherches menées dans les grandes entreprises renverrait une réponse pertinente. En outre, plus de la moitié de ces activités consisterait à recréer des informations déjà existantes, à collecter des documents sans les analyser, à effectuer des recherches sans trouver de résultats pertinents ou à convertir des informations sous d'autres formats. Ces activités ne sont pas productives, et donc non-rentables[2]. Au vu de ces études, la recherche d'informations est loin d'être optimale en entreprise. Les employés perdent un temps précieux qui pourrait être employé à des tâches plus importantes, et la firme perd en efficacité.

En effet, nombre de recherches sont infructueuses avec les moteurs de recherche syntaxiques, qui sont les moteurs dits « classiques ». Ils sont seulement fondés sur des techniques statistiques : lors d'une requête, le moteur de recherche va comparer une chaîne de caractères avec celles présentes au sein des documents de son index. Cet outil ne prend donc pas en compte le sens de la requête. Il y aura ainsi des problèmes de synonymie et de polysémie lors de la phase de la recherche, les taux de rappel et de précision des documents renvoyés ne seront pas totalement satisfaisants.

Une des solutions pour optimiser la recherche d'information est l'implantation d'une ontologie dans un moteur de recherche. Pour Jacques Chaumier :

> « Une ontologie définit les concepts d'un domaine (principes, idées, catégorie d'objet, notions abstraites) et les relations entre ces concepts ainsi que des règles et axiomes qui les contraignent. Elles sont orientées vers l'expression des connaissances et fournissent le vocabulaire spécifique à un domaine.[3]»

[2] DINET Jérôme. *La recherche d'information dans les environnements numériques.* Londres : ISTE éditions, 2014. 134 p. Systèmes d'information, Web et informatique ubiquitaire. ISBN 978-1-78405-018-4

Elles correspondent donc à un vocabulaire et un langage formel associés, c'est une sorte de grammaire qui indique la façon dont les termes peuvent être employés les uns avec les autres. Ces ontologies peuvent jouer le rôle de référentiel conceptuel utilisable par une machine dans le cadre d'une modélisation. Elles offrent donc une représentation structurée d'un domaine de savoir, qui pourra être exploitée par des applications diverses. C'est pour cela que les concepts du domaine et les propriétés possèdent le plus souvent des définitions lisibles par les machines [4].

En intégrant une ontologie dans un moteur de recherche, celui-ci deviendra alors sémantique : l'ontologie peut améliorer la pertinence d'une recherche d'information en renvoyant des documents concernant un concept précis, au lieu de se baser sur des mots-clés qui peuvent être ambigus. Elle permet donc de désambiguïser les requêtes. Par exemple, dans l'ontologie, il sera spécifié que w*indow* (la fenêtre) et *Windows* (système d'exploitation de Microsoft) n'ont pas le même sens, ce qui évitera de nombreuses confusions lors d'une recherche. Des fonctions supplémentaires pourront également être proposées : auto-complétion et expansion des requêtes, correction orthographique, etc. Cette recherche basée sur les ontologies est donc une recherche intelligente qui prend appui sur la sémantique des ressources et sur les concepts contenus dans les documents.

Néanmoins, le processus d'élaboration d'une ontologie est long et complexe : il faut modéliser un domaine de connaissances, de manière parfois précise. Des connaissances sur le domaine en question sont alors demandées, il est également indispensable de faire de nombreux choix de modélisation. Des compétences informatiques sont en outre attendues pour éditer l'ontologie, la formaliser et l'implanter dans un Système d'Information (SI). Dans beaucoup de projets, plusieurs acteurs peuvent intervenir à différentes étapes de la construction de l'ontologie ; l'équipe est donc pluridisciplinaire. Khadim Drame explique alors qu'il est nécessaire « d'utiliser des méthodes ou méthodologies pour seconder le processus de construction des ontologies [5]», processus très demandeur en temps et en ressources humaines. Jusqu'en 1995, les premières ontologies ont été construites de manière artisanale, leurs concepteurs n'ont pas suivi de méthodologie prédéfinie. À partir de 1998, des cadres méthodologiques de plus en plus élaborés et qui font maintenant

---

3 CHAUMIER Jacques. Les ontologies. Antécédents, aspects techniques et limites. *Documentaliste-Sciences de l'Information*, 2007, Vol. 44, n°1, p. 81-83. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-81.htm

4 Ibid.

5 DRAME Khadim. *Contribution à la construction d'ontologies et à la recherche d'information :application au domaine médical* [en ligne]. Thèse pour obtenir le grade de Docteur en informatique et santé. Université de Bordeaux, 2014. 187 p. [Consulté le 20/02/2017]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-01166042/document , p.25

référence sont conceptualisés[6]. Pour développer une ontologie, les professionnels peuvent donc s'appuyer sur de nombreuses recommandations et méthodes solides.

Pourtant, on peut se demander si ces méthodologies sont encore employées avec exactitude par les concepteurs d'ontologies. En effet, beaucoup de méthodes datent du début du siècle et sont peut-être obsolètes, d'autant plus que ces dernières années, de nombreuses ontologies ont été élaborées dans divers secteurs, notamment celui de l'industrie. Ainsi, des pratiques professionnelles plus innovantes sont probablement apparues. On peut aussi supposer que quelque méthodologies sont trop rigides, trop coûteuses en temps et se plient moins au développement d'une ontologie dans un contexte professionnel. Il est aussi possible que les concepteurs jugent ces méthodologies trop séquentielles. Ils pourraient décider de s'en affranchir pour élaborer de manière davantage itérative une ressource qui serait plus adaptée aux besoins des utilisateurs finaux de l'ontologie.

Nous avons alors formulé trois hypothèses principales auxquelles nous répondrons au fil de cet écrit :

- Les méthodologies de construction d'ontologies conceptualisées dans la littérature ne sont plus vraiment utilisées, même si des éléments méthodologiques sont repris dans les projets.

- Les méthodes utilisées aujourd'hui sont moins formalisées et plus itératives pour que les ontologies soient davantage adaptées aux usages des utilisateurs futurs.

- Il y a une différence de points de vue et de conceptions entre les chercheurs et « les professionnels » (informaticiens, documentalistes…), leur but n'est pas le même lors de la construction d'une ontologie.

Tout d'abord, nous ferons un état de l'art sur la question en abordant, entre autres, le rôle des ontologies dans le Web des données, leurs diverses applications pour aider à la recherche d'information et les différentes méthodologies de construction conceptualisées dans la littérature. Ensuite, notre méthodologie d'expérimentation sera présentée. En effet, pour vérifier nos hypothèses, nous avons mené des entretiens semi-directifs auprès de six personnes qui ont participé à des projets de construction d'ontologies dans des domaines aussi divers que la presse, le médical ou l'industrie. Pour finir, l'analyse des résultats de notre expérimentation nous permettra de voir de quelle manière l'élaboration d'ontologies est aujourd'hui appréhendée. Nous nous centrerons notamment sur l'articulation - parfois complexe - entre pratiques « officielles », réalité professionnelle et adéquation aux usages des publics cibles.

[6] Ibid.

1. Les ontologies et la recherche d'information : état de l'art

Cette première partie vise à poser les fondements de notre sujet et à l'ancrer dans un contexte plus global. Ainsi, nous discuterons des problématique et enjeux que soulève l'intégration des ontologies dans les SI afin d'améliorer la recherche d'information. Nous nous focaliserons également sur les différentes méthodologies de construction d'ontologies.

Dans le Web 1.0 ou Web documentaire, on constate de nombreuses limites dans la recherche d'information, ce qui conduit fréquemment au bruit ou silence documentaire. Un des objectifs de la mise en place du Web sémantique - ensuite appelé plus justement « Web des données » - est l'amélioration de la recherche, il est notamment possible de rechercher par concepts ou de trouver des informations d'une granularité plus fine. Dans ce contexte, les Systèmes d'Organisation des Connaissances (SOC) occupent une place prépondérante puisqu'ils permettent de définir des termes ou concepts (et éventuellement des relations entre eux) afin qu'un vocabulaire commun soit partagé par des usagers. L'indexation et la formulation des requêtes sont alors améliorées. Les ontologies informatiques constituent le SOC le plus complexe et le plus formalisé. Après avoir présenté l'étymologie et différentes définitions du terme, nous découvrirons les fonctionnalités relatives à la recherche qu'elles peuvent améliorer : indexation et annotation sémantiques, aide à l'appariement entre les documents et la requête, reformulation et expansion des requêtes, etc. Chaque projet est unique et spécifique dans une organisation, une ontologie est ainsi souvent conçue spécialement pour celui-ci. Nous verrons alors comment les méthodologies de construction sont choisies en fonction des spécificités du projet, afin de développer l'ontologie la mieux adaptée aux usages auxquels elle se destine.

1.1 Un changement de paradigme de la recherche d'information avec le Web des données

1.1.1 Du Web documentaire au Web des données

Le Web documentaire

Le Web 1.0 a été créé par Tim Berners-Lee et Robert Cailliau au CERN (Centre européen de recherche nucléaire). Ainsi, à la fin de l'année 1990, lors du projet WebCore, le premier serveur et le premier navigateur sont testés via une connexion internet ; le navigateur se nomm*e*

WorldWideWeb[7]. Lors de la naissance du Web, la métaphore « bibliothèque universelle[8] » est largement employée : cette nouvelle invention est vue comme un grand système documentaire au sein duquel tous les documents sont reliés les uns aux autres par des liens hypertextes, il est très facile d'accéder aux documents et à l'information. On navigue ainsi dans cette toile de pages Web en suivant des liens et on peut marquer d'un signet les pages qui ont éveillé notre intérêt, comme on le ferait grâce à un marque-page.

Selon Fabien Gandon, l'architecture du Web ne désigne pas l'objet Web, la toile que nous parcourons durant notre navigation, elle se réfère plutôt aux standards définissant l'infrastructure technologique du Web ; ces standards vont permettre son développement et sa normalisation. En 1994, le Massachusetts Institute of Technology (MIT), l'Université de Keio au Japon et l'Institut national de recherche en informatique et en automatique (INRIA) - où le projet WebCore du CERN a été transféré - se réunissent et créent le World Wide Web Consortium (W3C) afin de standardiser l'architecture du Web. Le W3C a notamment formalisé les trois notions fondamentales à l'origine et qui constituent le cœur du Web documentaire :

- L'*Universal Resource Identifier* (URI) : c'est un format d'identifiants uniques permettant de nommer n'importe quelle ressource sur le Web. De plus, si l'identifiant offre un chemin d'accès vers une représentation de la ressource, c'est une *Uniform Resource Locator* (URL) ou adresse Web.
- L'*HyperText Transfer Protocol* (HTTP) permet via une adresse URL d'accéder à une page Web identifiée et localisée par cette URL.
- *HyperText Markup Language* (HTML) est un langage de balisage que l'on utilise pour représenter, mettre en forme et publier des pages Web[9].

Vers le Web sémantique

Rapidement, des limites du Web 1.0 apparaissent : les ressources sont publiées sur le Web sans autre traitement que leur mise en forme et on peut simplement interagir en activant des liens hypertextes ; on dispose seulement du contenu du document en lui-même et du format de publication. En effet, il n'y a pas de traitement informatique pour préciser, par exemple, le type ou la structure du document, ni de terme ou concept pour caractériser le contenu d'une ressource. Pour Bruno Bachimont, « Le Web 1.0 fait ses traitements à l'aveugle, en ne prenant en compte que

---

7 Traduit par « Toile d'envergure mondiale »

8 BACHIMONT Bruno et al..Enjeux et technologies : des données au sens, *Documentaliste-Sciences de l'Information*, 2011, Vol. 48, n°4, p. 24-41. Également disponible en ligne à l'adresse : http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2011-4-page-24.html, p.27

9 BACHIMONT Bruno et al. Op.cit. p.27

le format de codage des contenus, mais non la sémantique de ces derniers[10].» Les liens entre les ressources sont construits au niveau de textes constitués de chaînes de caractères ; les liens ne sont pas situés au niveau des concepts qui caractérisent une ressource. Ainsi, les ressources sont une succession de chaînes de caractère, mais on ne dégage pas le sens de celles-ci. De plus, habituellement, l'utilisateur cherche moins les documents que l'information qu'ils contiennent.

Dès la première conférence du W3C en 1994, Tim Berners-Lee déclare que penser le Web seulement comme « un espace documentaire avec des liens entre les documents » est réducteur, cela revient à ne prendre en compte qu'un seul aspect de la problématique. Effectivement, les internautes ne naviguent pas sur le Web de manière aléatoire, ils mobilisent des modèles de connaissances, « une carte mentale » qu'ils possèdent à propos du monde qui les entoure. Si la machine arrive à comprendre les modèles de représentation des utilisateurs - même partiellement -, les interactions entre les internautes et le Web pourraient être grandement facilitées. C'est ce que le W3C appelle le Web sémantique, il faut enrichir les ressources de connaissances supplémentaires concernant le sens des contenus afin d'exploiter les documents de manière plus précise et performante. C'est tout l'objet des ontologies informatiques : elles consistent à capturer la représentation d'un domaine de connaissances afin d'aider la machine à fournir une réponse adéquate à la requête d'un internaute[11].

Il faudra environ dix ans au W3C pour concevoir les outils du Web sémantique, ils sont présentés ci-dessous :

Un extrait du « sandwich du Web sémantique » In. BACHIMONT, 2011, p.28

10 BACHIMONT Bruno et al. Op.cit. p.24

11 BACHIMONT Bruno et al. Op.cit. p.28

Le Web sémantique va alors constituer une extension du Web documentaire car il repose sur deux de ces notions fondamentales, les URI et le protocole HTTP, ainsi que sur un nouveau standard, *Ressource Description Framework* (RDF) :

- Les URI sont utilisées pour identifier non seulement les documents, mais aussi les lieux, les organisations, les personnes ou d'autres objets dont la matérialisation est différente de celle des documents, pages ou sites Web.
- HTTP permet de rendre toutes les URI consultables via son architecture de type client-serveur[12].
- Le langage RDF est la première brique des standards du Web sémantique, il désigne un modèle mais aussi plusieurs syntaxes, dont une en *Extensible Markup Language* (XML). Fabien Gandon déclare que « RDF est au Web de données ce que HTML était au Web documentaire dans sa métaphore initiale : le langage dans lequel on décrit, représente et relie des ressources à échanger sur le Web[13]». Effectivement, il permet de décrire n'importe quelle ressource, notamment sur le Web, par exemple en précisant son auteur, sa date de création, les droits de diffusion d'un film, etc. Son fonctionnement est basé sur la forme de triplets sujet-prédicat-objet.

Au-dessus de ces trois notions fondamentales, on trouve des recommandations supplémentaires :

- Le langage d'interrogation *Protocol and RDF Query Language* (SparQL) qui permet de construire des requêtes sur des données en RDF.
- Les schémas *RDF Schema* (RDFS) et *Web Ontology Language* (OWL) permettent de déclarer et décrire des types de ressources manipulées, les classes et les relations entre ces classes. Par exemple, avec RDFS, on peut définir des vocabulaires utilisés dans les graphes RDF. Ces schémas sont aussi utilisés pour formaliser des ontologies[14].
- La dernière brique recommandée par le W3C est la norme *Rule Interchange Format* (RIF) qui permet d'échanger des règles d'inférence (ou règles de raisonnement) dans le Web sémantique[15].

---

12 PAQUETTE Gilbert, *Introduction aux technologies sémantiques* [en ligne]. Cours Technologies sémantiques pour la gestion des connaissances. Université TELUQ (Québec), 2015. 25 p. [Consulté le 25/11/2017]. Disponible à l'adresse : http://inf6070.teluq.ca/teluqDownload.php?file=2013/07/INF6070_M1_a5_ApplicationTechnologiesSemantiquesGC.pdf, p.7

13 BACHIMONT Bruno et al. Op.cit. p.28

14 Cf. « Langages informatiques et niveaux de complexité » , p.34

15 BACHIMONT Bruno et al. Op.cit. p.28

Le Web des données

Aujourd'hui, l'emploi de l'expression « Web des données » est privilégié par rapport à l'expression « Web sémantique ». En effet, les machines n'ont pas encore réussi à saisir le sens des contenus mais le Web des données peut être perçu comme « la première phase de déploiement massif du Web sémantique »[16]. Il utilise les technologies du Web sémantique pour relier les données entre elles. En utilisant cette terminologie, on insiste sur la possibilité d'ouvrir les données engoncées dans les silos de tailles diverses (des carnets d'adresses jusqu'aux gigantesques bases de données), puis de les échanger et les relier selon les besoins des internautes. Ainsi, l'objectif principal du Web des données est de constituer une toile dans laquelle les internautes pourraient aisément naviguer d'une donnée à l'autre[17].

Le Web des données, qui transcende le Web documentaire et s'appuie sur des outils du Web sémantique, constitue un nouveau terrain d'expérimentation. Les liens ne sont plus simplement présents entre deux documents, ils peuvent exister entre des objets, des concepts ; les internautes peuvent notamment accéder à une information d'une granularité plus fine. Nous constatons alors que le Web des données promet d'importants changements concernant la recherche d'information. Nous aborderons ce thème en présentant la recherche classique. Celle-ci souffre de nombreux défauts. Pour résoudre ces limites, la recherche sémantique apparaît comme une solution efficace.

1.1.2 Entre recherche classique et recherche sémantique

La recherche d'information classique

Pour l'Association des professionnels de l'information et de la documentation (ADBS) dans *Le Vocabulaire de la documentation,* la recherche de l'information est circonscrite à « l'ensemble des méthodes, procédures et techniques ayant pour objet d'extraire d'un document ou d'un ensemble de documents les informations pertinentes[18].» Comme le font remarquer Nicole Boubée et André Tricot, la recherche d'information n'est pas une fin en soi et s'inscrit dans une tâche et un contexte plus globaux. Par exemple, des internautes recherchent des statistiques pour nourrir un rapport, qui est lui-même un document appuyant un projet plus large[19].

---

16 GANDON Fabien, FARON-ZUCKER Catherine, CORBY Olivier. *Le Web sémantique : comment lier les données et les schémas sur le Web ?*. Paris : Dunod, 2012. 206 p. ISBN 978-2-10-057294-6, p.20

17 BACHIMONT Bruno et al. Op.cit. p.30

18 BOULOGNE Arlette (coordonné par). *Vocabulaire de la documentation*. Paris : ADBS Éditions, 2004. (Sciences et techniques de l'information). ISBN 2-84365-071-2, Définitions également disponible à l'adresse : http://www.techno-science.net/?onglet=glossaire&definition=11203

Ingrid Pamela Mafokoua Tchigui ajoute que l'objectif de la recherche d'information « est de fournir à un utilisateur interrogeant un système, les documents les plus pertinents par rapport à sa requête[20]». Ce sont des Systèmes de Recherche d'Information (SRI) que les internautes vont utiliser ; ces SRI permettent de faire le lien entre une requête et des ressources. Ce sont donc des interfaces mettant en relation une source (collection) contenant potentiellement un nombre considérable de documents et des internautes cherchant, via des requêtes, des informations susceptibles de se trouver dans cette collection de ressources[21]. Lors d'une recherche, ces SRI visent à diminuer le silence documentaire - c'est-à-dire l'absence de documents pertinents dans les résultats de la requête - ainsi que le bruit documentaire qui désigne la proportion de documents non-pertinents parmi ceux renvoyés par le SRI. Pour remplir ces objectifs. Trois processus sont mis en œuvre au sein du système :

- Le processus d'indexation vise à offrir un modèle de représentation et de description pour décrire le contenu d'un document ou d'« une granule d'information plus fine », comme un chapitre ou un passage dudit document.
- Le processus d'appariement sert à sélectionner des documents pertinents par rapport à une requête spécifique. Cette pertinence est évaluée par une fonction d'appariement qui attribue un ordre de pertinence à des ressources pour une requête, le « score de pertinence » est généralement compris entre 0 et 1. Ainsi, cette fonction permet de définir quelles sont les ressources pertinentes et leur classement dans la page de résultats.
- Le processus de reformulation des requêtes prend en compte les retours des internautes concernant les résultats renvoyés par le SRI ou les paramètres d'une requête (modification de la recherche, proposition de correction orthographique, etc.) afin d'améliorer celle-ci et de fournir des résultats plus pertinents aux utilisateurs[22].

Le moteur de recherche est au cœur du SRI. Le principal objectif de cet outil d'aide à la recherche d'information est de fournir des documents répondant à une requête formulée par un utilisateur.

---

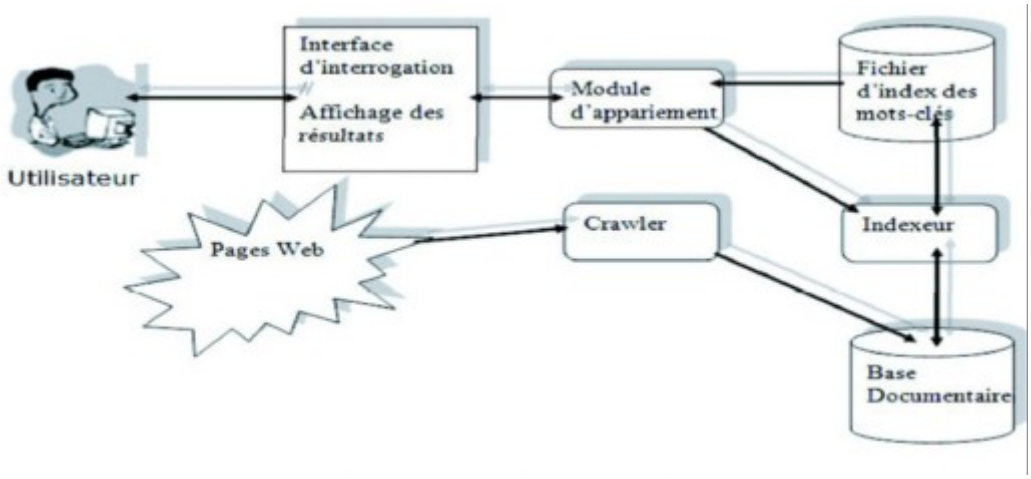
19 BOUBEE Nicole, TRICOT André. *Qu'est-ce que rechercher de l'information ?*. Villeurbanne : Presses de l'ENSSIB, 2010. 978-2-910227-83-8, p.59, Définition également disponible à l'adresse : http://books.openedition.org/pressesenssib/805

20 MAFOKOUA TCHIGUI Ingrid Pamela. Recherche d'informations dans le Web sémantique. *Supinfo.com* [en ligne]. 30 octobre 2016. [Consulté le 25/11/2017]. Disponible à l'adresse : https://www.supinfo.com/articles/single/3373-recherche-informations-clans-Web-semantique , p.1

21 SY Mohameth-François. *Utilisation d'ontologies comme support à la recherche et à la navigation dans une collection de documents* [en ligne]. Thèse pour l'obtention du grade de docteur en Informatique. Université de Montpellier II, 2012. 135 p. [Consulté le 25/11/2017]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-00822516 , p.14

22 Ibid., p.5

Il est composé de cinq modules, l'interface graphique, le *crawler*, le module d'indexation ainsi que le module d'appariement :



Architecture d'un moteur de recherche / MAFOKOUA TCHIGUI Ingrid Pamela. Op.cit p.4

- Sur l'interface graphique, la requête formulée par l'utilisateur est récupérée. Les résultats sont ensuite affichés : en cas de succès, le titre de la ressource et un lien pour y accéder sont généralement proposés. En cas d'échec de la requête, des suggestions pour mieux présenter la demande apparaissent.
- Aussi appelé *spider*, le *crawler* parcourt des documents en suivant des liens hypertextes et moissonne les données structurantes (titre, nom des chapitres, termes spécifiques). Celles-ci sont ensuite emmagasinées dans des bases de données ou bases documentaires, où elles seront indexées. Le crawler joue surtout un rôle important dans des moteurs de recherche parcourant le Web et moins dans des SRI d'entreprises ou de collectivités.
- L'indexation, via le module d'indexation, est une des fonctions primordiales des moteurs. Les mots contenus dans les différentes ressources ou pages Web visitées par le crawler, sont triés pour construire un index. Dans cet index, on associe à chaque mot l'ensemble des ressources où il est présent ; cela revient à créer un fichier de type mots-clés = adresses URL. Avant de prendre place dans l'index, ces mots-clés sont traités automatiquement par lemmatisation[23], troncature[24]. Ces éléments indexés se transforment alors en chaînes de caractères qui n'ont plus besoin d'avoir de sens pour les

[23] Filtrage des mots ou groupes de mots non significatifs (« mots vides »), dont les articles et les prépositions

[24] Choix d'une unité d'indexation unique pour différentes formes d'un mot (variantes morphologiques). Par exemple, un verbe à l'infinitif pourra être le représentant de toutes les formes conjuguées de ce verbe (« manger » représentant de « mange », « mangeait », « mangeâmes »)

humains. Ils sont conçus pour optimiser l'association entre une requête d'utilisateur et les ressources correspondant le plus à sa requête[25]. Ici, il s'agira alors d'indexation classique ou syntaxique, on n'analyse pas le sens des termes, ce sont simplement des chaînes de caractères. Une approche lexicale est ainsi mise en place ; par exemple, si l'occurrence d'un terme est élevée dans une page Web, l'utilisateur qui saisit ce terme dans sa requête aura plus de chance de voir apparaître cette page Web en bonne position dans les résultats de sa requête.

- Le module d'appariement tient le même rôle que dans un SRI, c'est l'algorithme du logiciel de recherche qui récupère les termes de l'internaute dans sa requête, les analyse et les compare avec les termes de l'index, afin que les documents les plus pertinents soient renvoyés dans l'interface[26].

Limites de la recherche classique

Les principaux SRI sont majoritairement basés sur l'utilisation du langage naturel pour la représentation des documents et dans les requêtes. Par exemple, comme dit précédemment, la statistique d'occurrence de termes dans les ressources est très utilisée pour évaluer leur importance. Un document dans lequel un mot est répété de nombreuses fois sera obligatoirement pertinent pour des personnes recherchant ce mot. Les SRI se limitent à la comparaison d'une chaîne de caractères avec une autre chaîne pour renvoyer les documents les plus adaptés. Les ressources que le *crawler* va visiter - et dont les termes vont être indexés dans des bases de données - sont donc vues comme des « réservoirs de termes ». Cette approche classique syntaxique souffre pourtant de limites récurrentes[27].

Toutes les disciplines qui utilisent le langage comme matériel doivent faire face aux problèmes d'ambiguïté du langage naturel : un mot ou une expression possèdent souvent plusieurs sens selon le contexte dans lequel on les emploie. Les SRI vont alors se heurter à des obstacles linguistiques et sémantiques.

Il existe également des problèmes de synonymie, ce sont des mots différents mais qui ont le même sens. Lorsqu'un utilisateur cherche les termes « voiture », « bagnole » ou bien « véhicule », il devrait pouvoir consulter les mêmes résultats de recherche puisque ces termes ont le même sens. Cependant, les chaînes de caractères à rechercher ne seront pas les mêmes et l'internaute

25 AUSSENAC-GILLES Nathalie. Le Web sémantique, quel renouvellement pour la recherche d'information ?. In : BOUGHANEM Mohand, SAVOY Jacques. *Recherche d'information: état des lieux et perspectives.* Cachan: Lavoisier, 2008. p. 97-132. ISBN 978-2746220058. Chapitre également disponible en ligne à l'adresse : https://www.irit.fr/publis/IC3/aussenac-LivreBoughanem2007_18janv.pdf , p.10

26 MAFOKOUA TCHIGUI Ingrid Pamela. Op.cit p.4

27 SY Mohameth-François. Op.cit. p.119

pourra manquer certains résultats s'il n'a pas cherché un terme exact. En effet, les utilisateurs ont l'habitude d'employer une grande variété de termes pour désigner le même concept. Trouver le document ou l'information exacte devient alors très compliqué[28].

À l'inverse, la polysémie désigne des mots identiques mais qui ont plusieurs sens. Le terme « glace » fait référence à un miroir mais aussi à une crème glacée[29]. De même, le mot « Java » peut désigner aussi bien un langage informatique, qu'une danse ou une île : la personne recherchant ce terme rencontrera du bruit documentaire. Selon le sens qu'elle place derrière le terme, il faudra probablement qu'elle consulte plusieurs documents ou qu'elle précise sa requête pour obtenir l'information qui lui sera utile.

De plus, via les moteurs de recherche syntaxiques actuels, il est impossible d'obtenir une réponse complète à des requêtes complexes comme « Où pourrais-je aller en vacances, pour une semaine, le mois prochain, avec deux enfants pour moins de 2000 $? »[30]. La réponse à cette question demanderait une combinaison de plusieurs sources, cette tâche est aujourd'hui laissée aux internautes, même si plusieurs outils d'aide à la recherche ont été développés par Google Search. Dès 2008, Quick Answers a été mis en place pour que les internautes obtiennent rapidement et simplement des réponses à leurs questions spécifiques. Par exemple, un usager peut rechercher « altitude du Mont Everest » et la réponse « 8 848 m » lui sera fourni dès la page de résultats de Google sans qu'il n'ait besoin de consulter des sites Web[31].

La recherche classique, appuyée par des outils syntaxiques présents dans la majorité des SRI, se heurte à de nombreuses limites liées principalement à l'ambiguïté du langage naturel. La recherche sémantique, grandement facilitée avec le Web des données, est alors une opportunité pour résoudre ces problèmes. On espère dépasser les limites des recherches basées sur des requêtes formulées en langage naturel, et sur des index, qui s'apparentent, selon Nathalie Aussenac-Gilles, à des « sacs de mots tronqués et pondérés »[32].

---

28 FRONTERE Mikhail. *Assistance intelligente à la recherche d'information : élaboration d'un projet de moteur de recherche au service de la connaissance dans l'organisation [*en ligne]. Mémoire pour obtenir le Titre professionnel « Chef de projet en ingénierie documentaire ». INTD-CNAM, 2015. 170 p. [Consulté le 25/11/2017]. Disponible à l'adresse : https://memsic.ccsd.cnrs.fr/mem_01309438/document, p.75

29 MAFOKOUA TCHIGUI Ingrid Pamela. Op.cit. p.5

30 PAQUETTE Gilbert, Op.cit. p.3

31 https://www.1and1.fr/digitalguide/Web-marketing/search-engine-marketing/les-resultats-de-recherche-google-de-1998-a-aujourdhui/

32 AUSSENAC-GILLES Nathalie. Op.cit. p.8

La recherche sémantique

Les études en recherche d'information se sont tout d'abord centrées sur la mesure de l'efficacité des moteurs de recherche, sur les systèmes d'indexation ou sur les pratiques des utilisateurs pour mieux adapter les interfaces. Depuis une quinzaine d'années, l'angle d'approche évolue et les travaux sont largement influencés par les domaines de l'Intelligence Artificielle (IA) ou de l'Ingénierie des Connaissances (IC). On ne traite plus des textes comme des ensembles de termes agencés entre eux, que l'on analyse et que l'on pondère ; l'approche n'est plus strictement statistique.

L'objectif est maintenant de mettre en place des outils pour déterminer de quoi parle un texte, c'est-à-dire comprendre quels concepts sont développés dans un document. Ces concepts permettent de mettre en lumière des unités de sens exprimées dans les textes. Une fois cette étape franchie, il sera possible de retrouver des documents, parties de documents ou données abordant un sujet particulier ; il sera également aisé de construire des dossiers thématiques ou de répondre à des requêtes complexes et précises. En utilisant des concepts, il est possible de résoudre les problèmes de polysémie et d'ambiguïté des mots-clés.

Dans le cadre du Web des données, la recherche d'information va alors être renouvelée selon différents points de vue : l'information est présente en très grande quantité et sous des formes variées (données non-structurées dans des textes ou *e-mails*, données structurées dans des tableurs, etc.). Que ce soit sur le Web ou sur les serveurs des entreprises, il faut traiter de très gros volumes de documents, prendre en compte des formats divers et des ressources hétérogènes ainsi que l'interopérabilité[33] des systèmes informatiques[34]. Mettre en place des fonctions de recherche sémantique demande un outillage complexe : des SOC jouent un rôle important pour l'aide à la recherche de l'information.

1.1.3 Utiliser des SOC pour plus d'efficacité dans la recherche d'information

Pour indexer et rechercher de l'information grâce à des concepts, les professionnels peuvent s'appuyer sur des ressources, comme les thésaurus qui fournissent un vocabulaire contrôlé dans lesquels les termes sont structurés et reliés entre eux par des relations. Avec l'avènement du Web des données, l'accent est davantage placé sur un autre genre de SOC : les ontologies. Implantées dans un SRI, elles permettent de raisonner sur des connaissances et appuient la recherche

[33] Interopérabilité informatique : « capacité de matériels, de logiciels ou de protocoles différents à fonctionner ensemble et à partager des informations » selon larousse.fr

[34] AUSSENAC-GILLES Nathalie. Op.cit, p.15

d'information[35]. Des éléments de définition puis une comparaison des différents SOC s'imposent avant de présenter plus en détail les ontologies dans une seconde partie. Ce sont les SOC les plus formalisés, elles sont aujourd'hui de plus en plus utilisées et leurs possibilités d'exploitation sont les plus poussées.

Définitions

Dès les années 1990 avec l'avènement d'Internet, l'usage des outils de bureautique s'est généralisé, les échanges se sont internationalisés et le nombre de documents produits sous forme électronique a augmenté de façon exponentielle. Pour produire, indexer, rechercher ou exploiter ces documents, les professionnels de l'Information-documentation ont besoin de ressources décrivant les termes et concepts d'un domaine[36], ce sont les SOC[37] ou RTO[38]. L'expression est issue du monde des bibliothèques, et plus particulièrement de celui des bibliothèques numériques.

Pour Manuel Zacklad, les SOC comprennent « tous les types de schémas permettant d'organiser des informations et de promouvoir la gestion des connaissances ». Cette notion englobe ainsi des outils diversifiés avec des niveaux de formalisation très disparates. Effectivement, en prenant appui sur cette définition, les schémas suivants sont considérés comme des SOC : schémas de classification organisant l'information à un niveau global, vedettes-matière qui permettent un accès plus détaillé, glossaires ou encore fichiers d'autorité qui contrôlent les variantes orthographiques de noms géographiques ou de noms propres. On peut aussi inclure dans les SOC les vocabulaires davantage structurés comme les thésaurus, les réseaux sémantiques ou les ontologies[39]. Manuel Zacklad compartimente les SOC au sein de trois catégories, les listes de termes, les classifications et catégories, puis les listes de relations :

Listes de termes	Classifications et catégories	Listes de relations
Fichiers d'autorité	Vedettes-matière	Thésaurus
Glossaires	Schèmes de classification*	Réseaux sémantiques
Dictionnaires	Taxinomies*	Ontologies
Répertoires géographiques	Schèmes de catégorisation*	

Typologie de systèmes d'organisation des connaissances, ZACKLAD Manuel, Op.cit p.8

35 AUSSENAC-GILLES Nathalie. Op.cit. p.8

36 BOURIGAULT Didier, AUSSENAC-GILLES Nathalie. Construction d'ontologies à partir de textes. *Conférence TALN 2003*, Juin 2003 [en ligne]. p. 11-14.[Consulté le 25/11/2017]. Disponible à l'adresse : http://www.atala.org/taln_archives/TALN/TALN-2003/taln-2003-tutoriel-002.pdf, p.2

37 En anglais : *Knowledge Organization System* (KOS)

38 Ressource Termino-ontologique, expression employée au début des années 2000, on lui préfère aujourd'hui l'expression « Système d'Organisation des Connaissances »

39 ZACKLAD Manuel, Systèmes d'organisation des connaissances hétérogènes pour les applications documentaires, *Document numérique*, 2010, Vol. 13, n°2, p. 7-12. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-document-numerique-2010-2-page-7.htm , p.7

Comparaison entre différents SOC

Mikhail Frontère approfondit la comparaison entre ces différents SOC en s'attachant à leur niveau de formalisation. Les listes de termes sont les plus simples et se contentent de fournir une liste de termes et/ou d'identités nommées, sans établir de relations entre les termes du référentiel. C'est un de leur principal inconvénient, la sémantique (le sens) de l'association des termes entre eux n'est pas renseignée. D'autres SOC sont plus élaborés, ils sont composés des termes et/ou concepts, mais aussi de relations de sens entre ces entités. Ainsi, ces liens sémantiques relient les composants, cela permet de connaître l'organisation de la connaissance au sein du domaine qui est décrit. Dans cette seconde catégorie de référentiels, on peut placer les réseaux sémantiques, thésaurus et ontologies. Parmi les SOC proposant des relations sémantiques, ces deux derniers sont les plus utilisés[40].

On peut donc différencier ces outils en prenant appui sur leur degré de formalisation, mais ce n'est pas le seul critère ; d'autres différences existent concernant la nature de leur contenu et leur structure. Au premier abord, certains schémas de données peuvent paraître semblables, il est alors important de les comparer et de passer en revue leurs similitudes et différences.

Taxinomies et ontologies

Le terme taxinomie, aussi orthographié taxonomie, a été créé dans les années 1800 par un botaniste ; cet outil sert à classer et décrire les êtres vivants de manière hiérarchique. La définition de ce terme s'est élargie et depuis la fin des années 1990, il désigne majoritairement des arborescences guidant la navigation au sein des sites Web et des portails intranet des entreprises. Il existe quelques similitudes entre les ontologies et les taxinomies : ces dernières, tout comme les ontologies, sont arborescentes et peu sont construites par des professionnels de l'information-documentation. De plus, l'organisation d'une ontologie est habituellement taxonomique, c'est-à-dire que ses catégories sont placées les unes sous les autres et reliées par des relations de subsomption, autrement dit des liens hiérarchiques.

Pourtant, Fabien Gandon fait remarquer que « les connaissances ontologiques dépassent largement les connaissances taxinomiques [41]» : dans la majorité des ontologies, il existe d'autres genres de relations. Elles sont beaucoup plus diverses et complexes, elles peuvent aussi être composées de définitions complètes, de contraintes, de fonctions de calcul, etc.[42] Les ontologies ne sont pas visibles par les usagers d'une application, elles sont directement intégrées dans un SRI

---

40 FRONTERE Mikhail. Op.cit.p.76

41 GANDON Fabien. Ontologies informatiques. *Interstices.info.* [En ligne]. 22 mai 2006. [consulté le 25/11/2017]. Disponible à l'adresse : https://interstices.info/jcms/c_17672/ontologies-informatiques

42 Cf. « Les relations entre les concepts », p.33

et rendent possible des raisonnements sur des connaissances pour servir la recherche d'information. En revanche, les taxinomies sont directement visibles par les usagers et facilitent seulement la navigation dans une application. Ainsi, même si ontologies et taxinomies peuvent parfois avoir une forme similaire, leurs buts principaux et leurs niveaux de formalisation varient considérablement[43].

Thésaurus et ontologies

Il convient tout d'abord de définir ce qu'est un thésaurus. Voici la définition présente dans la norme de 2011 sur les thésaurus et traduite par Sylvie Dalbin :

> « Thésaurus : vocabulaire contrôlé et structuré dans lequel les concepts sont représentés par des termes, organisés de façon à ce que des relations entre les concepts soient explicitées, et dont les termes préférentiels sont accompagnés par des entrées vers leurs synonymes ou quasi-synonymes.[44]»

Ainsi, ces termes représentent les concepts d'un domaine de connaissances, ils vont alors constituer un langage contrôlé pour l'indexation des documents et la recherche de ressources documentaires.

Les ontologies et thésaurus sont similaires sur certains points : ils représentent les éléments d'un domaine de connaissances particulier à l'aide de concepts. Ils sont également structurés selon des relations sémantiques similaires : hyponymiques (relations hiérarchiques générique/spécifique) et méronymiques (relations d'association)[45]. Mais même si les ontologies reprennent la structuration des thésaurus et les relations entre les concepts du domaine, elles le font de façon formelle et plus précise. On peut aussi ajouter qu'ontologies et thésaurus sont reliés entre eux puisqu'un thésaurus peut être vu comme un genre d'ontologie ou peut constituer une base pour la construction d'une ontologie.

Un thésaurus est principalement destiné à indexer et rechercher des documents, et certains peuvent lier des concepts et des documents entre eux. Les fonctions d'une ontologie diffèrent, elle va plutôt permettre de modéliser un champ de connaissances et de raisonner sur ces connaissances en utilisant des règles d'association et des propriétés rattachées aux concepts ; au

---

43 PICARD Anne-Claire (Le). *Le cycle de vie d'une ontologie : évaluation de l'ontologie du domaine de la Toxicologie Nucléaire. [*en ligne]. Mémoire pour obtenir le Titre professionnel « Chef de projet en ingénierie documentaire ». INTD-CNAM, 2014.174 p. [Consulté le 25/11/2017]. Disponible à l'adresse : https://memsic.ccsd.cnrs.fr/mem_01128938, p.35

44 DALBIN Sylvie. La norme "ISO 25964-1(2011) - Thésaurus pour la recherche documentaire" est publiée. *Dossierdoc.typepad.com*[En ligne]. 25 août 2011. [consulté le 21/03/2017]. Disponible à l'adresse : http://dossierdoc.typepad.com/descripteurs/2011/08/norme-iso-25964-1-thesaurus-publication-officielle.html

45 Cf. « Les relations entre les concepts », p.33

départ cet outil n'était pas conçu pour indexer des ressources. On peut alors voir l'ontologie comme « une extension des thésaurus » puisque des règles de déduction sont incorporées dans cet artefact formalisé[46].

Enfin, les termes d'un thésaurus sont exprimés dans un langage humain, même si on peut ensuite utiliser des technologies pour les rendre compréhensibles par des machines. Les ontologies sont élaborées pour être lisibles par des machines, le langage naturel est souvent remplacé par des identifiants qui ne sont pas directement compréhensibles par le cerveau humain. Ainsi, les ontologies sont rattachées au domaine de l'informatique ; les thésaurus font majoritairement partie du monde de l'Information-documentation et sont tout d'abord conçus pour être exploités par des humains[47]. Ontologies et thésaurus ont donc une structuration similaire et quelques relations communes sont utilisées dans ces deux outils ; pourtant leurs objectifs, leurs domaines d'application et leurs niveaux de complexité diffèrent.

Ainsi, la recherche d'information classique souffre de nombreuses limites car avec les moteurs de recherche syntaxiques, l'usager cherche simplement si un terme, qui est en fait une chaîne de caractères, est présent dans une des pages Web ou autre document indexé par un *crawler*. Ainsi, il y a souvent bruit ou silence documentaire à cause notamment des problèmes de synonymie ou de polysémie. Une des solutions est d'utiliser des SOC, des ressources décrivant les termes et/ou concepts d'un domaine de connaissances. Même si certains paraissent similaires, leurs niveaux de formalisation, leurs structures et leurs buts peuvent être différents. Il est donc important de choisir l'outil qui appuiera de manière satisfaisante les fonctionnalités de recherche que l'on souhaite mettre en place.

Dans un contexte informationnel où les technologies du Web sémantique se développent pour que les machines puissent comprendre « le sens » des informations, raisonner sur celles-ci et ainsi que la recherche d'information soit facilitée, les ontologies sont de plus en plus privilégiées. En effet, ce sont les SOC les plus complexes et formalisés, ils ont justement étaient pensés pour que la modélisation fine de domaines de connaissances soit possible et que la machine puisse faire des inférences. En outre, Nathalie Aussenac-Gilles défend l'idée que « les ontologies sont des représentations des connaissances d'autant plus pertinentes pour la recherche d'information qu'elles comportent une dimension terminologique ». Comme nous le verrons par la suite, plusieurs méthodologies de construction d'ontologies prennent appui sur des textes et sur le

---

46 LUSTREMENT Amandine. *Thésaurus et Web sémantique : quelles problématiques de mise en œuvre ?.* Mémoire pour l'obtention du Master esDOC. Université de Poitiers, 2016. 89 p., p.11

47 FRONTERE Mikhail. Op.cit. p.85

Traitement Automatique du Langage (TAL). La richesse terminologique des ontologies facilite l'association des concepts aux expressions linguistiques présentes dans des textes indexés par un *crawler.* Les ontologies sont donc des outils précieux pour faciliter une recherche d'information qui deviendra plus fine et efficace.

Après une rapide mise en contexte historique, quelques éléments de définition et une présentation des composants d'une ontologie, nous verrons en détail les fonctionnalités de recherche qu'une ontologie peut appuyer au sein d'un SRI, notamment au cœur du Web des données[48].

1.2 Les ontologies, des SOC performants au service de la recherche d'information

1.2.1 Mise en contexte

Approche historique

Le terme « Ontologie » a été construit en combinant deux racines grecques : *ontos* (ce qui existe, l'existant) et *logos* (le discours, l'étude). Ce terme se rapporte tout d'abord au domaine philosophique, l'Ontologie est une branche de la métaphysique abordant la notion d'existence et s'intéressant aux propriétés générales de ce qui existe. L'informatique a emprunté ce terme à la philosophie dès le début des années 1990, on est ainsi passé de la science ontologique à l'objet informatique, et de l'Ontologie à l'ontologie[49].

Les ontologies informatiques ont notamment été adoptées par la communauté internationale dont l'objet d'étude est l'acquisition des connaissances, cette communauté est aujourd'hui nommée Ingénierie des connaissances (IC). Le projet de l'IC consiste à créer des Systèmes à Base de Connaissances (SBC) qui effectuent des tâches ou résolvent des problèmes relevant d'activités intellectuelles ou cognitives. Le champ d'application des SBC est large puisqu'il comprend aussi bien des applications permettant l'automatisation du traitement d'une tâche, une modélisation conceptuelle d'un domaine ou encore l'assistance portée à un utilisateur[50]. L'IC est une catégorie de l'Intelligence Artificielle (IA), une branche de l'informatique en relation avec les sciences

---

48 AUSSENAC-GILLES Nathalie. Op.cit. p.8

49 GANDON Fabien. Op.Cit.

50 BACHIMONT Bruno, I*ngénierie des connaissances et des contenus. le numérique entre ontologies et documents.* Paris : Hermes science publications-Lavoisier, 2007. 279 p.. Collection Science informatique et SHS. ISBN 978-2746213692 Également disponible en ligne à l'adresse : https://stph.scenari-community.org/nf29/res/2007Bachimont_IngenierieDesConnaissancesEtDesContenus.pdf , p.48

cognitives[51]. Depuis les années 1970, un des objectifs de l'IA est de créer des systèmes experts, des logiciels capables d'effectuer des inférences à partir de règles prédéfinies. Dès les années 1990, les ontologies connaissent donc leurs premiers développements et tendent maintenant à remplacer les systèmes experts[52]. En effet, en IC, les ontologies répondent à des besoins de formalisation de connaissances et de standardisation des modèles pour favoriser leur interopérabilité et leur réutilisation ; elles peuvent alors faciliter les échanges de connaissances entre des systèmes formels et des applications informatiques. Ces fonctionnalités vont d'autant plus être exploitées dans un contexte où le Web des données se développe[53].

Afin de mieux appréhender la complexité des ontologies et avant de présenter leur rôle au sein du Web des données, il est important d'introduire des éléments de définition. En effet, les définitions du terme « ontologie » varient selon les angles d'étude concernant cette ressource.

Définitions

Nous prendrons principalement appui sur les définitions fournies ou retranscrites par Jean Charlet dans son mémoire *L'ingénierie des connaissances : développement et application pour la gestion de connaissances médicales*. Les ontologies sont donc apparues au sein de la communauté IC pour permettre de construire mieux et plus efficacement des SBC en réutilisant des connaissances d'un domaine. Jean Charlet dégage une première définition simple qui met en lumière le fait que le choix des objets présent dans une ontologie relève d'une décision personnelle :

> « Ontologie (déf. 1) : Ensemble des objets reconnus comme existant dans le domaine. Construire une ontologie, c'est aussi décider de la manière d'être et d'exister des objets.[54]»

Il est aussi important de rappeler que l'étude et les travaux sur les ontologies se sont développés dans un contexte informatique, domaine dans lequel un objectif principal est de construire des artefacts informatiques. Mentionner ce contexte est primordial pour saisir les buts poursuivis par les concepteurs d'ontologies et les contraintes qui se posent à eux lors de cette mise en œuvre. Lorsqu'on élabore un artefact, la question de la conceptualisation est cruciale : il est nécessaire de définir et de spécifier les concepts qui auront leur place dans l'ontologie. Une deuxième définition, dégagée au moment de l'élaboration de l'ontologie, prend alors davantage en compte la question

---

51 PICARD Anne-Claire (Le). Op.cit. p.20

52 FOURCASSIER Eric. *Des ontologies pour les humanités : Les ontologies de domaine à l'épreuve des humanités numériques.* Mémoire pour l'obtention du Master esDOC. Université de Poitiers, 2016. p.17

53 AUSSENAC-GILLES Nathalie. Op.cit. p.8

54 CHARLET Jean. *L'ingénierie des connaissances : développement et application pour la gestion de connaissances médicales.* Mémoire d'Habilitation à diriger des recherches. Université Pierre et Marie Curie, Paris, 2002. 143 p. [Consulté le 25/11/2017]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-00006920/document , p.44

du sens. Cette définition a été proposée par Thomas Gruber en 1993, c'est la plus synthétique et la plus couramment citée :

> « Ontologie (déf. 2) : Spécification formelle et explicite d'une conceptualisation partagée [55]»

Cette définition sous-entend tout d'abord qu'une communauté d'individus (« conceptualisation partagée ») doit se mettre d'accord sur une définition des concepts et des relations au sein d'un domaine de connaissances particulier afin de créer une ontologie. Quand ces éléments sont validés, il faut ensuite les structurer avec des outils de formalisation logique et informatique, c'est-à-dire les coder grâce à un langage opérationnel, exécutable (« spécification formelle ») : ces relations et concepts pourront être lisibles et exploités par des machines[56].

Pourtant, cette définition offre une vue figée de l'objet ontologie, l'étape de l'élaboration de l'ontologie est passée sous silence. En 1995, Nicola Guarino et Daniele Giaretta ont conceptualisé des définitions de l'ontologie durant son processus de création, ils voient l'ontologie comme « la représentation d'un système conceptuel ». Mike Uschold et Michael Gruninger se sont appuyés sur ces réflexions pour présenter d'autres éléments de définition :

> « Ontologie (déf. 3) : Une ontologie implique ou comprend une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts - entités, attributs, processus -, leurs définitions et leurs interrelations. On appelle cela une conceptualisation.[...] Une ontologie peut prendre différentes formes mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification. […] C'est une spécification rendant partiellement compte d'une conceptualisation. [57]»

Selon Fabien Gandon et Rose Dieng-Kuntz, on peut définir une conceptualisation comme « une structure sémantique intensionnelle qui capture les règles implicites contraignant la structure d'un morceau de réalité »[58]. L'ontologie est donc une représentation explicite partielle car sont seulement présents dans cette ressource, les éléments de conceptualisation d'un domaine de connaissances indispensables au bon fonctionnement du système. Les constructeurs d'ontologies ne visent pas à reproduire « le doublon formel d'un savoir », mais tendent à satisfaire les besoins

---

55 GRUBER Tom. À translation approach to portable ontology specifications. *Knowledge acquisition*. Vol. 5, n°2. 1993. p.199–220
Définition originale : *An ontology is an explicit specification of a conceptualization*

56 FOURCASSIER Eric. Op.cit. p.21

57 USCHOLD Mike, GRUNINGER Michael. Ontologies: Principles, Methods, and Applications. *Knowledge Engineering Review*. Vol. 11. N°2. Mars 1996. pp. 93–155.

58 GANDON Fabien, DIENG-KUNTZ Rose. Ontologie pour un système multi-agents dédié à une mémoire d'entreprise. *IC'2001, Ingénierie des Connaissances,* plateforme AFIA'2001, Juin 2001, (Grenoble, France) [en ligne]. [Consulté le 25/11/2017]. Disponible à l'adresse : https://hal.inria.fr/hal-01145808, p.5

informationnels d'une communauté d'utilisateurs. Effectivement, ils ne prétendent pas pouvoir créer une représentation, même partielle, du monde. Leur objectif est plus simplement de représenter « une conceptualisation portant sur un domaine circonscrit du savoir ». Ils décrivent des relations formelles entre des concepts et non des relations matérielles entre les objets représentés : les ontologies sont donc des objets documentaires qu'il faut traiter selon une approche strictement pragmatique[59].

En outre, toujours dans cette troisième définition, on peut lire qu' « une ontologie implique ou comprend une certaine vue du monde par rapport à un domaine donné ». Elle est élaborée et existe seulement grâce à un contexte donné qui lui donne son sens et sa structure. Les personnes en charge de la construction d'une ontologie prennent forcément des partis pris sur la conceptualisation d'un domaine. Deux ontologies centrées sur le même domaine de connaissances, mais élaborées dans des contextes différents ne seront jamais similaires. Comme le soulignent Jean Charlet, Bruno Bachimont et Raphaël Troncy, « Construire une ontologie, c'est décider de la manière d'être et d'exister des objets.[60]»

Les trois définitions proposées précédemment sont complémentaires puisqu'elles mettent toutes en lumière des points cruciaux concernant les ontologies. Ainsi l'ontologie est une conceptualisation parce que des concepts y sont définis ; elle devient par la suite un artefact informatique puisque les concepts et relations sont codés afin de devenir lisibles et exploitables par une machine ; enfin, la conceptualisation de l'ontologie reste partielle car elle dépend de la représentation du domaine par les concepteurs de l'ontologie, seuls les éléments nécessaires au bon fonctionnement de l'application sont représentés[61]

Plusieurs genres d'ontologies

Différents genres d'ontologies existent, ces ontologies ont toutes des objectifs pluriels que nous allons expliciter. Ainsi, Mustapha Baziz distingue plusieurs familles d'ontologies ayant chacune des caractéristiques propres :

- les ontologies de domaine : elles modélisent des concepts et relations propres à une discipline particulière, à un domaine du savoir. Elles sont réutilisables dans un domaine donné. Leur objectif est d'offrir une représentation structurée des connaissances d'un domaine pour que leur exploitation par différentes applications soit possible.

---

59 FOURCASSIER Eric. Op.cit. p.20

60 CHARLET Jean, BACHIMONT Bruno, TRONCY Raphaël. Ontologies pour le Web sémantique. *Revue Information, Interaction, Intelligence I3 [En ligne].* 2004. [Consulté le 25/11/2017]. Disponible à l'adresse: http://www.eurecom.fr/~troncy/Publications/Troncy-revue_i304.pdf, p.4

61 CHARLET Jean. Op.cit. p.44

- les ontologies de tâche : elles fournissent un vocabulaire des termes utilisés pour résoudre les problèmes de tâche, ces tâches peuvent être effectuées au sein d'un ou plusieurs domaines. Elles offrent un ensemble de termes pour décrire la manière de résoudre un type de problème.
- les ontologies d'application : ces ontologies applicatives contiennent suffisamment de connaissances pour structurer un domaine de connaissances particulier. Elles sont plus spécifiques et rassemblent l'ensemble des connaissances requises pour assurer le fonctionnement d'une application particulière, elles ne sont pas conçues pour être réutilisables. Elles peuvent également inclure une ontologie de domaine[62].
- les ontologies de haut niveau : aussi appelées « ontologies communes », elles présentent des concepts et relations d'un haut degré de généralité comme le temps, l'espace, la matière, l'objet, l'événement, la causalité, etc. Elles restent très abstraites et peuvent donc « coiffer » le sommet de l'architecture conceptuelle de ces ontologies de domaine[63].

Niveaux de granularité

Le degré de détail des composants de l'ontologie, aussi appelé niveau de granularité, est utilisé pour caractériser les ontologies. L'objectif opérationnel de l'ontologie va orienter ce degré de détail. Effectivement, suivant les contextes d'utilisation, il est parfois « indispensable de mobiliser une conceptualisation très détaillée ou au contraire de se concentrer sur des caractères plus généraux en évacuant des détails trop fins ». Le niveau de granularité choisi est aussi dépendant de la cible qui utilisera la ressource : des utilisateurs très spécialisés dans un domaine donné auront souvent besoin d'une ontologie plus détaillée que des usagers moins initiés[64].

Gayo Diallo propose une typologie des ontologies selon leur niveau de granularité et distingue alors deux types d'ontologies :

Ontologie à granularité fine : elle est très détaillée et permet de décrire des connaissances de manière précise. Les ontologies de domaine, de tâche ou d'application sont souvent de ce genre :

---

62 BAZIZ Mustapha, Application des Ontologies pour l'Expansion de Requêtes dans un Système de recherche d'information. Rapport de DEA Informatique de l'Image et du Langage (2IL). Université Paul Sabatier et Institut National Polytechnique de Toulouse, 2002 In
ABDERRAHIM Mohammed Alaeddine. *Exploitation des Ontologies dans les Systèmes de recherche d'information Arabes* [en ligne]. Thèse pour l'obtention du grade de Docteur en sciences. Université Aboubakr Belkaïd, Tlemcen, 2016. 94 p. [Consulté le 25/11/2017]. Disponible à l'adresse : http://dspace.univtlemcen.dz/bitstream/112/8632/1/Exploitation_des_Ontologies_dans_les_Systemes_de_recherche_dinformations_Arabes.pdf , p.21

63 FOURCASSIER Eric. Op.cit. p.23

64 FOURCASSIER Eric. Op.cit. p.25

ce niveau de détail dans la conceptualisation d'un domaine est exigé pour les ontologies très spécialisées. Elles s'adressent généralement à des utilisateurs initiés au domaine traité.

Ontologie à granularité large : elle est moins détaillée, les ontologies de haut niveau font par exemple partie de cette catégorie. Elles sont positionnées au sommet de la hiérarchie conceptuelle et doivent donc s'en tenir à des notions suffisamment larges car les notions qu'elles décrivent sont génériques[65].

1.2.2 Composants d'une ontologie

Les concepts, termes et propriétés

Il est maintenant intéressant de décrire les éléments présents dans une ontologie pour mieux comprendre sa structure. Nous pourrons voir ensuite de quelle manière ces composants varient selon la nature des fonctionnalités que l'ontologie va appuyer. Fabien Gandon a décrit de manière précise ces différents composants.

Une ontologie définit des concepts, qui sont « des constituants de la pensée, nés de l'esprit », ils peuvent être des principes, objets, idées, catégories d'objet ou notions abstraites. Ces unités sémantiques sont aussi nommées « classes » ou « catégories ». Les concepts sont représentés par des termes et sont reliés entre eux par des relations ; des règles et axiomes contraignent ces éléments.

L'ensemble des propriétés relatives à un concept est décrit comme son intention ; l'extension est constituée des objets ou êtres que ce concept englobe. Fabien Gandon cite l'exemple d'un concept anonyme nommé C et du domaine automobile pour illustrer ces notions. Dans ce cas, l'intension comprend les propriétés communes à des individus ou objets faisant partie d'un concept, ces propriétés permettent de définir le concept. Par exemple, l'intension du concept C peut être « C'est une sous-catégorie de véhicules de transport automobile, conçus et aménagés pour le transport d'un petit nombre de personnes (...) »[66]. Pour aboutir à une représentation du monde et de ses contraintes, les intensions sont organisées, structurées et contraintes. Effectivement, les propriétés d'une ontologie sont contraintes par des règles d'inférence, des règles de construction possibles ou interdites et des règles de déduction. Grâce à ces concepts,

---

[65] DIALLO Gayo. Une architecture à base d'ontologies pour la gestion unifiée des données structurées et non structurées. Thèse. Université Joseph Fourier – Grenoble I, 2006 In DRAME Khadim. *Contribution à la construction d'ontologies et à la recherche d'information : application au domaine médical [*en ligne]. Thèse pour obtenir le grade de Docteur en informatique et santé. Université de Bordeaux, 2014. 187 p. [Consulté le 20/02/2017]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-01166042/document, p.12

[66] GANDON Fabien. Op.cit.

aux propriétés et caractéristiques des concepts, aux relations entre concepts et aux règles relatives à ces relations, un SRI va pouvoir raisonner à partir d'une ontologie[67]. Par exemple, « une voiture est forcément un véhicule » ou « une voiture a quatre roues » peuvent constituer des intensions.

L'extension comprend les entités que l'on peut placer dans cette catégorie puisqu'elles respecteront l'intension du concept C. Ainsi, « la Twingo de Rose, le Kangoo d'Olivier, la 306 d'Isabelle, la Clio d'Alain… » entrent dans cette catégorie.

Dans une ontologie, les concepts et leur manifestation linguistique sont dissociés ; en effet, un concept n'est pas un terme et vice-versa. Un concept ne comprend qu'une définition, il n'a qu'un seul sens alors qu'un terme peut être ambigu. Un concept est désigné par « une représentation symbolique » qui est linguistique ou verbale. Dans les ontologies les plus formalisées, une URI est utilisée pour renseigner les concepts. Ici, les termes du concept C sont synonymes, ils peuvent être « voiture », « automobile », « auto », « véhicule automobile », ou encore « tacot », « bagnole »[68].

Les relations entre les concepts

Les relations sont les liens qui articulent les concepts entre eux. C'est notamment grâce à la richesse de ces relations que les ontologies se différencient des langages contrôlés comme les taxonomies ou thésaurus. Il existe plusieurs genre de relations, dont :

- Les relations hiérarchiques de subsomption : elles sont du genre « is a », c'est-à-dire « est un » et sont par exemple employées pour placer une espèce sous un genre.
- Les relations de partonomie : elles se différencient des relations de subsomption car ce sont des entités spécifiquement ontologiques, elles distinguent donc les ontologies des taxinomies ou thésaurus. Ce sont des relations « tout/partie », elles sont usitées pour signifier qu'il y a un rapport de composition entre deux classes. Par exemple, la classe « automobile » est composée de différents éléments comme le moteur, les roues ou encore le carburateur[69].
- Les relations inverse : ce sont des relations dont le sens peut être inversé, par exemple « fait partie » est l'inverse de « inclut ». Si une relation s'applique, elle s'appliquera de manière inverse dans l'autre sens, « si un cartilage fait partie d'une articulation, alors l'articulation inclut le cartilage, et vice-versa ».

---

67 FRONTERE Mikhail. Op.cit. p.87

68 GANDON Fabien. Op.cit.

69 FOURCASSIER Eric. Op.cit. p.28

Le concepteur de l'ontologie pourra également définir lui-même les relations au sein de son ontologie, c'est-à-dire décider des relations présentes dans le référentiel, il a donc la possibilité d'intervenir sur le langage d'indexation des ressources[70].

Une ontologie peut alors être présentée comme un réseau de concepts reliés par un nombre de relations qui peut être élevé. Dans ce réseau, des propriétés sont attribuées aux concepts et elles ont certaines valeurs; des règles permettent aux machines de faire des raisonnements[71].

Langages informatiques et niveaux de complexité

Dans le cadre du Web des données, il est essentiel que les ontologies soient formalisées grâce au même langage informatique afin de permettre une meilleure interopérabilité pour leur partage, leur modification et leur intégration dans diverses applications et pour diverses fonctionnalités. Ainsi, des langages ont fait l'objet de recommandations par le W3C pour représenter formellement des domaines de connaissances. RDFS et OWL sont considérés comme les plus adaptés car ils permettent un haut niveau d'expressivité dans les ontologies. Ils s'appuient sur le langage de balisage XML qui constitue un élément fondamental du Web sémantique.

RDFS constitue le langage le plus léger, il permet de déclarer et de décrire les types de ressources manipulées (les classes) et les types de relations entre ces classes. RDFS offre la possibilité « de définir des vocabulaires utilisés dans les graphes RDF et d'en nommer les primitives avec des URI ». On peut nommer de manière formelle les types de relations existantes entre les instances de ces classes. Grâce aux URI, l'interopérabilité des ontologies est facilitée[72].

OWL est une extension de RDFS et dépasse largement ce schéma car, grâce à ce langage plus récent, l'ontologie devient interprétable automatiquement par une machine ; la machine peut donc faire des inférences sur les connaissances de l'ontologie. Effectivement, pour le W3C, OWL a été conçu « pour représenter explicitement la signification des termes des vocabulaires [au sens de la logique des prédicats] et les relations entre ces termes ». Le vocabulaire du langage OWL est également plus riche que celui de RDFS : des relations entre classes précises ou des axiomes (contraintes sur les concepts ou les relations) sont présents.

Employer un langage informatique particulier pour formaliser une ontologie fait varier son niveau de complexité et d'expressivité, ainsi que les fonctionnalités proposées par celle-ci (description du domaine, inférences évoluées sur les concepts de l'ontologie, etc.). Choisir le langage informatique le plus adapté au niveau de complexité de la future ontologie est indispensable.

---

70 FRONTERE Mikhail. Op.cit. p.88

71 FRONTERE Mikhail. Op.cit. p.23

72 BACHIMONT Bruno et al..Op.cit. p.30

1.2.3 Apport des ontologies pour une recherche d'information améliorée

Nous allons découvrir que les ontologies, indépendamment de leur genre et leur niveau de granularité, ont de nombreux rôles à jouer au cœur du Web des données. Nous nous focaliserons ensuite sur les fonctionnalités précises qu'elles peuvent appuyer afin d'améliorer la recherche d'information.

Ontologies et Web des données

Avec l'émergence des technologies du Web sémantique et le développement du Web des données, les ontologies connaissent un véritable essor. Elles sont perçues « comme un moyen de disposer de modèles de connaissance partageables [73]» pour Nathalie Aussenac-Gilles. Nous avons également vu qu'elles permettaient de raisonner sur des connaissances. Elles peuvent ainsi être d'une grande utilité pour plusieurs points que nous allons développer. De nouveaux rôles - permis notamment grâce aux technologies du Web sémantique - apparaissent pour les ontologies.

Elles ont tout d'abord un rôle important d'aide à la communication : les ontologies peuvent tout d'abord faciliter la communication entre humains en permettant la création d'un vocabulaire standardisé (qui sera par exemple utilisé au sein d'un groupe ou d'une entreprise). Néanmoins, elles servent principalement la communication entre Hommes et machines ; dans ce cas, l'ontologie est davantage formelle.

Lors d'un échange d'informations entre deux personnes ou entre une personne et une machine, il est nécessaire que l'émetteur et le destinataire fassent appel à une conceptualisation partagée. En effet, lors d'un processus de communication, des individus ou entités doivent pouvoir faire les mêmes raisonnements sur les informations qui sont échangées afin qu'un message soit bien compris. Pour illustrer cela, Fabien Gandon prend l'exemple d'une conversation très simple : « - Tu connais un restaurant proche ?
- Il y a une pizzeria au coin de la rue.
- Merci. »
On constate que la première personne a employé le concept de restaurant dans sa requête. Pour lui répondre, son interlocuteur a alors utilisé une taxonomie de concepts sans en avoir même conscience. Autrement dit, dans sa représentation du monde, une pizzeria appartient au concept de restaurant et il en a déduit que sa réponse fournie était donc pertinente. En effet, la catégorisation (le fait d'identifier des catégories) et l'identification (le fait de déterminer si une entité appartient à une catégorie) sont des inférences élémentaires que l'on fait à longueur de

[73] AUSSENAC-GILLES Nathalie. Op.cit. p.16

journée. Cette conceptualisation du monde est la plupart du temps partagée et implicite lors d'un échange entre deux êtres humains : ici, la deuxième personne n'a pas besoin de préciser une nouvelle fois qu'une pizzeria est un restaurant puisqu'elle suppose que sa réponse est comprise comme telle.

Lors d'un échange entre une personne et une machine, une ontologie devient un outil très précieux car elle fournit une conceptualisation partagée et permet alors que la machine fasse des raisonnements sur les concepts présents. L'ontologie constitue alors une sorte de « carte mentale » des représentations de l'utilisateur que la machine va déplier et parcourir pour choisir le chemin débouchant sur la réponse la plus pertinente possible à une requête.[74].

Nous pouvons nous appuyer sur un extrait d'un autre exemple de Fabien Gandon : si une personne recherche simplement le mot « livre » dans un SRI, de nombreux résultats qui répondaient pourtant à sa requête ne seront pas renvoyés. En effet, le système recherchera la chaîne de caractères « l-i-v-r-e » et pas une autre chaîne de caractères qui pourrait aussi être pertinente comme « r-o-m-a-n ». Mais si, grâce à une ontologie, on explique quelques aspects de la réalité à la machine, comme le fait que « roman et nouvelle [soient] des sous-types de livre », le système pourra se rendre compte que le roman est un livre. Ainsi, des résultats comprenant le mot « roman » seront aussi proposés à l'utilisateur et il y aura moins de silence documentaire[75].

L'interopérabilité des modèles et les échanges de connaissances entre utilisateurs et applications informatiques sont facilités grâce aux ontologies. Dans le cas d'une communication entre deux machines, l'ontologie répertorie l'ensemble des concepts échangeables par ces machines, même si elles sont développées sur des bases hétérogènes. On parle alors d' « interopérabilité sémantique »[76].

Elles aident également à l'échange de données : les ontologies peuvent constituer des langages partagés, des formats d'échange de données et peuvent ainsi servir de « langage pivot ». En effet, dans ce cas-ci, l'ontologie définit des données ou schémas de bases de données à un niveau conceptuel et formel. Via une seule interface de recherche, l'utilisateur d'un SRI va interroger plusieurs sources d'information, sans se préoccuper du fait que les modèles de données des ressources interrogées seront peut-être différents. La formulation des requêtes est alors facilitée et davantage uniforme.

Enfin, elles aident à l'indexation et à la recherche : nous nous focaliserons sur cet aspect dans la suite de notre propos. Ici, les ontologies deviennent des ressources fournissant des métadonnées,

74 GANDON Fabien. Op.cit.

75 Ibid.

76 CHARLET Jean. Op.cit. p.48

autrement dit, elles sont des « creuset(s) de mots-clés servant à définir des métadonnées » ; le contenu des ressources sera alors caractérisé et les documents seront indexés grâce à ces ontologies. Finalement, les ontologies permettent de mieux présenter les documents répondant à une requête, ils seront plus pertinents, mieux classés et mieux ordonnés sur la page de résultats[77].

Avec l'intégration d'ontologies dans des SRI, on espère alors que la recherche d'information sera facilitée puisque : le traitement des documents sera plus efficace grâce à l'indexation sémantique ; la reformulation des requêtes visera à réduire le bruit et le silence documentaire ; enfin, la restitution des résultats de recherche se fera de manière plus visuelle. Ces points seront développés dans les pages suivantes.

Apports pour l'indexation sémantique

Avec les ontologies, on espère traiter plus précisément et efficacement les documents. En tant que formalisations explicites et partagées des concepts d'un domaine, les ontologies permettent d'améliorer le processus d'indexation et par conséquent le processus d'appariement. Effectivement, utiliser des concepts issus d'une ontologie permet de résoudre d'éventuelles ambiguïtés, cela permet aussi de trouver des liens entre la requête d'un utilisateur et des concepts présents dans les sources d'un SRI[78].

Selon l'Association française de normalisation (AFNOR), « L'indexation est un processus destiné à représenter par les éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document ou d'une question ». Il s'agit de décrire le contenu d'une ressource de manière plus ou moins formalisée pour permettre à l'utilisateur d'un SRI de rechercher et de trouver des documents répondant à son besoin d'information. L'objectif fondamental de l'indexation est donc « le signalement optimal du contenu des documents »[79].

L'indexation sémantique est un cas particulier du processus d'indexation, elle va notamment être rendue possible grâce aux ontologies. Elle se place à un niveau différent : les documents du système ne se réduisent plus à « des chaînes de caractères pondérées », on les traite à un niveau conceptuel. Ainsi, des liens sont tissés entre les mots présents dans les documents et les notions qu'ils désignent. Dans un index sémantique, la présence de ces notions (exprimées sous forme de concepts) dans chacun des textes sera stockée. Pour passer à ce « niveau d'abstraction supérieure », l'utilisation d'une ontologie est primordiale car elle fournit un modèle de notions présentes au sein des documents, ainsi que les expressions linguistiques (les termes) associées à

---

77 AUSSENAC-GILLES Nathalie. Op.cit. p.16

78 SY Mohameth-François. Op.cit p.119

79 FRONTERE Mikhail. Op.cit. p.52

ces notions. On fait donc appel à une ressource normalisée pour décrire le contenu informationnel des ressources[80].

Avec l'implantation d'une ontologie dans un SRI, on espère alors améliorer la qualité de l'indexation en regroupant les termes synonymes sous un même concept ou en distinguant des mots similaires, dont le sens est pourtant différent. Olivier Gagnon illustre ce dernier cas à l'aide de l'exemple suivant : sur une page Web, un internaute consulte un article centré sur les derniers exploits d'une équipe de *baseball*, « les Tigers de Détroit ». Cependant, un autre article suggéré par le site comme étant relié à ce dernier document se focalise sur le golfeur Tiger Woods. Une indexation syntaxique a effectivement été utilisée : la machine considère que ces deux documents sont voisins puisqu'ils contiennent tous les deux la chaîne de caractères « tiger », cela est faux mais reste difficile à évaluer pour une machine. L'indexation par ontologie résout ce problème car le contexte d'un document va être analysé pour déterminer s'il aborde plutôt le thème du *baseball* ou bien du golf. Par exemple, dans l'article centré sur les Tigers de détroit, on trouvera certainement des termes comme « terrain », « batte », « deuxième base », etc. En comparant ces mots aux termes contenus dans une ontologie du sport, le système constatera que cet article parle davantage du *baseball*. Ainsi, si une ontologie est utilisée, cet article centré sur le *baseball* aura peu de chance d'être relié à un article parlant d'un joueur de golf[81].

Aide à la formulation des requêtes

Les ontologies sont également très utiles dans l'interface de recherche afin que l'utilisateur puisse formuler plus facilement une requête, les documents retournés par le SRI seront davantage en conformité avec son besoin d'information.

Dans une interface de recherche classique, la requête est formulée en langage libre, mais nous avons découvert que ce mode d'interrogation n'était pas optimal à cause de nombreuses limites comme la polysémie ou la synonymie. Une ontologie implantée dans un SRI sert de langage pivot, plusieurs solutions sont alors envisagées pour formuler différemment des requêtes en prenant appui sur une ontologie. Les usagers peuvent formuler directement leur requête en utilisant le langage formel de l'ontologie ; par contre, ce mode de requête est destiné aux initiés connaissant déjà ce langage. Les utilisateurs peuvent aussi formuler leur requête en langage naturel, celle-ci

---

80 REYMONET Axel, THOMAS Jérôme, AUSSENAC-GILLES Nathalie. Modélisation de Ressources Termino-Ontologiques en OWL. *Journées Francophones d'Ingénierie des Connaissances*, Juillet 2007 (Grenoble, France). Cépaduès Éditions, p.169-180, 2007. Également disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-00365888 , p.3

81 GAGNON Olivier, I*ndexation de documents Web à l'aide d'ontologies,* Mémoire pour obtenir le diplôme de Maîtrise Es Sciences Appliquées (Génie informatique). École polytechnique de Montréal, 2013. [Consulté le 25/11/2017]. 98 p. Disponible à l'adresse : https://publications.polymtl.ca/1131/1/2013_OlivierGagnon.pdf , p.3

sera ensuite traduite sous forme de concepts à partir des termes de l'ontologie. Cette dernière solution semble être la plus simple pour les usagers d'un SRI[82].

Une des fonctionnalités importantes rendue possible grâce aux ontologies est également la reformulation de requêtes. L'utilisateur peut exprimer son besoin plus facilement : afin de le guider, une formulation avec des termes plus appropriés peut aussi lui être suggérée. De plus, selon Nathalie Aussenac-Gilles, la formulation écrite des requêtes par les usagers constitue une limite puisque celles-ci sont souvent courtes et non-formalisées car le langage naturel est employé. Dans certaines interfaces de recherche, ces requêtes vont alors être automatiquement étendues ou reformulées : des documents ne contenant pas exactement les termes de l'usager mais répondant à sa recherche d'information seront retournés[83].

L'objectif de la reformulation est, soit de limiter le silence documentaire, soit de limiter le risque de bruit. Dans le premier cas, un processus d'expansion de la requête va être mis en place : la recherche est élargie grâce à des termes similaires aux termes de la requête initiale. Au contraire, pour limiter le bruit, la requête est modifiée grâce à des termes qui ajoutent de l'information supplémentaire. Un plus grand nombre de contraintes s'applique donc à l'expression du besoin documentaire. Il y a donc plusieurs manières de reformuler une requête selon le résultat que l'on souhaite obtenir[84].

Finalement, les ontologies pourront améliorer les fonctions d'appariement : en utilisant la structure d'une ontologie (la manière dont les concepts sont organisés entre eux), on peut mettre en place des « mesures de similarité sémantique ». Ces mesures peuvent indiquer dans quelle mesure deux concepts sont proches en fonction de la distance qui les sépare l'un de l'autre. Par extension, il est possible de savoir si un document indexé par des concepts peut répondre à la requête d'un utilisateur. Ainsi, les ontologies peuvent permettre un appariement plus juste entre les ressources du SRI et les besoins d'information des usagers[85].

---

82 AUSSENAC-GILLES Nathalie, HERNANDEZ Nathalie, BAZIZ Mustapha. Ontologies pour la recherche d'information, importance de la dimension terminologique. In EL HADI Widad Mustafa. *Terminologie et accès à l'information*. Paris : Hermès Science Publications, 2006. 262 p. (Traité des sciences et techniques de l'information). ISBN 2-7462-1295 Chapitre également disponible en ligne à l'adresse : https://www.irit.fr/publis/IC3/Aussenac-traiteWidad_2006.pdf , p.12

83 AUSSENAC-GILLES Nathalie. Op.cit. p.17

84 HERNANDEZ Nathalie et al. RI et Ontologies – État de l'art 2008 [en ligne]. Rapport interne. IRIT, Université de Toulouse, 2008. 45 p. [Consulté le 25/11/2017]. Disponible à l'adresse : https://www.irit.fr/publis/SIG/2008_RA-14-FR_HHMR.pdf , p.30

85 SY Mohameth-François. Op.cit. p.33

Des modes variés de visualisation des résultats

Dans un SRI, la liste constitue la forme la plus commune de visualisation des documents. Chaque résultat est habituellement présenté dans un bloc contenant quelques métadonnées comme le titre ou l'auteur de la ressource, un texte peut aussi résumer brièvement le contenu du document. Dans le cas d'un SRI où une ontologie est implantée, des labels décrivant des concepts rattachés aux documents peuvent être mis en valeur dans ce bloc[86].

D'autres modes de visualisation plus novateurs sont proposés : les ontologies peuvent également être utilisées « comme guides sémantiques ». Les résultats peuvent être fournis sous forme de cartes sémantiques pour que l'utilisateur comprenne plus facilement la manière dont les ressources retournées sont liées les unes aux autres. Par exemple, les résultats pourront être disposés de telle sorte que la distance physique entre eux soit proportionnelle à leur distance sémantique au sein de l'ontologie. Ainsi, deux documents indexés avec des concepts identiques seront proches dans l'interface de recherche[87].

Une catégorisation des résultats est parfois présente : certaines ressources sont classées dans la même catégorie car elles traitent d'un concept similaire. Ainsi, l'usager peut voir facilement quels concepts sont présents dans les résultats retournés ; il peut comprendre plus facilement en quoi les documents qui lui sont proposés correspondent à sa requête. Ce mode de présentation offre une vue plus globale et claire qu'une liste présentant des résultats[88].

1.2.4 Choix au sein de l'ontologie pour des fonctionnalités spécifiques

En s'appuyant sur quelques exemples concrets où des fonctionnalités précises sont présentes, nous voyons que la structure de l'ontologie, ses relations et son niveau de formalisation doivent être bien pensés pour que l'application fonctionne de manière optimale.

Dans le cas d'une fonctionnalité d'expansion de requêtes, la nature des relations est primordiale. Ce service paraît très commode au premier abord mais il conduit aussi à prendre le risque de générer du bruit dans les résultats de la requête. Effectivement, pour élargir le périmètre de la requête, d'autres termes proches des mots utilisés par les utilisateurs peuvent être introduits dans le SI ; néanmoins, ceux-ci peuvent être inadaptés, ainsi des documents non-pertinents pour les usagers seront potentiellement retournés[89]. L'approche de Mustapha Baziz, Nathalie Aussenac-

86 SY Mohameth-François. Op.cit. p.54

87 SY Mohameth-François. Op.cit. p.34

88 SY Mohameth-François. Op.cit. p.54

89 AUSSENAC-GILLES Nathalie. Op.cit. p.121

Gilles et Mohand Boughanem consiste alors à procéder à une « expansion prudente » à l'aide de la base de données lexicales Wordnet, afin d'améliorer systématiquement les résultats ; ce processus est transparent pour l'utilisateur. En analysant les termes de la requête d'un utilisateur, des concepts Wordnet sont reconnus ; grâce à des relations sémantiques entre ces concepts, la requête est ensuite élargie. Par contre, plusieurs études ont montré que la nature des relations sémantiques avait une influence significative sur l'expansion des requêtes : il faudrait seulement exploiter les relations hyponymiques de type is-a pour obtenir des résultats satisfaisants dans ce genre de fonctionnalité. Ces relations devront seulement être exploitées sur un seul niveau autour de chaque concept pour éviter le bruit. Au contraire, les relations méronymiques (relations d'association) ou antinomiques (relations contradictoires) détérioreront la qualité des résultats pour l'expansion de requêtes. En outre, il faut minimiser le nombre de concepts utilisés pour représenter la requête pour que la fonction d'extension soit efficace[90].

À l'inverse, pour des tâches relevant davantage de l'expertise documentaire comme la classification ou l'aide à l'activité de veille, l'utilisateur est un spécialiste du domaine et/ou un professionnel de l'information-documentation. Ainsi, il sait formuler précisément ses centres d'intérêt ; le degré de couverture du domaine de connaissances dans l'ontologie peut donc être plus élevé et les concepts, plus nombreux, sans que cela génère davantage de bruit documentaire[91].

Les ontologies sont donc des spécifications formalisées d'une conceptualisation d'un domaine de connaissances, ainsi elles peuvent servir de SOC performants afin d'appuyer l'indexation et la recherche de documents. Leurs composants, et notamment des relations plus élaborées, leur permettent d'être plus formalisées, plus complexes que les autres SOC. Leurs fonctionnalités d'aide à la recherche d'information sont alors nombreuses : appui à l'indexation, à la formulation des requêtes ou encore fonctionnalité de visualisation des résultats renvoyés par le SRI. Cependant, niveau de granularité, degré de couverture du domaine et nature des relations entre les concepts doivent être mûrement réfléchis afin que l'ontologie soit la mieux adaptée aux fonctionnalités auxquelles elle se destine. Nous allons maintenant nous pencher sur la construction d'ontologies, un processus complexe et fastidieux que de nombreux chercheurs ont décrit dans la littérature scientifique. Des éléments généraux sur le cycle de vie des ontologies et sur les grands principes ont été décrits. Des approches et méthodologies hétérogènes existent

---

90 BAZIZ Mustapha, AUSSENAC-GILLES Nathalie , BOUGHANEM Mohand. Désambiguïsation et Expansion de Requêtes dans un SRI, Etude de l'apport des liens sémantiques. *Revue des Sciences et Technologies de l'Information (RSTI*) série ISI. Hermes : Paris, 8 (4) 2003, p. 113-136 In AUSSENAC-GILLES Nathalie, 2008, p.121

91 AUSSENAC-GILLES Nathalie. Op.cit. p.120

aussi : l'une sera privilégiée par rapport à l'autre en fonction des moyens mis à disposition et des spécificités de la future ontologie.

1.3 Des approches variées et des méthodologies formalisées pour la construction d'ontologies

Il existe diverses manières de construire des ontologies. Pour Fabien Gandon et Rose Dieng-Kuntz, lorsque des personnes se mettent d'accord « sur l'utilisation et la théorie spécifiée par l'ontologie », on parle « d'engagement ontologique ». « L'ingénierie ontologique », quant à elle, est davantage centrée sur les aspects pratiques puisqu'elle est focalisée sur l'étude « des méthodes, techniques et outils pour traiter les différentes phases de développement d'une ontologie[92] », c'est ce thème que nous allons maintenant aborder[93].

1.3.1 Le cycle de vie des ontologies

Pour Fabien Gandon, les ontologies sont des objets vivants, ainsi elles ont un cycle de vie composé de plusieurs étapes : détection des besoins, conception, gestion et planification, évolution, diffusion, utilisation puis évaluation. Nous allons détailler ces sept activités, menées idéalement de manière itérative et cyclique ; ces étapes constituent alors « un cercle vertueux ».

- Détection des besoins : tout d'abord, il s'agit de se demander les raisons pour lesquelles on construit l'ontologie, il faut aussi faire un état des lieux de l'existant.
- Conception : cette phase centrale englobe le choix des solutions pour créer l'ontologie, l'acquisition des connaissances (par du TAL, de l'analyse de texte, etc.) sur lesquelles s'appuieront les concepteurs, la conceptualisation et la modélisation, la formalisation puis l'implantation de l'ontologie.
- Gestion : cette activité de gestion et de planification est permanente puisque dans ce genre de projet complexe et exigeant, un travail sérieux de suivi conditionne la réussite du projet.
- Evolution : elle pose souvent des problèmes de maintenance et d'intégration technique car, quand l'ontologie change, ces changements impactent nécessairement tout ce qui a été construit au-dessus de ces modifications.
- Diffusion : on s'intéresse ici au déploiement et à la mise en place de l'ontologie. Cette phase est très contrainte par l'architecture des solutions techniques choisies (serveurs

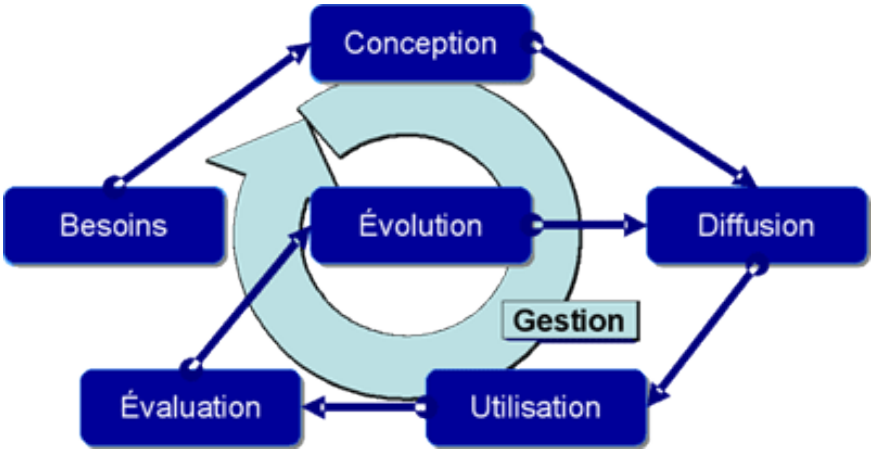
92 DRAME Khadim. Op.cit. p.22

93 GANDON Fabien, DIENG-KUNTZ Rose. Op.cit. p.6

Web, services Web, pair à pair, agents, etc.). Elle pose aussi le problème de l'ergonomie de la plateforme qui doit faciliter et encourager l'interaction des utilisateurs avec le système.

- Utilisation : elle concerne toutes les activités qui prendront appui sur l'ontologie finalisée. On peut citer l'annotation de ressources, la résolution de requêtes ou encore la déduction de connaissances.
- Évaluation : cette activité fondamentale prend appui sur des retours d'utilisation, il faut évaluer plusieurs fois.

Les différentes phases constituant le cycle de vie d'une ontologie selon Fabien Gandon sont schématisées. On retrouve les activités de développement et celles liées à la gestion de projet, comme l'évaluation continue de la qualité et l'évolution du produit[94].



Le cycle de vie d'une ontologie, GANDON Fabien, Op.cit

Natalya F. Noy et Deborah L. McGuinness approfondissent cela et mettent l'accent sur quelques règles fondamentales à suivre lors de la conception d'une ontologie. Selon elles, ce processus constitue « une simple méthodologie de génie cognitif ». Tout d'abord, il n'y a pas qu'une seule manière de modéliser un domaine de connaissances, il y a toujours plusieurs alternatives. Ainsi, la méthode choisie dépendra presque toujours de l'application que les concepteurs souhaitent mettre en place et des évolutions qu'ils anticipent sur ce système. D'autre part, le développement d'une ontologie est « nécessairement un processus itératif » pour ces deux auteures ; on ne peut pas se contenter d'appliquer successivement plusieurs étapes sans revenir sur celles-ci par la suite. Finalement, les concepts de l'ontologie doivent être semblables aux objets physiques du domaine : beaucoup de choix de modélisation lors de l'élaboration de l'ontologie seront guidés

[94] GANDON Fabien. Op.cit.

par son but d'utilisation et son degré de détail. Il faudra donc privilégier la méthode la plus intuitive, extensible et la plus adaptée à la tâche que l'ontologie appuiera[95].

Nous avons parcouru quelques éléments concernant le cycle de vie des ontologies et les principes généraux guidant leur construction. Plusieurs approches méthodologiques plus spécifiques ont été également proposées pour guider ce processus. L'approche privilégiée pour un contexte donné est fonction de l'usage auquel l'ontologie est destinée, mais également des ressources disponibles pour sa construction. Khadim Drame distingue quatre grandes catégories d'approches que nous allons développer : la construction à partir de zéro, la construction à partir de textes, la réutilisation de SOC existants puis la construction basée sur le *crowdsourcing*.

Diverses méthodologies conceptualisées par des chercheurs s'inscrivent dans chacune de ces approches. Dans les sous-parties suivantes, nous en présenterons certaines faisant référence au sein de la communauté et étant les plus citées dans la littérature.

1.3.2 Construction d'ontologies en partant de zéro

Dans les années 1990, les premières ontologies étaient construites de manière manuelle avec l'aide d'experts du domaine modélisé, les concepteurs partaient donc de zéro. À partir des années 2000, l'approche de la construction en prenant appui sur des textes a été largement utilisée. Cette approche se base sur les méthodes et outils du TAL et vise « à alléger le processus de construction d'ontologies en automatisant certaines étapes grâce aux textes ». La réutilisation de SOC déjà constitués et l'utilisation du *crowdsourcing* se développent aujourd'hui pour simplifier la construction d'ontologies[96].

Dans ces méthodologies de construction, les connaissances sont principalement formalisées par des humains, notamment des experts du domaine que l'ontologie va modéliser. Les réunions de *brainstorming* ou les interviews d'experts constituent par exemple des techniques d'acquisition de connaissances pour définir les concepts qui seront contenus dans la future ontologie ; pourtant, rien ne garantit que les connaissances décrites seront exhaustives.

Une des premières approches a été proposée en 1995 par Uschold et King, elle se base sur le développement d'*Enterprise Ontology* qui modélise les activités de l'entreprise. Elle est constituée de quatre phases : tout d'abord, la phase d'identification des objectifs et des utilisations futures de

---

95 NOY N.F., McGuinness D.L., *Développement d'une ontologie 101 : Guide pour la création de votre première ontologie,* Université de Stanford, 2005. 26p. Traduit de l'anglais par Anila Angjeli, BnF, Bureau de normalisation documentaire. [Consulté le 20/06/2017]. Disponible à l'adresse : http://www-limics.smbh.univ-paris13.fr/GBPOnto/data/documents/2010/6babahamedcontribution.pdf , p.4

96 DRAME Khadim. Op.cit. p.23

l'ontologie. Ensuite la phase de développement consiste en l'identification des concepts et relations du domaine à modéliser, puis à leur représentation explicite dans un modèle formel. L'exploitation éventuelle d'ontologies déjà existante est aussi comprise dans cette étape. L'évaluation vise à analyser l'ontologie construite pour voir si elle apporte satisfaction en vue des utilisations prévues. Enfin la phase de documentation a pour objectif de faciliter la réutilisation et le partage de la présente ontologie. Cette méthodologie comporte cependant quelques zones d'ombre. En effet, les techniques utilisées sont peu ou pas spécifiées. Par exemple, on ne sait pas comment déterminer les concepts-clés de l'ontologie, ce qui constitue un véritable frein.

Ensuite, de nombreuses méthodes sont apparues. Gruninger et Fox en ont formalisé une en 1996 dans le cadre du projet TOVE (*Toronto Virtual Enterprise*) centré lui aussi sur le domaine de l'entreprise. L'élaboration d'une ontologie doit prendre appui sur des problèmes qui se posent dans le domaine d'application, dans ce cas-ci, l'entreprise. Ces points problématiques sont formulés sous forme de questions auxquelles la future ontologie devra pouvoir répondre. Des termes seront extraits de ces questions et permettront de « spécifier la terminologie dans un langage formel ». Néanmoins, certaines limites sont toujours présentes : les étapes de construction et les techniques ne sont pas spécifiées[97].

Par ailleurs, Fabien Gandon et Rose Dieng-Kuntz présentent trois options complémentaires qui peuvent être utilisées lors de l'élaboration d'une ontologie, principalement lorsque celle-ci est créée à partir de zéro :

- L'approche ascendante : l'ontologie est construite par généralisation des concepts les plus spécifiques ; cette approche est adaptée pour la construction de ressources spécifiques.
- L'approche descendante : l'élaboration se fait par spécialisation en partant des concepts les plus généraux présents dans les couches hiérarchiques les plus élevées ; cette approche encourage la réutilisation d'ontologies.
- L'approche centrifuge : les concepteurs identifient tout d'abord les concepts centraux, ceux-ci vont ensuite être généralisés et spécifiés pour créer une ontologie complète. Grâce à cette troisième option, des sous-domaines thématiques apparaissent plus facilement au sein de l'ontologie.

Les auteurs ne pensent pas qu'il faille opposer ces trois options lors d'un projet ; au contraire, elles peuvent constituer « trois perspectives complémentaires d'une méthodologie complète ». Ces

[97] Ibid.

options peuvent être utilisées en parallèle pour un même projet afin de construire une ontologie très riche et complète[98].

La construction d'ontologies de façon manuelle étant coûteuse en temps et en ressources, d'autres approches permettant d'optimiser le processus de développement ont été mises en œuvre ces quinze dernières années ; elles permettent l'automatisation de certaines étapes. C'est le cas pour les méthodes d'élaboration d'ontologies à partir de textes, ces documents deviennent alors des sources de connaissance desquelles des concepts peuvent être extraits.

1.3.3 Construction d'ontologies à partir de textes

Avec les avancées du TAL, les méthodes et outils d'extraction de connaissances à partir de textes se sont fortement développés ; l'analyse et le traitement des documents textuels sont ainsi facilités. Des techniques linguistiques, statistiques et d'apprentissage sont souvent combinées pour extraire des connaissances ontologiques à partir de textes. La construction à l'aide de textes est principalement adaptée pour les ontologies de domaine : en constituant au préalable un corpus représentatif du champ de connaissances, on réussit à extraire certains, voire tous les constituants d'une ontologie (concepts, relations, instances, propriétés) à l'aide des outils de TAL. Ainsi, l'ontologie sera en totale adéquation avec les documents du corpus, cette ressource sera parfaitement adaptée si on souhaite indexer ces documents. En outre, les ontologies construites à partir de textes comportent souvent « une composante terminologique plus riche » que celles construites manuellement : des formulations différentes des concepts sont directement extraites des textes[99].

Terminae a été mis en place dans le début des années 2000 par le groupe de recherche Terminologie et Intelligence Artificielle du Laboratoire d'Informatique de Paris Nord (LIPN), cet outil constitue à la fois une méthodologie et un outil de construction semi-automatique d'ontologies à partir de textes. Cette approche comprend quatre phases constituant l'ensemble du processus de développement de l'ontologie : constitution du corpus (comprenant par exemple des documents techniques, de la littérature scientifique, des comptes rendus, etc.), étude linguistique (extraction des termes et de leurs relations), normalisation sémantique des termes extraits et conceptualisation (désambiguïsation des concepts et relations) puis formalisation pour préciser, compléter et valider le modèle construit lors de la conceptualisation. Effectivement, à toutes les étapes de la construction de l'ontologie, il est nécessaire de faire appel à des experts du domaine pour faire valider les documents, les candidats termes, les concepts, la bonne place des relations,

98 GANDON Fabien, DIENG-KUNTZ Rose. Op.cit. p.12

99 AUSSENAC-GILLES Nathalie. Op.cit. p.13

puis l'ontologie finale. Ainsi, un traitement humain reste indispensable pour lancer les différentes étapes, éradiquer les erreurs et enrichir l'ontologie.

En 2001, Maedche et Staab ont conceptualisé une méthode itérative pour la construction semi-automatique d'ontologies ; elle peut aussi être employée pour enrichir des ontologies déjà existantes. Cette méthodologie combine des techniques d'apprentissage automatique, des méthodes statistiques et des technologies de TAL. Elle fournit alors un ensemble d'algorithmes permettant d'extraire des concepts, attributs, relations à partir de textes. Ici, la phase de conceptualisation est automatique, l'ontologie est générée automatiquement. Il est ensuite possible de l'enrichir et de la raffiner avec la collaboration d'un expert qui ajoutera ou supprimera des concepts selon leur pertinence. L'outil de construction d'ontologies Text-To-Onto englobe cette méthode[100].

Néanmoins, ces méthodes d'élaboration peuvent seulement fonctionner si le corpus de textes est assez conséquent pour faire fonctionner les outils de TAL[101]. Des SOC déjà existants peuvent aussi être réutilisés pour alléger et simplifier le processus de construction d'une ontologie.

1.3.4 Construction basée sur la réutilisation de SOC existants

Cette approche se centre sur l'exploitation de l'ensemble ou d'une partie des informations contenues dans des SOC principalement informels pour développer ou enrichir des ontologies formelles. À la fin des années 1990, des outils ont été créés afin de permettre la réutilisation d'ontologies formelles : des concepteurs peuvent élaborer une ontologie commune de manière collaborative avec des groupes travaillant à distance. Par exemple, les utilisateurs peuvent utiliser le serveur Ontolingua afin de créer, publier et éditer des ontologies et de les réutiliser ensuite à grande échelle[102].

En 2008, Jiménez-Ruiz a testé une méthodologie basée sur la réutilisation d'ontologies. Les parties pertinentes des ontologies sources sont extraites et fusionnées avec des parties de l'ontologie cible, elles sont finalement intégrées dans cette dernière. De manière plus détaillée, pour extraire des fragments des ontologies sources, des concepts de référence déjà présents dans l'ontologie cible sont souvent utilisés et alignés aux concepts des sources. D'autres entités sont ensuite extraites afin de définir des sous-concepts. Par exemple, grâce à cette méthodologie, le thésaurus

[100] DRAME Khadim. Op.cit. p.25

[101] DRAME Khadim. Op.cit. p.26

[102] DRAME Khadim. Op.cit. p.28

NCI (National Cancer Institute) et l'ontologie médicale GALEN ont été utilisés pour aboutir à la construction d'une ontologie du domaine de l'arthrite chronique juvénile[103].

Néanmoins, selon Fabien Gandon et Rose Dieng-Kuntz, la réutilisation d'ontologies « est à la fois séduisante (…) et difficile ». Elle devrait permettre d'économiser du temps et des efforts puisque des contenus déjà créés pourraient être agrégés dans de nouvelles ontologies. Pourtant, cela est complexe puisque les conceptualisations doivent être réajustées entre l'ontologie réutilisée et celle désirée. Effectivement, les objectifs, les contextes de modélisation et d'utilisation des ontologies rendent celles-ci difficilement transposables d'un contexte à un autre. De plus, il n'est pas possible d'importer directement ou de traduire automatiquement une ontologie vers une autre, cela demande d'importants efforts de réajustement et une supervision humaine est obligatoire[104].

Une approche plus récente est apparue au début des années 2010 pour notamment réduire le temps d'élaboration d'une ontologie, elle est basée sur le *crowdsourcing.*

1.3.5 Construction basée sur du crowdsourcing

> « Le *crowdsourcing* consiste à externaliser des tâches traditionnellement effectuées par un agent désigné (comme un employé ou un entrepreneur) en faisant appel à l'intelligence et au savoir-faire d'un grand nombre de personnes ».

Nous avons vu que le processus de construction des ontologies était fastidieux et demandait énormément de temps ; ainsi il est judicieux d'impliquer un large groupe d'utilisateurs pour simplifier cette tâche de conception, surtout quand il s'agit d'ontologies importantes et complexes. Des chercheurs ont alors proposé des méthodes pour que des utilisateurs réalisent diverses tâches permettant le développement d'ontologies, telles que l'évaluation de la qualité de la ressource ou la production de nouvelles ontologies.

Getman et Karasiuk ont proposé une méthode reposant sur du *crowdsourcing* pour permettre la construction d'une ontologie dans le domaine du droit. Vingt étudiants en droit ont travaillé sur une ontologie durant un semestre, chaque utilisateur était chargé d'un domaine pour que la tâche soit simplifiée. À la fin du semestre, l'ontologie réalisée a été évaluée : les chercheurs ont estimé que le niveau de couverture des concepts s'élevait à 90 %. Ces résultats sont concluants même si quelques problèmes ont été soulevés. Les branches de l'ontologie élaborées par des usagers différents sont peu connectées entre elles, ou des concepts synonymes ne sont pas reliés. Les auteurs en concluent que le *crowdsourcing* est bénéfique pour la construction et l'enrichissement

[103] DRAME Khadim. Op.cit. p.30

[104] GANDON Fabien, DIENG-KUNTZ Rose. Op.cit. p.9

de ressources, même si une phase de « polissage » par des professionnels est nécessaire à la fin du projet.

Cette technique a fait ses preuves et peut être complémentaire des trois autres approches plus classiques : construction d'ontologies dans leur intégralité, à l'aide d'un corpus de textes ou encore réutilisation de RTO, même si cette dernière approche semble quelque peu difficile à mettre en place[105].

Afin que le Web des données se développe, des technologies du Web sémantique sont utilisées pour que des machines puissent raisonner sur les informations contenues dans des documents, celles-ci ne sont plus seulement des suites de caractères dénuées de sens. Dans ce contexte, les ontologies, des spécifications formalisées de domaines de connaissances, sont amenées à jouer un rôle de plus en plus important. Ce sont en effet les SOC les plus complexes, elles sont formalisées grâce à des langages du Web sémantique et appuient un grand nombre de fonctionnalités d'aide à la recherche d'information : aide à l'indexation, à la formulation et à l'expansion de requêtes, présentation améliorée des résultats pour l'utilisateur. Cependant, pour Nathalie Aussenac-Gilles, l'intérêt d'utiliser des ontologies pour améliorer la recherche d'information « n'est ni immédiat, ni systématique ». Les concepteurs doivent au préalable dégager les exigences de la ou des fonctionnalités qui seront appuyées par une ontologie[106]. Une réflexion approfondie sur les usages et attentes des personnes qui utiliseront l'ontologie est alors centrale et nécessaire.

Dans la littérature, de nombreuses et diverses méthodologies théorisées par des chercheurs régissent la construction d'ontologies. Elles seront choisies en fonction du temps de travail qui peut être consacré à cette lourde tâche, des ressources humaines disponibles et aussi en fonction du niveau d'exigence et de complexité que l'on souhaite donner à l'ontologie. Néanmoins, en échangeant avec des professionnels dont l'une des missions était la construction d'ontologies pour aider à la recherche d'information, nous nous sommes rendu compte qu'ils ne connaissaient pas l'existence de ces méthodologies pourtant dites « de référence », qui n'étaient pas systématiquement utilisées dans un cadre professionnel. L'approche de ces professionnels était très centrée sur les contenus qui seraient valorisés grâce à l'ontologie et sur les usages futurs des publics cibles. Ainsi, nous pouvons nous demander s'il n'y a pas un différentiel, voire une césure, entre les pratiques et méthodologies préconisées de manière théorique et les pratiques effectives dans un milieu professionnel.

[105] DRAME Khadim. Op.cit. p.28

[106] AUSSENAC-GILLES Nathalie. Op.cit. p.28

Nous avons mis en place une expérimentation qui nous permettra de répondre à cette supposition et aux hypothèses que nous détaillerons par la suite. Dans une seconde partie, nous énoncerons alors les objectifs de notre expérience et présenterons en détail notre expérimentation.

2. L'expérience des concepteurs d'ontologies : présentation de l'enquête

2.1 Objectifs de l'enquête et hypothèses

Nous avons donc souhaité explorer le différentiel qui pourrait éventuellement exister entre théorie et pratiques professionnelles lors du développement et du peuplement d'ontologies conçues dans le but de faciliter la recherche d'information.

Il paraît également indispensable de s'intéresser à la prise en compte des usages avant de modéliser la ressource. En effet, une ontologie peut avoir de multiples utilités comme l'aide à l'indexation, la formulation-reformulation de requêtes ou encore la visualisation des résultats. Nous supposons donc que dans le cadre des projets que nous avons choisis et qui s'inscrivent dans différents domaines, les ontologies auront des finalités diversifiées. Il est alors indispensable que l'ontologie - et donc sa méthodologie de construction - s'adaptent en amont à ces différents usages, mais nous pouvons nous demander dans quelle mesure cet impératif est respecté.

Ainsi, nous avons dégagé trois hypothèses auxquelles l'analyse des six entretiens que nous avons menés va nous permettre de répondre :

- Les méthodologies de construction d'ontologies conceptualisées dans la littérature ne sont plus vraiment utilisées, même si des éléments méthodologiques sont repris dans les projets.

Nous supposons que les personnes élaborant des ontologies de nos jours prennent peu, plus ou pas du tout appui sur des méthodes théorisées dans la littérature scientifique. Néanmoins, des étapes ou éléments généraux présents dans la littérature pourraient être repris lors de l'élaboration de ladite ressource. Des concepteurs pourraient aussi prendre appui sur une/des méthodologies spécifiques et s'en éloigner de différentes manières (omission ou ajout d'étapes, choix d'un autre logiciel que celui préconisé, niveaux différents de formalisme ou de couverture d'un domaine, etc.)

- Les méthodes utilisées aujourd'hui sont moins formalisées et plus itératives pour que les ontologies soient davantage adaptées aux usages des utilisateurs futurs.

Nous pensons qu'en 2018, les concepteurs prennent davantage en compte les usages pour concevoir des ontologies les mieux adaptées aux utilisateurs finaux. Ainsi, par rapport à celles du début des années 2000, les méthodes suivies aujourd'hui pourraient être moins linéaires et les tâtonnements, plus nombreux. Effectivement, à différentes étapes de la construction, les

concepteurs se repositionneraient continuellement sur les usages pour ne pas en dévier, afin de créer un modèle réellement orienté utilisateurs.

- Il y a une différence de points de vue et de conceptions entre les chercheurs et « les professionnels » (informaticiens, documentalistes…), leur but n'est pas le même lors de la construction d'une ontologie.

Nous faisons l'hypothèse que les objectifs des chercheurs et des professionnels sont parfois différents lors de la construction d'une ontologie. Les chercheurs souhaitent peut-être mettre en place un modèle exhaustif et très formel ; ils s'appuieraient davantage sur des méthodologies de construction dites « de référence ». Les professionnels rechercheraient alors une source plus simple, adaptable et rapide à construire dans un cadre industriel. Ils se baseraient moins sur la littérature scientifique pour définir leur méthode de construction. Dans le cas d'une collaboration entre chercheurs et professionnels, il pourrait donc y avoir décalage entre l'ontologie théorique, préconisée par des chercheurs, et l'ontologie effectivement implantée dans un système concret.

2.2 L'échantillon : des concepteurs d'ontologies diverses

Lors de la construction de l'échantillon, nous nous sommes fixée un premier impératif : les personnes que nous allions interroger devaient toutes avoir participé à la conception d'une ontologie pour aider à la recherche d'information. En effet, nous souhaitions qu'elles aient été placées au cœur du projet et qu'elles puissent ainsi nous donner des éléments précis sur la méthodologie de construction utilisée, le niveau de granularité de la ressource, les logiciels utilisés, etc.

Nous avons également tenté de trouver un équilibre entre différents corps de métier : nous voulions entendre les points de vue de professionnels et de chercheurs puisqu'une de nos hypothèses oppose leurs manières de faire. Au sein des professionnels, nous avons interrogé une documentaliste et des informaticiens pour comparer leurs réponses et voir si leurs professions avaient une incidence sur les points qu'ils mettaient chacun en relief.

Grâce au stage que nous avons effectué au sein du service Banque de Contenus (BDC) à Ouest-France, nous avons eu l'opportunité de faire la connaissance de plusieurs professionnels ayant participé à l'élaboration d'ontologies. Cette proximité a donc été un atout pour la prise de contact et pour le déroulé des entretiens, qui ont eu lieu directement sur notre lieu de stage. Nous avons également interrogé un chercheur qui avait collaboré avec les personnes de ce service sur une des ontologies.

Cependant, il était intéressant de ne pas se limiter à un seul contexte, en l'occurrence celui de la presse, et d'explorer d'autres domaines. De cette manière, nous pouvions comparer l'expérience et le ressenti de professionnels dans d'autres structures. Nous voulions aussi savoir si la nature du domaine à modéliser dans l'ontologie influait sur la manière de concevoir cette ressource. Lors de nos lectures et de notre veille, sur le site de l'événement SemWeb.Pro[107] notamment, nous avions pris connaissance de projets innovants faisant appel à des ontologies. Grâce à notre réseau ou en envoyant directement des *e-mails*[108], nous avons pu entrer en contact avec ces personnes. Ces messages comprenaient des informations sur notre sujet de mémoire et nos questionnements. Nous leur apprenions également que nous souhaitions mener des entretiens dans ce cadre. Nous avons ensuite envoyé des *e-mails* plus brefs aux personnes ayant répondu positivement pour se mettre d'accord sur la date et l'heure d'entretien.

Il faut ajouter que pour les projets dans lesquels les concepteurs d'ontologies se sont impliqués, les questions de conception et de modélisation d'ontologies devaient être primordiales. Effectivement, cette question est centrale dans cet écrit et nous devions avoir une matière importante à analyser sur ce sujet afin de répondre à nos hypothèses de départ. Dans le cadre des ontologies construites à Ouest-France, nous savions déjà que cela avait été le cas. Pour les autres projets, lors d'un échange d'*e-mails* et d'une conversation téléphonique préparatoire à l'entretien avec un chercheur, nous nous sommes assurée que ces questions avaient occupé une place centrale.

Au total, nous avons sollicité neuf personnes, six ont répondu positivement. Une n'a pas répondu, l'autre manquait de temps pour un entretien, la dernière a pu nous renvoyer vers un collègue que nous avions déjà contacté et qui avait davantage collaboré sur un projet précis.

Nous allons maintenant partir des projets faisant appel aux ontologies et présenter les personnes de notre échantillon qui ont participé à la conception de ces ressources.

2.2.1 Domaine de la presse : ICODA, ontologie Socle et projet DataMaritime

Ouest-France est un quotidien régional français vendu dans les régions de l'ouest de la France, 53 éditions locales sont éditées. C'est le premier quotidien français avec un peu moins de 700 000 exemplaires diffusés chaque jour. Ce quotidien fait partie du groupe SIPA-Ouest-France également propriétaire, entre autres, de quotidiens régionaux comme *La Presse de La Manche* et les trois

[107] http://semWeb.pro/, journée de rencontre annuelle réunissant des professionnels du domaine du Web sémantique

[108] Cf. Annexe n° 1 , p.98 : *E-mail* de prise de contact envoyé à notre échantillon

Journaux de Loire (*Le Courrier de l'Ouest, Le Maine-Libre, Presse Océan*). Le groupe détient aussi les titres spécialisés *Le Marin* ainsi que *Voile et voiliers*.

Au sein du siège social de *Ouest-France* à Rennes, le service informatique Banque de Contenus (BDC) est implanté, il est dirigé par Michel Le Nouy. Dans ce service, des informaticiens, *data scientists* et une documentaliste développent notamment « Troove ». Cette banque de contenus du groupe SIPA-Ouest-France contient et valorise les articles des journaux du groupe ; plus de 36 300 000 articles y sont disponibles[109].

Un annuaire comprenant des lieux, personnes et sociétés est aussi créé et mis à jour par ce service. Les entités présentes dans l'annuaire sont mises en valeur dans les articles de Troove. En effet, les utilisateurs peuvent sélectionner un lieu, une personne ou une société ; ils auront ensuite accès à une visualisation dans laquelle l'entité sera rapprochée d'autres instances (selon le nombre de cooccurrences entre ces entités au sein des articles de la banque de contenus). Par exemple, on constate que Jacques Auxiette est souvent cité dans les articles où La Roche-sur-Yon apparaît également[110]. En effet, il a été maire de cette commune de 1977 à 2004.

Pour valoriser davantage ces entités, et donc les contenus du groupe, une ontologie de haut niveau, l'ontologie « Socle » a été pensée et élaborée au sein du service BDC. En prenant appui sur l'architecture « lieu/personne/société » déjà présente puis en créant, enrichissant et détaillant d'autres branches, l'annuaire sera à terme enrichi et les articles seront catégorisés plus finement.

Le service BDC a travaillé en collaboration avec des chercheurs de l'INRIA pour modéliser cette ontologie dans le cadre du projet de recherche ICODA. Ce projet prend la forme d'un partenariat entre l'INRIA, l'Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), Ouest-France, le programme *Les décodeurs du Monde*[111] et l'Agence France-Presse (AFP). Les collaborateurs étudient principalement la question des *fake news*, *du data journalism* et conçoivent des outils pour que les journalistes puissent vérifier facilement la fiabilité de leurs sources et la véracité des informations.

Des chercheurs du Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) de l'INRIA sont membres du projet ICODA et ont centré leurs recherches sur la vie politique locale et plus précisément sur la démission des élus municipaux. Ils ont alors élaboré l'ontologie Socle avec le service BDC de Ouest-France en prenant appui sur ce cas d'usage.

---

109 Chiffre datant du 09/07/2018

110 Cf . Annexe n °2, p.100 : Visualisation de l'entité « La Roche-sur-Yon » sur Troove

111 https://www.lemonde.fr/les-decodeurs/, « Les décodeurs du Monde.fr vérifient déclarations, assertions et rumeurs en tous genres », les journalistes répondent aux questions des internautes et replacent l'information dans son contexte

Par exemple, grâce à l'ontologie, il sera possible de répondre à des questions telles que « Quels conseillers municipaux ont démissionné de 2014 à 2017 en Vendée ? ». Les données du groupe, dont se servent les chercheurs, seront plus facilement extraites, classifiées et interrogées. Ainsi, les ontologies de Ouest-France et du LIRMM ont des bases communes, mais celle du LIRMM ne répond pas aux mêmes objectifs.

Le service BDC concentre aussi ses efforts pour valoriser des contenus de domaines plus spécifiques. Le projet DataMaritime vise à valoriser les données centrées spécifiquement sur le domaine du maritime, données provenant principalement du *Marin*, hebdomadaire du groupe de presse dont les contenus sont lisibles sur Troove. Ensuite, l'objectif était de constituer une interface qui permette d'accéder quotidiennement aux contenus centrés exclusivement sur ce domaine. Le potentiel client de DataMaritime était le Cluster Maritime Français, une organisation conçue pour rassembler tous les acteurs du maritime. Un de ses objectifs est de « créer des synergies entre acteurs du maritime » pour que l'économie du secteur puisse bénéficier des innovations de tous. Les visualisations de DataMaritime permettent notamment de voir quelles entreprises sont en relation les unes avec les autres. L'objectif de l'ontologie était d'optimiser l'organisation des données.

Marie-Paule Cochet

Marie-Paule Cochet travaille à Ouest-France depuis 20 ans, elle est chef de projet au sein du service BDC depuis sa création en 2014. Elle joue un rôle d'intermédiaire entre certains journalistes, les autres services informatiques de Ouest-France, le service documentation et le service BDC afin que la banque de contenus et l'annuaire soient les plus intègres et complets possible. Marie-Paule Cochet a suivi une formation en informatique de gestion dans une école d'ingénieurs. Aujourd'hui, elle occupe un poste moins technique qu'au début de sa carrière, celui-ci est axé sur la valorisation des contenus. Dans ce cadre, elle a participé à la modélisation de l'ontologie Socle.

Henri-Maxime Suchier

Data scientist, ses missions sont centrées sur l'extraction, la valorisation et la vérification de la qualité des données. Il travaille spécifiquement sur les questions liées au Web sémantique. Par exemple, il modélise un annuaire sous forme sémantique, en RDF, et va le peupler avec les contenus issus de Ouest-France mais aussi issus d'autres sources comme Wikidata. Doté d'un doctorat en informatique, il s'est tout d'abord dirigé vers le milieu de l'Enseignement supérieur et de la recherche pour ensuite rejoindre des sociétés privées. Henri-Maxime Suchier a lui aussi

participé à l'élaboration de l'ontologie Socle, il était notamment chargé des questions plus techniques comme l'intégration des données dans l'ontologie.

Nadia Fafi

Après un master 2 en reprise d'études en documentation et un stage de fin de master dans le service BDC, elle est devenue architecte des données à Ouest-France. Elle valorise les contenus du groupe. Elle a notamment travaillé sur les catégories servant à classer les articles dans Troove. Elle a créé de nombreux corpus de documents sur Troove pour un algorithme d'apprentissage qui classe ensuite les articles automatiquement. Nadia Fafi participe aussi à la modélisation de l'ontologie Socle. Cependant, nous l'avons plus spécifiquement interrogée sur l'ontologie du domaine maritime qu'elle a élaborée pour le projet DataMaritime.

Michel Chein

Il est chercheur en informatique à l'Université de Montpellier et membre de l'équipe *Graphs for Inferences on Knowledge* (GraphIK) du LIRMM. Ses sujets de réflexion sont, entre autres, le raisonnement à partir de graphes, la représentation de connaissances et l'identification d'entités nommées. Il fait partie du projet ICODA et collabore avec les professionnels du service BDC à Ouest-France pour créer l'ontologie Socle. Avec d'autres chercheurs, il s'est basé sur le cas d'usage des démissions des conseillers municipaux pour réfléchir sur cette ontologie.

2.2.2 Domaine médical : projet LERUDI

Dans le domaine médical, et notamment dans celui des urgences, le repérage et l'analyse instantanés d'informations sont vitaux pour le patient. Le projet LEcture Rapide en Urgence du Dossier Informatisé du patient (LERUDI)[112], qui a débuté début 2009, répond à cet objectif : un prototype de moteur de fouille de textes a été élaboré et évalué. Il est capable d'extraire presque instantanément des informations utiles du Dossier Médical Partagé[113] (DMP) informatisé du patient, dossier plus ou moins structuré. L'application sera utilisée par des médecins urgentistes. Dans ce projet, l'ontologie est implantée au cœur du système de TAL qui sert à indexer les contenus du DMP.

Jean Charlet

Jean Charlet est chargé de mission recherche à l'Assistance publique-Hôpitaux de Paris et membre du Laboratoire d'Informatique Médicale et Ingénierie des Connaissances pour l'e-santé (LIMICS). Il

[112] http://esante.gouv.fr/actus/services/le-projet-lerudi-fiche-signaletique

[113] Anciennement Dossier Médical Personnel

centre ses recherches sur les systèmes d'aide à la décision et la représentation des connaissances en santé. Il est expert dans la construction d'ontologies du domaine médical et a dirigé une dizaine de ces projets, dont le projet LERUDI.

2.2.3 Domaine industriel : projet Open Food System

Le projet Open Food System (OFS)[114] réunit 25 partenaires. Ce grand projet de recherche et de développement centré sur la cuisine numérique réunit filière agroalimentaire, électroménager et services numériques. Il a débuté en 2012 et s'est terminé quatre ans plus tard. Au sein de ce projet, on trouve le programme *Nos recettes* dont le groupe SEB est le chef-de-file. L'objectif de ce programme est de proposer des recettes digitales compréhensibles par des appareils de cuisine connectés (comme le Cookéo). Ainsi, l'expérience des utilisateurs pourra être simplifiée et enrichie. Les recettes doivent être structurées de manière lisible, dans un format de données interopérable, c'est une des utilités de l'ontologie élaborée dans ce cadre. Aujourd'hui, le projet est devenu *Cooking recipe*, les travaux de recherche sont en train d'être industrialisés.

Florence Amardeilh

Travaillant aujourd'hui pour le projet *Ticket For Change*, elle a auparavant été Directrice Recherche et Développement chez Mondeca après avoir réalisé, au sein de cette même structure, une thèse combinant les outils du TAL et du Web sémantique. Mondeca est une entreprise élaborant des solutions pour gérer des référentiels, des connaissances et des outils faisant appel aux technologies du Web sémantique. Pour le projet OFS, Florence Amardeilh a mis en place le partenariat avec SEB puis a participé à la création de l'ontologie : modélisation, collaboration avec les partenaires, opérationnalisation…Elle a donc un profil hybride puisqu'elle vient du monde de la recherche et une de ses missions principales était de veiller à ce que l'ontologie puisse être implantée dans une application industrielle.

2.3 Grille d'entretien et déroulé des entretiens semi-directifs

2.3.1 Le choix des entretiens semi-directifs

Nous avons choisi de mener des entretiens semi-directifs pour laisser de la liberté à nos interlocuteurs, tout en nous laissant la possibilité de cadrer leur propos. Effectivement, faire remplir un questionnaire à notre échantillon n'était pas du tout adapté : les projets étaient trop divers les uns des autres pour qu'un seul modèle de questionnaire s'applique. De plus, une courte

[114] http://www.openfoodsystem.fr/home

réponse par écrit n'était pas suffisante pour saisir le point de vue complet des personnes interrogées et leur cheminement intellectuel.

Des entretiens individuels étaient donc plus judicieux. Nous aurions pu choisir de mener un entretien collectif avec les trois personnes du service BDC ayant participé à la modélisation de l'ontologie Socle. Pourtant, après réflexion, ce n'était pas le genre d'expérimentation le plus efficace. Une ou deux personnes auraient pu monopoliser la parole et la confisquer aux autres ; cela aurait aussi pu contraindre l'expression de certains points de vue du fait de la présence d'autres interlocuteurs.

Il nous a semblé nécessaire de cadrer les entretiens grâce à une grille reprenant les grandes thématiques et les principaux points à aborder lors des rencontres. La plupart des questions sont assez généralistes pour s'adapter aux différents projets (« Comment avez-vous construit l'ontologie ? », « Quelle est l'utilité de cette ontologie dans votre projet ?, etc.). Parfois, les questions étaient posées comme telles, parfois, nous relancions simplement les personnes ; la discussion était plus fluide de cette manière. Utiliser cette grille n'a pas été trop restrictif, ni contraignant. Par exemple, des personnes ont orienté certaines parties de l'entretien vers des sujets plus techniques. Même si ces points n'étaient pas mentionnés dans la grille d'entretien, nous pouvions relancer la discussion sur ces sujets s'ils éclairaient nos questions principales.

2.3.2 Conception de la grille d'entretien

La grille d'entretien finale[115] est subdivisée en quatre grandes catégories que nous avons organisées en fonction de leur niveau de spécificité.

Tout d'abord, assez classiquement, on trouve quelques questions concernant la formation et le parcours professionnel des personnes interrogées, pour les mettre en confiance et pour obtenir des informations complémentaires sur ces sujets. Effectivement, avant les entretiens, nous connaissions déjà en partie le cursus de notre échantillon puisque nous avions consulté leur profil sur les réseaux sociaux professionnels comme Viadéo ou Linkedin. Suite à des discussions avec les professionnels du service BDC, nous avions aussi appris quelques éléments.

La deuxième catégorie centrée sur « le projet faisant appel à une ontologie » permet d'entrer dans le cœur de l'entretien. Ici, le projet dans son ensemble est présenté, on demande aussi au concepteur de l'ontologie de parler de son rôle au sein de celui-ci. Ainsi, on peut se rendre compte de son niveau de responsabilité et des domaines sur lesquels il s'est davantage centré : dans la suite de l'entretien, on sait plus précisément quelles questions complémentaires poser. Avec le

[115] Cf. Annexe n° 4, p.103 : Grille d'entretien finale

point « L'ontologie au cœur du projet pour aider à la recherche d'information : rôle, utilité », on souhaite obtenir des informations contextuelles sur l'utilisation des ontologies. Elles peuvent avoir de multiples utilités pour aider à la recherche d'information, tout dépend du projet dans lequel elles sont implantées. Par conséquent, les méthodologies de construction pourraient elles aussi être multiples.

Le point où nous demandions le nom et la profession de la personne à l'origine de la décision d'utiliser une ontologie était également important. Nous voulions savoir qui menait en quelque sorte ce projet ; en effet, la personne à l'origine de la décision occupait probablement un rôle moteur. Il était intéressant de comprendre, qui des professionnels ou des chercheurs, avait pris l'initiative, pour déceler la présence d'un éventuel « rapport de force » entre ces professions dès le début de l'entretien.

La troisième thématique est la plus importante en termes de points à aborder et de contenus à obtenir de la part de notre échantillon. Elle est axée sur le point-clé de notre sujet : la construction et le peuplement de l'ontologie. En questionnant les concepteurs sur la manière dont ils avaient construit l'ontologie, nous voulions savoir s'ils abordaient d'eux-mêmes la question des méthodologies précises employées. Nous souhaitions aussi savoir si des étapes avaient été respectées pour pouvoir éventuellement les comparer à celles préconisées dans la littérature scientifique. Cela pouvait apporter des réponses à notre première hypothèse : « Les méthodologies de construction d'ontologies conceptualisées dans la littérature ne sont plus vraiment utilisées, même si des éléments méthodologiques sont repris dans les projets ».

Dans cette partie, nous avons aussi insisté sur le lien entre usages et conception de l'ontologie : nous leur avons demandé si l'interface dans laquelle sera implantée l'ontologie avait déjà été pensée, si les concepteurs avaient travaillé avec leurs futurs usagers dans le cadre du projet. Nous avons aussi abordé la question des possibles itérations lors des différentes phases de construction. Ainsi, nous souhaitions vérifier notre deuxième hypothèse : « Les méthodes utilisées aujourd'hui sont moins formalisées et plus itératives pour que les ontologies soient davantage adaptées aux usages des utilisateurs futurs. »

Enfin, dans une dernière thématique, nous nous sommes davantage centrée sur notre troisième hypothèse concernant les éventuelles différences de points de vue et d'objectifs des chercheurs et des professionnels. Il était intéressant d'aborder les points relatifs à la collaboration entre diverses professions : comment la collaboration avait été organisée ? Y avait-il eu quelques points bloquants lors de cette phase ? Nous avons aussi demandé si les concepteurs avaient choisi de suivre les directives ou remarques de leurs collaborateurs et quels choix ils avaient privilégiés. De

plus, en comparant les réponses données tout au long des entretiens par les professionnels et les chercheurs, nous avons pu analyser cette possible différence de conception concernant l'élaboration et l'utilité des ontologies.

2.3.3 Modification de la grille d'entretien

Dans la partie précédente, nous avons présenté la grille d'entretien dans sa version finale, néanmoins, elle a connu plusieurs modifications que nous avons effectuées après avoir constaté des dysfonctionnements dans la formulation ou le contenu des questions lors du premier et deuxième entretien[116].

Certaines des questions étaient trop générales et leur formulation, trop floue. Par exemple, nous voulions partir de formulations assez larges comme « Pour vous, une ontologie c'est ?.. », « Quand je vous dis "niveau de formalisme", qu'est-ce que ça évoque pour vous ? » pour laisser une plus large marge de manœuvre à nos interlocuteurs dans leur réponse. Mais cette manière d'aborder ces points était trop vague, les concepteurs ne comprenaient pas forcément l'utilité de ces questions et restaient flous. Nous avons donc recentré quelques questions sur les projets précis pour que ces questions deviennent plus concrètes, par exemple : « À quoi sert l'ontologie dans votre projet ? ». Puis nous avons noté si nos interlocuteurs faisaient référence au niveau de formalisme de l'ontologie dans leur discours.

Après deux entretiens, nous nous sommes rendu compte que les personnes interrogées parlaient spontanément de SOC (thésaurus, ontologies…) qu'elles avaient réutilisés afin de construire leur présente ontologie. Nous avons alors noté dans notre grille d'entretien ce point précis puisque les autres concepteurs pouvaient aussi l'aborder. Il était ensuite intéressant de comparer leur expérience, notamment le choix de ces SOC et les objectifs de la réutilisation de ceux-ci.

2.3.4 Déroulé des entretiens et retours

Les six entretiens se sont déroulés entre le 31 mai et le 11 juillet 2018. Nous avons interrogé les trois professionnels du service BDC dans les locaux de Ouest-France dans le cadre de notre stage. Nous avons mené les autres entretiens par visioconférence via Skype. Dès le début de la discussion, nous avons demandé aux interlocuteurs si nous pouvions les enregistrer pour ensuite retranscrire leurs propos, ils ont tous accepté. Il était plus pertinent de retranscrire les entretiens dans leur intégralité pour resituer les réponses des personnes dans leur contexte. En outre, il était

[116] Cf. Annexe n° 3, p.101 : Première version de la grille d'entretien (au 02/05/2018)

ensuite plus facile de réutiliser leurs citations dans la troisième partie de ce document[117]. Les entretiens ont duré entre trente minutes et une heure.

Lors des entretiens, nous craignons qu'il y ait eu parfois quelques biais. Par exemple, nous étions nous-même placée au cœur du service BDC lors de notre stage. Ainsi, nous connaissions déjà certaines problématiques auxquelles les professionnels avaient dû faire face : peut-être avons-nous dirigé les entretiens vers ces points précis que nous savions pertinents pour nourrir notre sujet ? À l'issue de notre premier rendez-vous téléphonique avec Jean Charlet au cours duquel nous lui avions présenté notre sujet de mémoire, il nous avait transmis quelques documents de présentation sur le projet LERUDI. Nous connaissions alors plusieurs éléments précis et techniques sur la construction de l'ontologie et avons pu orienter nos questions sur ces points. Ce n'était pas le cas pour le projet OFS : nous nous étions évidemment documentée sur le programme mais nous ne possédions pas d'élément aussi précis. Cela peut amener une disparité dans le niveau de précision des éléments abordés au cours de l'entretien.

Ces informations peuvent peut-être introduire des biais mais elles peuvent aussi constituer une force : nous connaissions mieux les projets et savions quels étaient les points intéressants à creuser. Nous pouvions alors aller plus en profondeur dans notre réflexion et poser des questions plus précises dès le début de l'entretien.

À posteriori, nous pensons qu'il aurait été intéressant d'entendre l'expérience d'autres personnes sur les projets LERUDI et OFS pour pouvoir comparer leurs propos et avoir des visions différentes. Cependant, Jean Charlet a surtout collaboré avec des médecins et Florence Amardeilh a notamment travaillé avec une chercheuse, ces personnes ont malheureusement un emploi du temps très chargé. Nous avons contacté la chercheuse mais elle n'était pas disponible pour un entretien.

[117] Cf. Annexes n° 5-10, p.105-154 : retranscription des six entretiens

2.4 Premiers constats

Après les entretiens, nous avons pu dégager des premières constatations que nous analyserons avec plus de précisions dans la troisième et dernière partie.

Peu de personnes interrogées citent des méthodologies de construction officielles, par contre, on peut noter que plusieurs grandes étapes assez générales sont communes à nombre de projets. Certaines méthodologies sont encore expérimentales et se construisent au fil des projets menés par les structures.

Les itérations sont omniprésentes lors des différentes phases de construction, les méthodologies ne sont pas linéaires. Un poids important est également donné à l'évaluation de la ressource.

La question de la prise en compte des usages pour l'élaboration de l'ontologie est centrale et de plus en plus prégnante. Les concepteurs insistent beaucoup sur ce point.

La collaboration entre personnes qui n'ont pas forcément la même profession est omniprésente. Cela est nécessaire et constitue un atout puisque chacun peut apporter son expertise. Il existe une réelle interdépendance entre ces différents acteurs. Elle génère néanmoins des contraintes : nécessité de formation, objectifs différents, souhait d'un niveau de granularité plus ou moins large selon les interlocuteurs, etc.

Pour finir, la frontière entre professionnels et chercheurs est assez tenue : lors de la construction d'ontologies, il n'y a pas forcément de vision conditionnée par la profession, mais plutôt des approches dépendant toujours du projet et de son contexte.

3. Plus de liberté et de diversité dans la conception d'ontologies : analyse des entretiens

Enfin, nous allons reprendre point par point nos trois hypothèses et voir si elles sont vérifiées. Pour cela, nous allons nous appuyer sur les entretiens que nous avons menés auprès des concepteurs d'ontologies, nous les avons intégralement retranscrits. Puis nous avons dégagé des grandes thématiques transverses et des points plus particuliers qui ont attiré notre attention. Nous avons ensuite élaboré un tableau avec ces thèmes principaux et l'avons complété avec les citations retranscrites. Ce tableau nous a servi de trame pour rédiger cette troisième et dernière partie.

Après avoir apporté une réponse à nos trois hypothèses de départ, nous pourrons conclure sur cette question principale : existe-t-il réellement une césure entre théorie et pratiques professionnelles lorsque l'on élabore des ontologies afin de faciliter la recherche d'information ?

3.1 Les méthodologies de construction d'ontologies conceptualisées dans la littérature ne sont plus vraiment utilisées, même si des éléments méthodologiques sont repris dans les projets.

3.1.1 Des méthodologies officielles peu utilisées

Pas de connaissance/pas d'intérêt pour ces méthodologies

Tout d'abord, nous constatons que de nombreuses personnes ne connaissent pas et n'utilisent pas de méthodologies officielles lors de l'élaboration d'ontologies. Ainsi, ces méthodes très cadrées et précises ne font pas référence dans certaines structures professionnelles.

En effet, les professionnels travaillant au sein du service BDC de Ouest-France ne citent pas de méthodologie officielle et n'ont pas lu à ce propos. Par exemple, pour Nadia Fafi, lors de la construction de l'ontologie du domaine maritime, elle n'a utilisé aucune méthodologie préexistante : « c'était vraiment la découverte » et « il n'y avait pas de méthode anticipée ».

Le chercheur Michel Chein a une vision plus catégorique et négative : il n'a pas lu sur « des méthodologies de construction dites "officielles" comme ARCHONTE, Terminae », ça ne lui dit « strictement rien » ; s'il a lu dessus « c'était il y a longtemps » ou il a tout oublié car cela ne lui « apportait rien ». Néanmoins, il sait que « plein de choses [...] ont été écrites sur "comment construire une ontologie", des méthodologies, des choses comme ça ... » mais il trouve cela « assez creux ». Pour lui, ces documents s'apparentent souvent à des manuels d'utilisation des

éditeurs d'ontologies tels que Protégé ou COGUI : « Des fois, des gens ont construit des outils et après, quand ils font "une méthodologie de construction d'ontologie", ça revient à une description de leur outil ». Ainsi Michel Chein n'utilise pas ces méthodologies consignées dans la littérature puisqu'il les juge inutiles et peu pertinentes.

Des méthodologies tout de même utilisées...

Néanmoins, deux chercheurs citent des méthodologies officielles et les ont utilisées dans le cadre de leurs projets. Pour toutes les ontologies médicales dont il a dirigé la construction, Jean Charlet a pris appui sur des méthodes très cadrées et rigoureuses. Il cite notamment ARCHONTE qui a principalement aidé son équipe « à mettre au point la raison pour laquelle [ils organisent] la hiérarchie des concepts ». Il a également utilisé l'outil d'analyse de traitement du langage SYNTEX-UPERY ; ce logiciel, constitué de deux modules, est cité dans de nombreux articles centrés sur la construction d'ontologies à partir de textes.

Florence Amardeilh, dans le cadre du programme *Nos recettes* du projet OFS a aussi utilisé une méthodologie officielle, « qui a été faite par des Espagnols il y a une dizaine d'années de ça, une méthodologie qui explique la manière de s'y prendre pour créer une ontologie ». Néanmoins, c'est Sylvie Després, enseignante-chercheuse au Laboratoire d'Informatique Médicale et Ingénierie des Connaissances pour l'e-santé (LIMICS), qui a eu l'idée d'utiliser cette méthode. En effet, elle a collaboré avec Florence Amardeilh sur ce projet, notamment sur la construction de l'ontologie. Florence Amardeilh avait aussi utilisé des méthodes officielles « pour construire des ressources termino-ontologiques » dans le cadre de sa thèse centrée sur l'informatique, la linguistique et le Web sémantique.

Des approches formalisées et officielles sont donc utilisées pour la construction des ontologies des projets LERUDI et OFS : peut-être que l'étendue de l'ontologie ou le nombre de personnes collaborant à la constitution de ces ressources nécessitent l'usage d'une méthodologie rigoureuse et cadrée ? À l'inverse, on peut supposer que pour des ontologies comprenant peu de concepts (moins de cent) telles que l'ontologie Socle ou celle du domaine maritime, il n'est pas indispensable de faire appel à des méthodologies formalisées.

...mais appliquées avec plus de liberté

Pourtant, même si Florence Amardeilh fait explicitement référence à une méthodologie officielle, celle-ci n'est utilisée que partiellement. En parlant de la méthode NEON, elle rapporte « après je ne suis pas sûre qu'on l'a suivie à la lettre ». Dans la suite de son propos, elle s'attarde davantage sur « des choix de modélisation » qui doivent être effectués. Ainsi, pour elle, le choix de la

méthode de construction n'est pas crucial, ce sont surtout les choix lors de la constitution de l'ontologie qui feront la différence et participeront à la qualité de la ressource : elle se demande « telle relation entre entités, est-ce que justement je la modélise par une relation ou est-ce que ça va plutôt être un attribut avec une *data-propertie* ? ».

Par ailleurs, une approche est partagée par plusieurs chercheurs : construire une ontologie ne demande pas nécessairement la lecture d'articles scientifiques expliquant la manière de procéder. Avec un peu de logique et quelques bonnes pratiques en tête, il est possible de s'en passer. Pour Michel Chein, l'élaboration d'ontologies « c'est souvent que du bon sens, c'est pas beaucoup plus que du bon sens. » Florence Amardeilh partage ce point de vue : « il n'y pas 36 manières non plus de construire une méthodologie ». Ils laissent donc entendre que dans certains contextes, les concepteurs pourraient faire l'impasse sur ces méthodes complètes et bien cadrées.

Pour Florence Amardeilh, les professionnels peuvent avoir besoin de se référer à des guides méthodologiques lorsqu'ils débutent dans la conception d'ontologies, mais peuvent rapidement les laisser de côté lorsqu'ils ont acquis davantage d'expérience. Elle nuance ses propos en précisant que chacun a sa propre expérience de travail et que certaines personnes, mêmes expérimentées, préfèrent continuer à se référer à des approches formalisées :

> « Disons que je pense que ça guide beaucoup quand on démarre la modélisation d'ontologies, quand on n'en a jamais trop fait avant, ça permet d'avoir un guide des principes de modélisation qui sont effectivement très bien à avoir en tête. Et après, plus on a de l'expérience, et on a moins besoin de regarder les guides. C'est vrai qu'aujourd'hui, je ne regarde plus trop les guides...En fait chacun dérive sa propre méthodo de sa propre expérience, selon le contexte aussi. »

Pour Florence Amardeilh et Michel Chein, il paraît moins pertinent de parler de « méthodologies de construction » que de « principes de modélisation », de « bonnes pratiques » ou de « normes communes ». Il ne s'agit pas de respecter scrupuleusement et dans son intégralité une méthode rigide, mais de prendre en compte des pratiques communes à de nombreux concepteurs. Selon Michel Chein, qui a participé à plusieurs projets dans différents domaines et collaboré avec des médecins, des agronomes, etc., « il y a des normes communes qui font partie du folklore. » Quelques bonnes pratiques seraient ainsi partagées de manière presque implicite par tous les professionnels du domaine des ontologies.

Néanmoins, le chercheur n'y accorde pas une très grande importance. Il place davantage l'accent sur l'éditeur d'ontologies COGUI que son équipe utilise ; en effet, leurs « bonnes pratiques découlent de l'utilisation de cet outil. » Le logiciel a renforcé les bonnes pratiques grâce à « un modèle théorique » implanté au cœur de celui-ci : c'est pour cela « qu'une relation doit commencer par une minuscule, les noms de classes commencent par une majuscule, il y a plein de

gens qui font ça ». Pourtant, pour lui, à part ces réflexions concernant davantage la forme des entités de l'ontologie, « il n'y en a pas plus de pratiques que ça ».

Donc nous constatons donc que seul Jean Charlet a appliqué à la règle une méthodologie de construction de référence pour constituer l'ontologie du domaine médical. Les autres concepteurs ne connaissent pas ces méthodes et/ou ne les utilisent pas. Ils peuvent aussi prendre des libertés par rapport à la méthode stricte, comme c'est le cas pour Florence Amardeilh. Ces premières constatations nous amènent à penser que les chercheurs ont une vision différente des « professionnels » puisqu'ils sont les seuls à prendre appui sur des méthodologies consignées dans la littérature scientifique. Néanmoins, la vision de Michel Chein - convaincu de l'inutilité de ces méthodes - contraste fortement.

Des bonnes pratiques semblent se dégager, même si elles n'occupent pas une place prépondérante pour Michel Chein. Nous pouvons alors supposer que les concepteurs accordent moins d'importance à une méthodologie dans son intégralité qu'à un ensemble de pratiques communes à tous. En outre, nous allons découvrir que même si ces méthodes sont peu utilisées, des étapes de construction conceptualisées dans des articles scientifiques sont communes à de nombreux projets, sans que les concepteurs n'en aient forcément conscience.

3.1.2 Néanmoins, des étapes de conception communes

Toutes les personnes interrogées font mention d'étapes de construction, même si elles ne citent pas explicitement une méthodologie cadrée. Lorsque nous demandons à Nadia Fafi ce qu'elle entend par « méthodologie de construction d'ontologies », elle répond « Moi quand tu dis "méthode", j'ai des étapes en tête, des étapes de travail ». Pour elle, ces méthodes regroupent alors plusieurs étapes successives à accomplir pour aboutir à une ressource finie. Florence Amardeilh partage cette opinion, pour elle « il y a de grandes étapes » même si elles sont très générales et peu précises : « […] mais après grosso modo, ces étapes-là de production d'ontologies sont assez classiques ». Des étapes peuvent être empruntées indifféremment à diverses méthodologies de construction.

Repérage des concepts : en relation avec les objectifs de l'ontologie

Une première et longue étape est le repérage des concepts-clés qui auront leur place au sein de l'ontologie. Pour Florence Amardeilh, « on va d'abord s'intéresser au vocabulaire, […] se demander s'il y a des concepts qui émergent et comment je peux les représenter ». La manière de mener à

bien cette phase va dépendre des objectifs finaux de la ressource. À Ouest-France, l'ontologie Socle et celle du domaine maritime viseront à valoriser les contenus du groupe SIPA-Ouest-France. Marie-Paule Cochet insiste sur cette fonction : l'ontologie Socle,

> « ...ça va servir à aller interroger les contenus plus facilement que par mots-clés, de façon plus pertinente […]. Donc le premier usage, c'est vraiment ça, c'est de pouvoir faciliter la fouille des contenus, faire ressortir des contenus, d'éclairer complètement les contenus parmi les 10 000 articles qu'on reçoit chaque jour. Ce n'est pas facile, juste avec des mots-clés. »

En plus de faciliter l'interrogation des contenus du groupe, l'ontologie permettra de désambiguïser le texte de certains articles et « d'extraire les entités des articles avec plus de précision » : l'annuaire pourra être enrichi avec de nouveaux concepts présents dans l'ontologie. Par exemple, on pourra à terme trouver la notion de « PDG » dans l'annuaire et extraire automatiquement le nom d'un PDG d'une entreprise, si ce nom est accolé au terme « PDG » dans un article :

> « Et même si on n'a pas les instances dans l'annuaire, on suggère que des entités sont pertinentes. Par exemple, dans la base de connaissances, on n'a pas encore le PDG d'Orange et si, dans un des articles, on voit « Jean Dupont, PDG d'Orange, dit… », [l'algorithme] peut suggérer qu'il est PDG d'Orange. »

Pour l'ontologie du domaine maritime, Nadia Fafi souligne également le fait que l'ontologie devait valoriser les contenus relatifs à ce domaine : « l'idée était d'optimiser l'organisation des données sur le domaine maritime ». Nadia a aussi dégagé des thématiques et notions principales « pour donner des points d'accès sur l'interface [DataMaritime] » ; comme elle l'explique, sur la plateforme, les contenus ne seront pas simplement classés sous une catégorie « Article » mais bien sous différentes thématiques pour qu'ils soient davantage valorisés.

Pour ces deux projets, il est donc essentiel que lors de la construction de l'ontologie, on prenne appui sur les contenus du groupe : le vocabulaire utilisé devra idéalement être similaire puisque la ressource valorisera des articles présents dans la base de connaissances. Ainsi, Nadia Fafi s'est toujours centrée et recentrée sur les contenus lors du repérage des concepts-clés de l'ontologie, elle accorde une grande importance à cela :

> « [...] pour moi, ce qui était vraiment important, et ce qui m'a beaucoup aidée, c'est que je ne m'éloignais jamais des contenus, c'était un élément qu'il ne fallait pas que je perde de vue. Mes données, c'est de là qu'elles vont venir donc je ne m'éloignais pas des contenus, j'allais vraiment voir quel type de données j'avais à chaque fois dans les articles ».

Avec Marie-Paule Cochet, elles se focalisaient sur un article du domaine maritime et allaient « à la pêche aux données dedans » pour voir quel type de données et de concepts étaient présents ; ces concepts pouvaient ensuite potentiellement être intégrés dans l'ontologie. Aujourd'hui, la

méthode préconisée par Marie-Paule Cochet est également de « partir des contenus ». Les professionnels du service BDC constituent des corpus sur des thématiques précises. En se centrant sur quelques-uns de ces articles riches en données, ils construisent une carte mentale avec les instances et relations présentes dans l'article. Donc, « des concepts et relations commencent à apparaître ». À partir de la carte mentale et des instances, ils essaient de « définir les concepts, les relations et de modéliser l'ontologie à partir de là. »

Nadia Fafi, Henri-Maxime Suchier et Florence Amardeilh ont aussi indiqué avoir utilisé une carte mentale ou *mindmap* lors du premier repérage des concepts, avant de formaliser ceux-ci dans un éditeur d'ontologies. Protégé a été utilisé dans le cadre des projets LERUDI et OFS ; le service BDC et Michel Chein ont employé COGUI, développé par le LIRMM, où travaille Michel Chein. L'usage de ces outils semble alors être la norme dans le domaine de la conception d'ontologies.

Pour l'ontologie Socle et celle du domaine maritime, on se rapproche davantage d'une construction en partant de zéro : en effet, les connaissances sont formalisées par des êtres humains. Néanmoins, les concepteurs prennent également appui sur des textes, même si des outils de TAL ne sont pas utilisés pour repérer les concepts et les relations de manière automatique ou semi-automatique. Comme nous l'écrivions dans notre état de l'art, « en constituant au préalable un corpus représentatif du champ de connaissances, on réussit à extraire certains, voire tous les constituants d'une ontologie (concepts, relations, instances, propriétés) à l'aide des outils de TAL ». Ici, l'extraction des entités se fait à la main, peut-être par manque de moyens, de temps ou d'expertise dans le domaine du TAL, puisque ces outils spécialisés sont complexes à mettre en place.

Même si Michel Chein n'en a pas conscience, il s'est inspiré d'une approche préconisée par les chercheurs Gruninger et Fox dans le cadre du projet TOVE (*Toronto Virtual Enterprise*) centré sur le domaine de l'entreprise[118]. Effectivement, ils conseillent de « prendre appui sur des problèmes qui se posent dans le domaine d'application […]. Ces points problématiques sont formulés sous forme de questions auxquelles la future ontologie devra pouvoir répondre ». Prenant place au sein du projet ICODA, cette ontologie est centrée sur la démission des conseillers municipaux. Les concepteurs étaient « toujours guidés par l'objectif final qui est celui de poser des questions concernant la démission de « qui ? quand ? où ? pourquoi ? quel est son profil ? ». De plus, selon lui, définir les concepts et les propriétés fondamentales « sans essayer de faire tourner un système à partir des questions que vont poser les utilisateurs, ce n'est pas très pertinent. » Le domaine de connaissances était « tellement compliqué » que l'équipe a choisi cette approche pour effectuer les choix nécessaires.

[118] Cf. « 1.3.2 Construction d'ontologies en partant de zéro », p.44

Ainsi, la phase de repérage des concepts et des relations est spécifique à chaque projet, elle est conçue en lien étroit avec les objectifs auxquels l'ontologie répondra.

La réutilisation de SOC : une constante

Une des approches communes à tous les projets, au-delà des divers objectifs de l'ontologie, est l'utilisation de Systèmes d'Organisation des Connaissances ou d'autres sources de connaissances pour repérer des concepts. Cette manière de procéder, présentée dans de nombreuses méthodologies de référence semble alors partagée par la communauté professionnelle.

Les chercheurs de notre échantillon accordent une place importante à l'utilisation de SOC. En effet, cet appui sur des ressources préexistantes permet de ne pas refaire ce qui a déjà été fait, et donc de gagner du temps et de partager des conceptualisations communes entre différentes structures. C'est une phase indispensable pour Jean Charlet : « En médecine, on sait qu'on a des référentiels incontournables dès qu'on commence à travailler […]. De toute façon, on sait qu'on va regarder ce qu'il y a là-dedans par rapport à la spécialité qui nous intéresse quand on va faire notre ontologie ».

Même si pour différents projets, il choisit de ne pas réutiliser une ontologie, « ce n'est pas tout à fait partir de zéro : […] quand [il] construit une nouvelle ontologie, [il] regarde toujours ce qui a déjà été fait dans le domaine » puisque les concepteurs se documentent toujours sur les ressources du domaine qui existent déjà. Ceci est peut-être lié au domaine médical ou des classifications sont reconnues et utilisées par de nombreux professionnels.

Michel Chein a pris appui sur « des ontologies utilisées beaucoup en documentation » et qui constituent selon lui « des normes », il a donc utilisé des concepts issus de l'*International Committee for Documentation Conceptual Reference Model* (CIDOC-CRM) ou du *Functional Requirements for Bibliographic Records* (FRBR). Il s'est aussi basé sur « des sources structurées » comme le vocabulaire de l'INSEE. Henri-Maxime Suchier s'est lui aussi basé sur des ressources de référence comme l'INSEE pour les lieux ou l'ontologie RDF *Friend of a friend* (FOAF) pour les personnes et organisations. Il dit avoir utilisé « des ressources génériques » et « classiques » puisque l'ontologie Socle « est sur quelque chose de globalement générique ». Pour lui, il est également intéressant de se renseigner sur ces ressources lorsqu'il rencontre des difficultés lors de la conception de l'ontologie, il veut savoir « comment [les concepteurs des ressources de référence] ont inversé le problème ».

Cependant, Florence Amardeilh présente tout de même les limites de ce type de réutilisation. Elle avait tenté de réutiliser « l'embryon d'ontologie » Taaable, également centrée sur le domaine de la cuisine. Néanmoins, cette phase n'a pas fonctionné puisque « ce qu'[ils avaient] récupéré ne

collait pas tout à fait avec les exigences de finesse qu' [ils devaient] prendre en compte pour le projet ». Comme nous l'avions vu dans la première partie, il est souvent difficile de réutiliser des SOC puisque les conceptualisations doivent être réajustées entre les deux ressources ; en outre, les objectifs des ontologies diffèrent souvent. Donc la réutilisation de ressources (terminologiques ou non) permet de gagner du temps et de s'inspirer de sources extérieures pour modéliser un domaine de connaissances précis. Cependant, il faut rester conscient des limites de ce genre d'approche.

L'évaluation de l'ontologie : une étape incontournable

Pour les chercheurs, l'évaluation de la ressource - une des sept étapes du cycle de vie des ontologies par Fabien Gandon[119] - est indispensable. Michel Chein est de l'avis qu'il faut faire des évaluations dès le début du projet et « le plus souvent possible », « à chaque instant, à chaque étape », il ne faut « pas attendre la phase finale pour faire l'évaluation ». L'évaluation est la phase la plus importante dans une méthode de construction d'ontologies, elle ne doit en aucun cas être négligée : « S'il y a quelque chose à développer au niveau de la méthodologie, à mon avis c'est sur l'évaluation. »

Jean Charlet partage ce point de vue ; la phase d'évaluation a été complète et effectuée en collaboration avec des experts du domaine :

> « La méthodologie d'évaluation a vraiment été importante, on a vraiment travaillé avec les médecins. Dès l'instant que les médecins étaient impliqués dans le processus de construction, que le système était évalué... On a fait des évaluations et on a enrichi l'ontologie quand on repérait des manques lors de la campagne d'évaluation. »

Pour le projet OFS, la phase de validation a également été très rigoureuse : les concepteurs ont mis en place une double validation. La chercheuse Sylvie Desprès et son équipe ont « [validé] la connaissance qui était modélisée dans l'ontologie ». Ensuite, à Mondeca, l'équipe de Florence Amardeilh a fait « la validation technique, opérationnelle, [ils se demandaient] "est-ce qu'on obtient bien les bons raisonnements attendus ? ». Donc les concepteurs se remettent constamment en question lors de cette phase d'évaluation et de validation.

Les professionnels du service BDC à Ouest-France n'ont pas abordé le sujet de l'évaluation de l'ontologie. En effet, l'ontologie du domaine maritime n'a pas fait l'objet d'évaluation. L'ontologie Socle n'a pas atteint sa forme définitive, le projet est moins avancé que ceux des chercheurs, ainsi une phase d'évaluation à ce moment était peut-être moins indispensable ? Comme nous le

[119] Cf. « 1.3.1 Le cycle de vie des ontologies », p.42

verrons par la suite, ce différentiel peut aussi s'expliquer par les divers niveaux de compétence et d'expérience des concepteurs d'ontologies de notre échantillon.

Ainsi, les méthodologies de construction officielles sont peu lues et/ou peu utilisées. Nous constatons que les concepteurs prennent davantage de liberté par rapport aux méthodes. Malgré cela, cette construction est toute même régie par des règles communes à ce domaine : utilisation de quelques bonnes pratiques, étapes de construction communes et faisant parfois partie de méthodologies officielles, poids très important de l'évaluation. Des « règles métier » existent donc et sont appliquées, même si les professionnels n'abordent pas d'eux-mêmes la question des méthodologies ou règles formalisées.

Nous avons constaté que les concepteurs évaluaient souvent leurs prototypes et remettaient constamment en question leur travail afin de créer l'ontologie la plus juste et la plus adaptée possible. Nous allons maintenant découvrir que ces personnes effectuent également de nombreuses itérations, ceci tout au long du projet d'élaboration de la ressource. Utiliser des méthodes plus tâtonnantes et moins formalisées permettrait de prendre davantage en compte les utilisateurs futurs de l'ontologie.

3.2 Les méthodes utilisées aujourd'hui sont moins formalisées et plus itératives pour que les ontologies soient davantage adaptées aux usages des utilisateurs futurs.

3.2.1 Des méthodes moins formalisées et moins strictes

Des méthodes davantage itératives

La grande majorité de notre échantillon construit des ontologies de manière itérative, il y a de nombreux retours en arrière, les méthodes sont moins strictes et arrêtées.

C'est le cas pour le projet de l'ontologie du domaine maritime : les tâtonnements sont partie prenante du projet de construction. Nadia Fafi préférait élaborer une ontologie complète, « aller au plus large » et ajuster ensuite le schéma selon les données qui n'étaient finalement pas nécessaires. Elle a aussi appris à utiliser les outils indispensables au fur et à mesure : « Pour la construction de l'ontologie, on a utilisé le logiciel Protégé. Je ne connaissais pas du tout comme c'était ma première ontologie. On a découvert petit à petit comment on allait l'utiliser. » Pour ce projet, Nadia Fafi, Marie-Paule Cochet et Henri-Maxime Suchier ont collaboré pour mettre eux-

mêmes en place une méthodologie de conception : « C'était ça qui était assez incroyable pendant ce stage ! […]. On découvrait ensemble comment on allait constituer notre première ontologie. » Elle ajoute « Oui, la méthode on ne l'a pas forcément anticipée ». Pour l'ontologie Socle, Marie-Paule Cochet reprend cette approche : « Après, on n'a pas bon du premier coup donc on se pose des questions. »

Les chercheurs utilisent eux aussi une manière de procéder très itérative. Michel Chein n'a pas de méthodologie arrêtée et utilise à plusieurs reprises le terme « bricolage » pour décrire la façon dont il a construit la ressource : « Ça, j'ai envie de dire, c'est un peu du bricolage ! Comment on a fait ? Donc, une partie a été faite par des gens de Ouest-France, on en avait parlé lors de la réunion. C'est vraiment du bricolage. » Florence Amardeilh se soucie moins de la méthodologie et des étapes précises que de l'évolution permanente de l'ontologie : « Pour tout vous dire, je ne me fie pas vraiment à la méthodo comme elle a été formalisée, je ne vais pas me dire "à quelle étape je suis de cette méthodo-là ?". […] Après, je construis ça de manière itérative. » Le projet n'est pas constitué d'étapes bien délimitées et linéaires ; les retours en arrière sont permanents, il faut constamment s'adapter et reconstruire certaines parties de la ressource. Cette façon de procéder a pu rendre le projet plus complexe à suivre :

> « ...du coup c'était un peu compliqué de suivre le train et de rattraper les wagons en fonction de ce qu'on nous demandait, au moment où on nous le demandait. Du coup, il y a des choix qui ont été faits, il fallait revenir sur l'ontologie, casser des arbres que l'on avait faits car le point de vue changeait complètement, il fallait refaire. »

Michel Chein est encore plus convaincu, utiliser une méthode itérative, « c'est absolument fondamental » : en effet, comme il a pu le constater dans le cadre de ses projets précédents, procéder de manière linéaire ne fonctionne pas du tout. « Une démarche complètement séquentielle » où les utilisateurs, puis les experts sont consultés, où « on représente les connaissances des experts et on implémente », ça ne marche pas, « ça se casse la figure ». Il ajoute :

> « ...mais c'est vraiment un processus itératif. Au début les gens pensaient qu'il pouvait y avoir des étapes séquentielles, ça c'est pas vrai du tout, on s'est rendu compte que ce n'était pas raisonnable, c'est pas possible...C'est la méthode itérative qui, le plus rapidement possible, prend en compte les problèmes qu'on veut résoudre. »

Les pratiques des concepteurs sont donc conformes aux règles fondamentales que Natalya F. Noy et Deborah L. McGuinness préconisent de suivre[120]. Selon elles, le développement d'une ontologie

[120] Cf. « 1.3.1 Le cycle de vie des ontologies », p.42

est « nécessairement un processus itératif », on ne peut pas passer d'étapes en étapes sans revenir sur celles-ci en cours de projet. Maedche et Staab avaient aussi conceptualisé une méthode itérative à partir de textes pour la construction semi-automatique d'ontologies[121].

Néanmoins, lorsque nous demandons à Jean Charlet s'il pense que par rapport aux années 2000, les méthodologies sont plus itératives et moins linéaires, il n'est pas convaincu : « Pour moi, c'est à peu près semblable. » Nous pouvons donc, pour l'instant, valider partiellement notre hypothèse en la nuançant. Les méthodes utilisées ne sont pas forcément *moins* formalisées et *plus* itératives qu'au début du développement de ces ressources dans les années 2000 ; par contre, il apparaît qu'elles sont *peu* formalisées et *très* itératives pour les projets sur lesquels nous avons porté notre attention.

Des modifications facilitées par les outils informatiques

Ces itérations sont facilitées par les langages et outils informatiques, il est donc plus facile de ne pas utiliser de méthodologie bien cadrée et de revenir à de nombreuses reprises sur les choix effectués.

Pour Henri-Maxime Suchier, une méthode bien linéaire n'est pas indispensable puisque dans le domaine informatique, des traductions sont facilement mises en place. Il assure « Oui, on peut toujours traduire d'un modèle vers un autre. […] Au final, c'est pareil...toutes les traductions sont possibles, après il y en a qui sont plus coûteuses que d'autres, mais tu peux toujours passer de l'une à l'autre » et « de toute façon, c'est tellement facile en informatique de passer d'une syntaxe à une autre, on sait faire des traductions ». RDF est aussi très utile puisqu'il est possible de « faire un choix à un moment donné et le remettre en question après », on peut notamment modifier le niveau de granularité de l'ontologie après coup : « Si ce n'est pas assez détaillé, je redétaille [le niveau de granularité] et je suis capable de redisperser toutes mes entités par rapport à ce nouveau niveau de granularité ».

Ainsi, pour lui, plusieurs modélisations potentielles sont possibles, il n'y a pas qu'un seul modèle envisageable, les concepteurs peuvent effectuer de multiples itérations pour éventuellement passer d'un modèle à un autre « […] ça montre bien qu'il n'y a pas juste une façon de faire. Il n'y a pas une seule modélisation possible mais plusieurs points de vue qui vont à chaque fois correspondre à un besoin ou une compréhension du problème. »

Les manières de construire une ontologie sont aussi fonction des moyens disponibles alloués au projet ainsi que de l'expérience et de l'expertise des concepteurs.

---

[121] Cf. « 1.3.3 Construction d'ontologies à partir de textes », p.46

3.2.2 Des méthodes dépendantes du projet et des concepteurs

Des méthodes dépendant des moyens mis en œuvre

La méthodologie choisie est tout d'abord très dépendante des contraintes techniques et temporelles du projet.

L'ontologie du domaine maritime était la première à être construite au sein du service BDC, principalement par Nadia Fafi, en stage à cette période-ci. Pour ces professionnels, « c'était vraiment la découverte » ; Nadia avoue que ce projet était « un peu expérimental » pour eux. Pour Marie-Paule Cochet, durant la construction de cette première ressource, ils ont « vu les balbutiements et le potentiel qu'il y avait » dans le domaine des ontologies. Par ailleurs, la construction de l'ontologie a dû être effectuée en moins de trois mois, soit la durée du stage de Nadia Fafi. Elle le fait remarquer avec des expressions qui traduisent le manque de temps : « c'était un peu l'urgence, j'étais stagiaire », « à la fin de mon stage, on était un peu dans le *rush*. » Les contraintes temporelles étaient donc fortes ; il a fallu rapidement improviser une méthodologie de construction.

L'importance du projet OFS et de la taille de l'ontologie a aussi conditionné la manière de la construire. En effet, cette ontologie a été pensée et élaborée sous la forme de six/sept modules (par exemple, la cuisine, les aliments ou la nutrition) « parce que c'était vraiment une très très grosse ontologie » : il y a « entre 6000 et 7000 concepts » selon elle. Cette ressource aurait été trop complexe à élaborer sans cette compartimentation sous plusieurs modules thématiques. Le projet de recherche a aussi duré plus longtemps que la construction de l'ontologie du domaine maritime, le nombre de collaborateurs était plus élevé, et le nombre de concepts n'était pas comparable. Il est alors normal qu'une méthodologie officielle et anticipée ait été suivie dans les grandes lignes, alors que ça n'a pas été le cas pour le projet DataMaritime.

Temps, caractéristiques techniques et moyens disponibles influent sur la manière de construire des ontologies. Nous allons voir que les niveaux de formation et de compétence des concepteurs conditionnent également cela.

Une question de formation et de compétences

Michel Chein et Jean Charlet partagent ce même point de vue : les projets de construction d'ontologies sont très ambitieux, coûteux en temps et demandeurs de solides compétences de la part des concepteurs. Pour Jean Charlet, ces points peuvent souvent devenir des contraintes pour la réalisation de l'ontologie :

> « Après, il faut mettre en balance le coût du développement de l'ontologie, et ça peut souvent être un problème. Ça peut aussi être un problème de compétence parfois parce que les gens n'ont pas les compétences. Et donc un problème de coût parce que dans tous les cas de figure, on a des gens qui travaillent des mois et des mois sur l'ontologie. »

Michel Chein est persuadé qu'il faut avoir des compétences assez élevées en logique et informatique : « ça c'est clair ! On ne peut pas construire une ontologie sans connaître un minimum de logique du premier ordre. Parce que si des gens [sans connaissance] construisent quelque chose, ils ne sauront pas comment cela sera utilisé ensuite. »

Jean Charlet a donc formé des médecins qui ont collaboré à la conception de l'ontologie. Sur le projet LERUDI en particulier, il n'y a « pas eu de point bloquant » parce que le médecin était intéressé par le domaine des ontologies. Par contre, pour d'autres projets, le temps de formation peut être plus long, « il y a une espèce d'acculturation », ce domaine peut être complexe à appréhender au début. Florence Amardeilh a aussi dû former des professionnels de SEB non-initiés à la question des ontologies et cela a demandé un temps assez important :

> « C'est plus avec les gens de chez SEB, que c'était des fois plus compliqué. Il fallait adopter le même vocabulaire qu'eux, ils ne savaient pas forcément ce qu'était une ontologie et à quoi ça servait, donc il fallait un peu les éduquer, en même temps avancer. Donc là, c'était un peu plus compliqué. »

Pour l'ontologie du domaine maritime, la majorité des collaborateurs avaient été peu formés sur ce domaine avant le commencement du projet, même si Marie-Paule Cochet avait assisté à des formations théoriques sur ce sujet. Seul Henri-Maxime Suchier possédait des compétences techniques sur les ontologies. Il fallait donc se confronter à la pratique ; Nadia Fafi relate cette expérience :

> « Quand je suis arrivée en stage, j'avais eu juste un cours de trois heures sur les ontologies...et ce n'était pas au centre de mon sujet de mémoire. Et comme c'était la première ontologie du service, j'avoue qu'au début c'était plutôt Henri-Maxime qui nous guidait. […] Moi, je découvrais vraiment avec lui. Il avait vraiment un bagage technique plus fort. »

Il est aussi nécessaire d'acquérir au fil du projet une expertise du domaine de connaissances pour faire les meilleurs choix de modélisation ; ça a été le cas pour Nadia Fafi : « Au départ, je ne connaissais pas grand-chose au domaine maritime et au fur et à mesure, je développais une expertise, je commençais à développer un modèle, j'avais des doutes de temps en temps parce que je découvrais en même temps le domaine maritime, toute son économie... »

Il est donc normal et compréhensible que des concepteurs, comme Florence Amardeilh ou Jean Charlet - ayant suivi des formations dans le domaine du Web sémantique, de l'Ingénierie des

Connaissances et possédant une expérience significative dans la construction d'ontologies - utilisent des méthodologies faisant référence et davantage formalisées. Michel Chein fait figure d'exception puisqu'il choisit de ne pas utiliser de méthodologie précise, même s'il a probablement lu sur ce sujet auparavant.

On peut supposer qu'en règle générale, plus les professionnels ont d'expérience dans la conception d'ontologies et plus ils utilisent une méthodologie construite et cadrée. Lors de la constitution de la deuxième ontologie du service, une petite méthodologie - même si elle n'est pas « officielle » - a été validée :

> « Par rapport au projet de recherche qui s'est mis en place avec ICODA, on a dit qu'il fallait reprendre le travail de Nadia. On voyait bien la démarche, on a validé la démarche. [...] Alors l'objectif, c'était de tout remettre à plat, de repartir un peu de zéro techniquement au niveau de l'ontologie, tout en ayant investi sur tout le travail. »

Dans la majorité des cas, les méthodes sont donc itératives et libres, il y a peu de méthodologies assez « rigides » mises en place, même si elles diffèrent selon les modalités du projet et l'expérience des concepteurs d'ontologies. De nombreuses itérations sont effectuées lors du développement de l'ontologie : elles peuvent permettre de revenir vers les futurs usagers pour les questionner de nouveau sur leurs attentes et besoins. Effectivement, il est important de construire des ontologies avec et pour des utilisateurs identifiés. Nous allons maintenant développer ce point et découvrir cependant que cette exigence n'est pas toujours respectée.

3.2.3 Construire des ontologies plus adaptées aux usages des futurs utilisateurs

Une prise en compte des usages

La majorité des personnes de notre échantillon ont consulté les usagers de la future ontologie et ont tenté de prendre en compte leurs remarques. Ce point semble être commun à la quasi-totalité des projets de construction d'ontologies impulsés aujourd'hui.

Avant de mettre en place le projet de l'ontologie Socle, les personnes de la BDC ont démarché les journalistes de Ouest-France pour connaître leurs besoins et les thématiques qui pourraient les intéresser. Ainsi, il y a eu « un brainstorming commun » entre des datajournalists et ce service. Ils ont pu évoquer « plusieurs sujets autour de l'exploitation des contenus via la base de connaissances » comme les thèmes de la politique locale ou des commerces. En effet, pour Henri-Maxime Suchier, « le fait d'impliquer des gens qui ne sont pas des informaticiens, mais qui sont des utilisateurs finaux des outils, c'est une façon d'avoir au final un outil qui soit adapté aux

utilisateurs ». Michel Chein s'est lui aussi appuyé sur la demande d'un datajournalist qui souhaitait étudier la démission des conseillers municipaux dans les contenus du groupe. Il insiste là-dessus : « Vraiment, on s'est appuyés sur une demande réelle d'un journaliste. »

Nadia Fafi accorde aujourd'hui beaucoup d'importance à la prise en compte des besoins des usagers : « Pour moi, j'ai beaucoup plus à cœur la question des usages maintenant. [...] j'ai aussi beaucoup plus en tête quand on fait un *brainstorming* avec Marie-Paule Cochet et Henri-Maxime Suchier, l'usage final pour essayer de voir si on a trop affiné ou pas […]. »

Jean Charlet a sollicité « très très souvent » un médecin-urgentiste pour ce projet, encore plus que pour les autres projets de construction d'ontologies médicales : en effet, peu de documents concernant le domaine des urgences « tel qu'il se pratique » existaient. Il a donc fallu faire davantage appel à une personne experte du domaine, qui est aussi un potentiel utilisateur de l'application future.

Pourtant, des difficultés à faire remonter les attentes des utilisateurs

Pourtant, dans beaucoup de projets, les concepteurs éprouvent parfois des difficultés à connaître réellement les besoins des usagers ou à les recontacter lorsque le projet est en cours de gestation.

Pour la première ontologie du service BDC centrée sur le domaine maritime, Nadia Fafi n'était pas en relation avec les usagers et n'a pas pris en compte leurs demandes :

> « Non, je n'avais pas d'impératif au niveau des usagers parce que je n'étais pas en relation avec eux. Je n'avais aucun moyen de communiquer avec les potentiels usagers. [...] Je ne prenais pas tellement en compte les besoins des usagers, pour être honnête, car je ne communiquais pas avec. »

En effet, elle suivait les recommandations et attentes d'Alexandra Turcat[122] qui jouait le rôle de médiatrice entre la conceptrice de l'ontologie et l'entreprise commanditaire (le Cluster Maritime) dans le cadre de ce projet. Michel Chein n'a plus de lien pour le moment avec les datajournalists qu'il avait interrogés en début de projet car son équipe « passe par l'intermédiaire des gens de Ouest-France » pour cette tâche, qui semble être difficile à pérenniser.

Pour l'ontologie Socle, les professionnels connaissent les thématiques qui intéressent certains journalistes ; néanmoins, Henri-Maxime Suchier nuance cela puisque, pour lui, « pour l'instant, les usages ne sont pas forcément bien définis. » Il faudrait donc se pencher encore davantage sur la question des usages attendus pour combler cette lacune. Pourtant, il « imagine que derrière, les usages en découleront » lorsque l'ontologie sera terminée et implantée dans l'annuaire. C'est donc

[122] Rédactrice en chef de l'hebdomadaire spécialisé *Le Marin*, principale interlocutrice de Nadia sur le projet DataMaritime

en quelque sorte une approche inverse à celle qui est préconisée dans de nombreuses méthodologies où l'analyse des usages doit précéder la conception de la ressource.

Florence Amardeilh a dû faire face à cette même problématique : à cause de la gestion du projet, il y a eu des difficultés à faire remonter les usages. Elle était dépendante d'autres services et attendait les usages pendant très longtemps ; son équipe et elle essayaient de prendre de l'avance en les devinant, mais cela était complexe et contraignant :

> « Donc il fallait adapter l'ontologie, la faire évoluer, on a été très très contraints parce que normalement pour qu'on élabore une ontologie, surtout dans ce genre d'application, elle doit être désignée pour un usage bien précis. Et cet usage on l'attendait pendant longtemps. Pour ne pas attendre, on essayait de voir avec les experts, avec les gens de SEB à quoi pourrait servir l'ontologie. On essayait d'anticiper les usages qui devaient redescendre après vers nous par les membres de l'UX...mais voilà il y avait toujours des décalages et c'était vraiment pas facile de travailler dans ce contexte-là avec autant de contraintes. »

De plus, les usagers ont parfois un emploi du temps très chargé et peu de temps à consacrer aux chercheurs. C'est notamment le cas dans le domaine médical, où les spécialistes sont peu nombreux et peu disponibles : Jean Charlet a parfois des difficultés à contacter ces professionnels. Tout d'abord, des spécialistes « il n'y en a pas tant que ça » puis « les médecins sont souvent très très occupés par leur activité. Si on arrive à dégager à temps, voire même qu'on décide dans le financement qu'un médecin soit payé pour travailler sur l'ontologie pendant un quart de son temps...c'est déjà très bien pour nous. »

Tous les concepteurs d'ontologies de notre échantillon sont d'accord : il faut prendre en compte les usages futurs avant de démarrer le projet de construction. Des actions ont été initiées dans ce sens, mais il semble encore difficile de pérenniser cette collaboration. Manque de disponibilité des usagers, gestion de projet complexe ou suivi des relations difficiles semblent freiner cette étape pourtant cruciale. Les concepteurs doivent parfois deviner les attentes des futurs utilisateurs, mais cette tâche est complexe et pas toujours efficace.

Utiliser un vocabulaire lisible par les usagers finaux

Néanmoins, la question du vocabulaire est une préoccupation commune dans de nombreux projets : il doit être adapté aux futurs utilisateurs de l'ontologie et parfois lisible par des publics cibles utilisant des vocabulaires différents.

Dans l'ontologie du projet OFS, Florence Amardeilh a utilisé un niveau de granularité assez fin pour modéliser les aliments « qui participent à la constitution des ingrédients de la recette ». En effet, cette ontologie servira, entre autres, à faciliter la recherche de recettes via des fonctionnalités

d'extension sémantique. Les utilisateurs devront idéalement retrouver le plus de recettes possible et ne pas être confrontés au silence documentaire, même s'ils ne cherchent pas le bon terme. Donc de nombreux synonymes sont utilisés au niveau des concepts pour que l'extension sémantique fonctionne : lorsque le futur usager recherchera « un aliment « Patate », [il trouvera] toutes les recettes qui contiennent « Pomme de terre » et vice-versa. » Elle ajoute que des termes et résultats ont été affinés « en fonction de ce que les gens cherchent » : l'expérience des destinataires est réellement prise en compte.

Pour ce même projet, il a aussi fallu définir d'autres synonymes, mais pour une raison différente : « effectivement, [ils voulaient] que des gens qui n'aient pas les mêmes vues sur la recette puissent quand même comprendre la recette ». Comme elle le fait remarquer, « un nutritionniste, un chef-cuisinier et un utilisateur lambda » n'utiliseront pas le même vocabulaire ; il faut donc s'adapter à ces personnes qui utiliseront toutes l'application, mais chercheront les recettes avec des termes plus ou moins compliqués. Un important travail a donc été effectué concernant le vocabulaire, et plus précisément les synonymes « qui permettent de donner plusieurs points de vue sur la même entité. »

Pour l'ontologie Socle, créée au sein du service BDC, et celle concernant la démission des conseillers municipaux, élaborée au LIRMM, la réflexion sur l'usage d'un vocabulaire métier est prégnante pour construire une ontologie adaptée aux utilisateurs finaux. Marie-Paule Cochet pensait que « l'ontologie pouvait être un outil pour converser avec les journalistes, voire les documentalistes », elle souhaitait donc implanter de nombreuses « notions métier » dans ce but. Pour elle, « c'est peut-être utopique », mais elle souhaiterait avoir une base de connaissances « entièrement lisible par les journalistes », ils utiliseraient alors les contenus de façon décuplée. Il ne faut donc pas perdre les usagers en ajoutant des notions qu'elle juge inutiles et trop complexes. Néanmoins, nous verrons ci-dessous que cette demande est souvent difficile à respecter du fait de contraintes plus techniques.

Michel Chein précise bien que son ontologie (du moins une des versions) « est destinée aux utilisateurs ». C'est à l'aide de cette ressource qu'ils pourront interroger des bases de données centrées sur la politique locale, ils constitueront leurs requêtes avec « le vocabulaire de l'ontologie ». Il faut donc « à tout prix que [le vocabulaire] soit le plus simple possible...enfin le plus compréhensible possible. » Il utilise aussi le concept de « nuance politique » qui fait partie du « jargon du ministère de l'Intérieur » puisqu'il souhaite s'exprimer comme les utilisateurs finaux, soit « les gens qui vont ensuite faire les statistiques sur les résultats des élections. »

Cependant, l'implémentation d'un vocabulaire entièrement compréhensible par les usagers - impératif mis en avant et respecté par beaucoup de concepteurs - entre parfois en contradiction avec des exigences techniques. Il ne faut pas oublier qu'un des objectifs principaux d'une ontologie est d'être compréhensible par une machine pour que celle-ci puisse faire des inférences sur les connaissances.

Mais construire également une ontologie pour la machine

Michel Chein préconise de construire différentes versions d'une même ontologie avec des niveaux hétérogènes de spécificité pour chacune d'entre elles. Par exemple, il y aura le niveau utilisateur, le plus simple, puis le deuxième niveau où des notions plus complexes, comme le temps et la provenance des sources, seront prises en compte. Le troisième niveau « sera lié à des problèmes d'efficacité » : des « notions purement techniques » seront ajoutées à la ressource.

Marie-Paule Cochet souligne le fait que les professionnels sont obligés de rajouter des concepts qu'elle juge « trop techniques » dans l'ontologie pour que la machine puisse comprendre la connaissance, même si c'est peu lisible. Elle a « parfois l'impression que [les professionnels de la BDC] sont allés un peu trop loin », concernant notamment le niveau de granularité qui devait rester très large puisque c'est une ontologie de haut niveau. Néanmoins, elle a conscience que cette couche supplémentaire est ajoutée pour répondre à « des besoins techniques. »

Nadia Fafi ajoute que pour l'ontologie du domaine maritime, des propriétés ont été ajoutées dans le modèle pour exprimer des dates. Elle a conscience du fait que ces informations ne sont pas toujours compréhensibles par les utilisateurs mais qu'elles sont essentielles pour que les « données restent fiables » et pour « qu'il y ait moins d'ambiguïté. » Henri-Maxime Suchier synthétise cette problématique en une phrase :

> « C'est un peu le compromis qu'on a à chaque fois, entre une ontologie qui doit pouvoir résoudre des problèmes techniques par une machine et une ontologie qui doit pouvoir être compréhensible, être manipulable par un être humain non-informaticien. »

Les manières de procéder pour construire une ontologie ne sont peut-être pas *plus* itératives qu'il y a quinze ans, mais elles restent peu cadrées et *très* itératives. Les concepteurs reviennent constamment sur les différentes étapes de conception afin de rectifier les choses, de prendre en compte de nouveaux paramètres ou d'évaluer les différentes étapes de travail. Pourtant, même si la majorité des personnes s'intéressent aux besoins des utilisateurs en termes d'informations et

adaptent le vocabulaire des ressources, ces tentatives manquent souvent de suivi. Il est parfois difficile de garder contact avec ces usagers peu disponibles.

Nous allons maintenant nous focaliser sur le lien et les possibles différences de points de vue entre les professionnels et les chercheurs, personnes souvent amenées à collaborer, comme nous l'avons vu précédemment. Certains points ont déjà été abordés en filigrane dans d'autres parties mais il nous semble crucial de consacrer une place plus importante à cette problématique potentiellement épineuse.

3.3 Il y a une différence de points de vue et de conceptions entre les chercheurs et « les professionnels », leur but n'est pas le même lors de la construction d'une ontologie.

3.3.1 Une collaboration nécessaire entre des personnes possédant des compétences diverses

S'enrichir de l'expertise d'autres professionnels

Notre échantillon partage un point de vue commun : lors d'un projet faisant appel à une ontologie, il est nécessaire de collaborer entre personnes de professions différentes afin de mutualiser les compétences et d'échanger divers points de vue.

Pour Michel Chein, quatre grands groupes de personnes sont amenés à collaborer lors de la construction d'une ontologie : les utilisateurs finaux, les experts du domaine, « les ingénieurs de la connaissance », puis les informaticiens. Les experts vont apporter la connaissance et éventuellement valider les choix des ontologistes ; les ingénieurs de la connaissance vont « modéliser les connaissances qui seraient utiles pour les utilisateurs » ; enfin, les informaticiens « se préoccupent des différents modèles techniques pour faire tourner tout ça. » Plusieurs rôles sont ainsi bien différenciés et complémentaires. Marie-Paule Cochet insiste sur cette complémentarité entre professions lors des différentes étapes de construction :

> « Au sein du service SIB[123], on travaille tous ensemble, on a besoin de tout le monde, on a besoin des journalistes, on a besoin des documentalistes, on a besoin des informaticiens. Il y a toute une démarche et à chaque étape, on se demande de qui on a besoin. »

Jean Charlet a collaboré avec une documentaliste dans le cadre d'un projet sur la toxicité nucléaire, celle-ci a effectué « un gros travail sur la qualité des sources. » Ces professionnels de

[123] Service Informatique Banque de Contenus

l'information-documentation ont donc un important rôle à jouer pour contrôler la véracité et la validité des données de l'ontologie. Ces professionnels pourraient aussi devenir ingénieurs de la connaissance, ces intermédiaires faisant le lien entre « ce qu'ont dit les experts, les gens du domaine » et « les raisonnements tels qu'ils sont implémentés dans la machine. » Par exemple, pour l'ontologie du domaine maritime, Nadia Fafi a en quelque sorte vulgarisé les connaissances qu'elle avait obtenues de la rédactrice en chef du *Marin* pour les rendre plus facilement accessibles aux informaticiens du service. Par ailleurs, ses collègues informaticiens sont « vraiment essentiels » pour faire avancer les projets de construction d'ontologies, pour les opérationnaliser. Pour elle, sans les informaticiens, « l'ontologie ne sera pas utilisée, elle restera sur une carte mentale ou dans Protégé. »

Ainsi, il est souvent nécessaire de faire appel à un expert pour mieux connaître un domaine de connaissances, vérifier la justesse des informations ou approfondir un sujet. Florence Amardeilh a sollicité « différents chercheurs des autres domaines, des nutritionnistes, des anthropologues, des chefs cuisinier de la structure Paul Bocuse, des gens de SEB » afin que pour chaque module de l'ontologie, les domaines soient caractérisés de la façon la plus juste possible.

Pour Henri-Maxime, il est rassurant de travailler avec des chercheurs spécialisés dans le domaine de l'Ingénierie des Connaissances, tels que Michel Chein. Effectivement, ceux-ci peuvent valider des modélisations déjà mises en place :

> « Et assez souvent, j'ai l'impression que c'est comme pour se dire "Bon, c'est bon, on ne fait pas n'importe quoi parce que nos choix sont confortés par ce qu'eux font". Comme à chaque fois que j'ai travaillé avec des gens qui sont experts dans un domaine, au final, tu tires profit de leur expérience et de leurs connaissances. »

Un rapport parfois disproportionné entre les collaborateurs ?

Des documentalistes, informaticiens, chercheurs, experts du domaine, collaborent donc sur des projets communs afin de mêler leurs compétences. Pourtant, les relations entre ces professionnels paraissent parfois régies par un rapport de force ; des personnes peuvent prendre la main sur les projets.

Henri-Maxime Suchier laisse entendre que les chercheurs du LIRMM sont « moteurs » et que les professionnels du service BDC sont davantage « clients » : « La collaboration est presque plus à un niveau où, nous, on est clients d'eux, dans le sens où, eux, font ce travail de conception et nous derrière on récupère. » Les professionnels posent de nombreuses questions : « Pourquoi ci ? Pourquoi ça ? Pourquoi avoir fait ce choix-là ? » et les chercheurs répondent, jouant ainsi le rôle de formateurs.

Pourtant, on peut nuancer ce possible rapport de force puisque, dans beaucoup de projets, le choix de créer une ontologie pour aider à la recherche d'information a été effectué de manière commune entre différents professionnels. Par exemple, la décision d'élaborer une ontologie pour le projet LERUDI « a été actée assez rapidement » lors de réunions où Jean Charlet était présent avec des professionnels représentant le financeur du projet (l'ASIP Santé[124]). Lors de réunions de travail, les chercheurs ont pris conjointement la décision d'utiliser certaines approches : privilégier une approche symbolique, ne pas réutiliser de ressource préexistante, etc.

Pour le projet OFS, les industriels travaillant pour SEB et Florence Amardeilh ont tout de suite pensé qu'il était « très valorisant » d'utiliser une ontologie afin de valoriser les recettes numériques.

Ainsi, certaines catégories de professionnels peuvent parfois mener le processus d'élaboration de l'ontologie. Néanmoins, dans davantage de cas, le rapport semble plus égalitaire, du moins concernant la décision d'utiliser une ontologie pour appuyer des fonctionnalités d'aide à la recherche d'information.

3.3.2 Des divergences de points de vue, d'objectifs et de pratiques ?

Même s'il est nécessaire que des personnes possédant des compétences diverses et complémentaires travaillent ensemble, cette collaboration peut parfois devenir complexe à cause de différents, particulièrement au sujet du niveau de granularité et des pratiques professionnelles.

Ces divergences de points de vue sont souvent normales puisque ces personnes ne partagent pas toujours les mêmes préoccupations, objectifs et domaines de compétence. Lors de sa collaboration avec des informaticiens, Nadia Fafi a par exemple rencontré quelques difficultés en raison de contraintes techniques ; en effet, les informaticiens devaient parfois freiner les projets auxquels elle prenait part :

> « Par contre, les contraintes avec lesquelles j'ai dû faire, c'est que je travaillais dans un service informatique, [...] je devais toujours faire face à une contrainte technique qui va être celle d'un collègue qui me dit que "non, là ça représenterait un volume trop important de données", qu'il n'est pas forcément pour les données en triplets... Donc je devais avancer en ayant en tête les contraintes techniques que mes collègues informaticiens pouvaient avoir autour de moi. »

[124] Agence des Systèmes d'Information Partagés de Santé

Des chercheurs plus pointilleux et exigeants

Les chercheurs et professionnels ont parfois des difficultés à choisir le niveau de détail le plus adapté pour l'ontologie. Les chercheurs pourraient être jugés trop pointilleux par les autres professionnels, et leur point de vue, trop théorique. Les professionnels rechercheraient davantage la simplicité.

Tout d'abord, l'ontologie Socle et celle pour le projet ICODA devaient au départ être similaires. Pourtant, au fil du temps, ces ontologies se sont éloignées puisque leurs domaines et leurs objectifs ne sont pas tout à fait les mêmes. Henri-Maxime explique : « Le sujet d'ICODA, ce sont les conseillers municipaux mais ce n'est pas notre sujet. Du coup, on a des objectifs un peu différents. » En effet, Socle est une ontologie de haut niveau qui servira à terme à valoriser tous les contenus des journaux du groupe ; l'ontologie ICODA se limite au cas d'usage des démissions des conseillers municipaux. Ainsi, selon Henri-Maxime Suchier, le travail des deux équipes s'est un peu scindé : « On s'est dit "il faut qu'on fasse cette ontologie" et on est tous plus ou moins partis chacun de notre côté pour commencer à modéliser ça. Donc, à un moment, on s'est retrouvés et chacun de notre côté, on avait travaillé sur une ontologie ». Marie-Paule Cochet met en avant les vitesses de développement hétérogènes des ontologies ; pour elle, les chercheurs vont quelquefois trop rapidement et les professionnels de Ouest-France ont plus de difficultés à suivre. Elle explique « on ne peut pas prendre l'intégralité de ce que [les chercheurs du LIRMM] ont décidé parce que c'est un peu trop tôt pour nous », ou encore :

> « Ils sont en avance de phase sur nous, donc je pense que c'est trop tôt et que ça ne change rien si on ne prend qu'un sous-ensemble de ce qu'ils ont fait. C'est vrai qu'ils ont envie d'aller vite parce qu'ils ont d'autres sujets derrière. Pour moi, ça va trop vite par rapport à ça. »

En plus de ces contraintes organisationnelles, décider du niveau de détail de l'ontologie prend du temps et cela ne convainc pas toujours toutes les parties. Pour Marie-Paule Cochet, « il fallait qu'[ils partent] sur la même chose qu'ICODA mais il y a des choses qu'[ils vont] avoir du mal à faire valider », notamment aux autres informaticiens. Les concepts doivent rester « lisibles » et simples pour qu'une première version de l'ontologie soit implantée au sein de l'annuaire. Cependant, certains concepts créés par les chercheurs deviennent trop complexes et « trop peu transparents » pour que l'ontologie soit utilisée simplement au sein du service BDC. Henri-Maxime Suchier partage ce point de vue : pour lui, « les gens de Montpellier avaient quelque chose de beaucoup plus fin et de beaucoup plus détaillé » et ils « vont aller chercher très loin dans les détails sur des trucs très précis. » Il se rend compte que les professionnels de la BDC n'ont « non

seulement pas besoin d'un tel niveau de détail et qu'en plus, ça pourrait polluer au final le modèle de données. »

Aussi, pour Michel Chein, « il faut faire de nombreux choix » des deux côtés pour tenter de contenter les parties. De plus, il recherche lui aussi la simplicité puisque l'ontologie « est destinée aux utilisateurs » : par exemple, son équipe va essayer de « limiter l'héritage multiple, c'est-à-dire une classe qui est sous-classe de plusieurs classes ». Ainsi, on peut nuancer les remarques des professionnels de la BDC qui mettent en avant la trop grande complexité et le manque d'utilisabilité de l'ontologie ICODA.

Lorsque nous interrogeons Jean Charlet à propos des points précis sur lesquels il avait porté son attention lors de l'élaboration de l'ontologie LERUDI, il répond que l'ontologie n'est pas « simple à utiliser », les concepts sont complexes et le langage est surtout conçu pour que la machine puisse raisonner sur les connaissances. Sa réponse dépend potentiellement du domaine de connaissances, le médical, pour lequel les ontologies sont souvent plus conséquentes et granulaires.

Ainsi, les chercheurs Michel Chein et Jean Charlet sont plus exigeants que les professionnels, et plus particulièrement sur les niveaux de détail. Cela peut occasionner des difficultés et incompréhensions pour certains professionnels qui souhaiteraient que l'ontologie soit plus simple et davantage compréhensible. On peut tout de même nuancer cette observation puisque Michel Chein souhaite lui aussi développer une ontologie la plus simple possible pour les utilisateurs, même si le niveau de granularité reste encore trop élevé pour les professionnels de Ouest-France.

Une forte exigence d'opérationnalisation

Lors de plusieurs entretiens, les professionnels ont mis en lumière l'impératif d'opérationnalisation de l'ontologie et de l'application ; impératif qui conditionne de nombreux choix de conception afin de créer une ressource plus pratique que théorique.

Ils insistent tout d'abord sur les contraintes de production auxquelles ils doivent faire face. Marie-Paule Cochet met en évidence les contraintes temporelles ; elle pense que « c'est un travail prospectif, plus sur le long terme et [ils ont besoin] d'avoir quelque chose qui tienne en tant que production.» Il est très important pour elle que l'ontologie puisse être facilement utilisée, c'est une de ses principales préoccupations. Elle explique :

> « Donc il faut valider qu'on respecte bien les contraintes de production...que le modèle soit viable tout en étant évolutif, donc compatible avec tous les sujets qu'on a en parallèle, le maritime, le commerce...Il faut qu'on puisse le mettre en production et qu'on ne fasse pas une usine à gaz ! »

Florence Amardeilh a aussi dû surmonter de nombreuses contraintes liées à la future opérationnalisation de l'ontologie. Sylvie Desprès et elle ont tenté de réutiliser l'ontologie de la cuisine, Taaable, déjà mise en place par des chercheurs. Pourtant, « le projet Taaable était juste un projet universitaire, académique, les exigences industrielles portées par SEB n'étaient pas les mêmes. » Ainsi, une ontologie élaborée dans un contexte universitaire pouvait difficilement être utilisée pour un autre projet pour lequel les exigences d'opérationnalisation étaient très fortes. Les objectifs n'étaient pas les mêmes et certaines informations plus pratiques, comme les valeurs nutritionnelles, manquaient dans l'ontologie source.

Elle a travaillé en collaboration avec de nombreux services, le projet était découpé en *work packages,* donc en plusieurs activités interdépendantes : construction de l'ontologie et de l'interface utilisateurs, opérationnalisation, etc. Dans le milieu industriel, elle a aussi fait face à des contraintes de production, moins présentes dans un contexte universitaire : « Il y a vraiment aussi le contexte industriel de SEB. C'est-à-dire que l'ontologie puisse répondre à certains besoins, en termes de volume, en termes de rapidité. » Elle devait convaincre les professionnels de SEB, « les gens du business » et les partenaires, de l'utilité de cette ontologie. Cette ressource devait donc être « la plus complète possible », « la plus utile possible », « la plus opérationnalisable possible », « la plus facile à manipuler » et « facile à maintenir aussi ». En effet, ces qualités précises devaient être respectées afin de « vraiment convaincre SEB que [l'utilisation d'une ontologie] était efficace et la meilleure solution pour eux. »

Florence Amardeilh pense qu'il est dommage que les chercheurs restent parfois trop cantonnés à la théorie. Elle n'approuve pas le fait qu'ils restent quelquefois trop rigides et trop centrés sur les détails :

> « Voilà, il y a vraiment un point qui bloque aujourd'hui et qui gêne peut-être l'acceptation des ontologies, les chercheurs ont leur approche chercheur très théorique, mais quand on fait une application, des fois on ne peut pas être super exhaustifs ou super nickel sur tous les angles de la connaissance qu'on souhaite modéliser. À un moment, il faut arrondir les angles. […] Mais voilà, ça reste compliqué parce que les chercheurs veulent un monde parfait, bien modélisé, et les industriels veulent juste quelque chose qui marche. »

Avec la chercheuse Sylvie Desprès, la collaboration s'est toutefois très bien passée puisqu'elles ont toutes deux compris « les exigences d'une onto ». Leurs deux approches étaient complémentaires : « Après, parfois, elle avait le côté modélisation connaissances pures, théoriques. Moi, j'avais plus le côté "oui, mais non, d'un côté pratique, ça ne va pas pouvoir se faire comme ça, car il y a telle contrainte technique…" On en rediscutait et on arrivait toujours à un compromis. »

Néanmoins, une frontière très poreuse entre théorie et pratique

Cependant, des exigences élevées d'opérationnalisation sont également présentes dans des projets « davantage universitaires ». Nous allons alors nous rendre compte que l'écart est très ténu entre professionnels et chercheurs, entre projets de recherche et projets industriels. Cela remet alors en cause la frontière rigide entre professionnels et chercheurs que nous avons instaurée.

Au départ, pour Florence Amardeilh, le projet OFS, « c'était quand même un projet de recherche. » Il a donc été nécessaire de convaincre les professionnels de SEB de l'utilité de l'ontologie afin que le projet soit industrialisé. Il est parfois difficile de mêler ces deux aspects, pourtant, chez Mondeca, l'entreprise où Florence était Directrice Recherche & Développement, les milieux de la recherche et de l'industrie se confondent :

> « C'est ça, souvent, dans le domaine du Web sémantique, un des problèmes aujourd'hui, c'est que ça reste beaucoup un sujet très théorique avec des chercheurs...ça a du mal à sortir sur des cadres applicatifs, après il y en a. Mondeca a maintenant presque vingt ans d'expérience et ils ne font pas que de la recherche, il y a deux tiers de projets clients. Il y a pas mal de gens qui s'y intéressent et qui font des applications concrètes. »

À Ouest-France, dans le cadre de certains projets, ces deux approches sont également entremêlées. Comme l'explique Nadia Fafi, pour le projet en collaboration avec les chercheurs de Montpellier :

> « Il y a une partie recherche où on essaie de développer des outils qui vont nous permettre d'avancer sur plein de problèmes rencontrés dans la presse. Donc là, ICODA c'est vraiment un projet de recherche où on a un partenariat avec des chercheurs. […] Mais, de plus en plus, on s'approche quand même d'une utilité, du fait de pouvoir utiliser ce qui ressort de ces recherches. »

Jean Charlet aborde brièvement ce même sujet ; le projet LERUDI va ensuite être industrialisé en relation avec d'autres professionnels. Il explique : « on voyait rapidement que certains documents étaient mal indexés et on vérifiait avec les industriels si c'était la faute du système, si le système de reconnaissance fonctionnait mal..ou si c'était la faute de l'ontologie à qui il manquait des classes. » C'est la même démarche que met en place Michel Chein : ils font « tout » dans leur équipe et souhaitent « faire la théorie, les outils et aller jusqu'aux applications réelles. »

Ainsi, dans le cadre de projets de recherche, en apparence théoriques et menés par des chercheurs, des ontologies vont être mises en production. Elles doivent donc répondre à des exigences de production, tout comme les ontologies conçues dans un cadre davantage « professionnel ».

En outre, nous avons compris que quelques personnes de notre échantillon ont des profils hybrides ; cela nous amène à remettre en cause la catégorisation stricte entre « professionnels » et « chercheurs ». Effectivement Henri-Maxime Suchier Florence Amardeilh ont tous deux obtenu un doctorat en informatique ; Henri-Maxime a également débuté sa carrière dans l'Enseignement supérieur. Ils se sont ensuite tournés vers le milieu industriel, donc ces deux domaines sont tout à fait compatibles. Comme Florence le dit elle-même, elle est « un peu à cheval entre les deux mondes. » Ça lui est bénéfique puisqu'elle fait facilement le lien entre la théorie et le milieu industriel : « Donc être à la croisée des deux mondes, c'est intéressant car on peut faire le pont entre les avancées théoriques et comment on les met en pratique pour répondre à de vrais besoins qui viennent des utilisateurs. »

Pour Michel Charlet, au LIRMM, les fonctions sont très mêlées et toutes les personnes « ont plusieurs casquettes. » Ainsi les rôles qu'il a auparavant mentionnés (experts du domaine, ingénieurs de la connaissance, informaticiens) « sont joués même s'il n'y a que deux personnes dans l'équipe. » Il faut s'adapter aux contraintes organisationnelles et être polyvalent.

Conclusion

Les ontologies ont donc de nombreuses utilités pour aider à la recherche d'information : aide à l'indexation, à la formulation et à l'expansion de requêtes, présentation améliorée des résultats dans l'application, etc. Pourtant, elles sont coûteuses en ressources financières et humaines ; elles restent souvent difficiles à mettre en place. Des méthodologies de construction bien cadrées et précises ont été présentées dans la littérature scientifique dès la fin des années 1990. Beaucoup sont séquentielles, même si certaines mentionnent que de nombreuses itérations sont nécessaires pour construire la ressource la plus adaptée aux fonctionnalités de l'application finale.

Nous nous demandions si ces méthodologies étaient vraiment appliquées ou si les professionnels concoctaient davantage « des méthodes maison ». Nous souhaitions également interroger le lien entre construction d'une ontologie et prise en compte des attentes des futurs utilisateurs : concrètement, ces attentes étaient-elles toujours prises en compte et de quelle manière les concepteurs les récoltaient ? Une scission semblait apparaître entre pratiques des chercheurs, plus cadrées et minutieuses, et pratiques des professionnels, plus simples et instinctives. Nous nous sommes également interrogée sur la véracité de cette hypothèse.

En interrogeant des professionnels et chercheurs ayant conçu des ontologies dans différents domaines - presse, médical, industrie - nous nous sommes rendu compte que les méthodologies de construction officielles restent peu utilisées. Quelquefois, les grandes lignes de ces méthodes sont tout de même suivies ; les professionnels réutilisent également - sans en avoir conscience - de grandes étapes de travail communes empruntées à des méthodologies diverses. Ainsi, beaucoup de concepteurs s'appuient sur des ressources préexistantes et évaluent continuellement leurs ontologies. Les manières de construire ces ressources sont en général plus libres et itératives. De nombreux tâtonnements rythment le processus de création : les méthodes utilisées ne sont pas séquentielles.

L'adaptation de l'ontologie aux publics cibles est une constante dans tous les projets : la quasi-totalité de notre échantillon a amorcé des actions pour prendre en compte les usages de ceux-ci. Pourtant, cette relation avec les utilisateurs potentiels semble difficile à pérenniser sur le long terme.

Il est nécessaire que différents professionnels (documentalistes, ingénieurs de la connaissance, informaticiens, experts) travaillent ensemble pour donner naissance à une ontologie juste et simple à utiliser. Dans le cadre de certains projets, cette collaboration pose néanmoins certaines difficultés : parfois, ces personnes n'ont pas les mêmes objectifs, compétences ou pratiques. En

outre, certains reprochent aux chercheurs leur niveau d'exigence trop élevé. Malgré ces différences, toutes les personnes de notre échantillon concourent à un but commun : construire des ontologies pour des applications qui seront mises en production à plus ou moins long terme. Il n'est pas vraiment pertinent de différencier de manière stricte « professionnels » et « chercheurs », d'autant plus que de nombreux professionnels ont un profil hybride. Ils visent tous à construire des ontologies opérationnalisables, adaptées aux différents domaines de connaissances et aux fonctionnalités de l'application cible.

Les choix de modélisation, les degrés de granularité et de couverture du domaine dépendent donc davantage des modalités du projet, et non de la profession des personnes développant la ressource. Par exemple, les projets LERUDI et OFS sont plus exigeants, rassemblent un nombre de collaborateurs plus important, les ontologies sont plus étendues en terme de concepts. Il est donc compréhensible que des méthodologies officielles et des niveaux de formalisation plus élevés soient utilisés afin de mieux cadrer ces projets. De plus, les concepteurs sont souvent contraints par le temps. Lorsque l'ontologie doit être construite en quelques mois, ils auraient tendance à développer leur propre méthode de travail, alors qu'ils choisiraient une méthodologie cadrée s'ils ont davantage de temps. La formation et l'expérience des professionnels dans ce domaine constituent aussi des points cruciaux : généralement, plus les personnes sont formées et ont l'habitude de créer des ontologies, et plus elles utilisent des méthodologies consignées dans la littérature scientifique.

Les méthodologies de construction évoluent avec l'appropriation des technologies du Web sémantique par les professionnels. Ces méthodes nécessitent d'être itératives et ces itérations sont rendues possibles par les outils du Web sémantique. De plus, la nouvelle recommandation officielle du W3C[125], *Shapes Constraint Language* (SHACL) pourrait faciliter la construction d'ontologies en ajoutant des contraintes à OWL. Effectivement, avec OWL, il est parfois difficile de mettre en place des contraintes concernant les raisonnements, puisque l'on raisonne dans un « monde ouvert » : tout ce qui n'est pas faux est vrai. SHACL permet de raisonner dans « un monde fermé », où est plus facile de faire des inférences : il faut indiquer que quelque chose est vrai pour pouvoir raisonner dessus[126]. En outre, cette conception est davantage partagée par les industriels qui se saisissent de plus en plus des ontologies pour créer des applications fiables et adaptées aux besoins des utilisateurs. Ainsi, avec l'utilisation de SHACL, de nouvelles méthodes de construction encore plus itératives pourraient voir le jour, permettant au domaine de l'Ingénierie des connaissances de se renouveler.

[125] Parue le 20 juillet 2017

[126] http://spinrdf.org/shacl-and-owl.html

Bibliographie

ABDERRAHIM Mohammed Alaeddine. *Exploitation des Ontologies dans les Systèmes de recherche d'information Arabes* [en ligne]. Thèse pour l'obtention du grade de Docteur en sciences. Université Aboubakr Belkaïd, Tlemcen, 2016. 94 p. [Consulté le 25/11/2017]. Disponible à l'adresse : http://dspace.univtlemcen.dz/bitstream/112/8632/1/Exploitation_des_Ontologies_dans_les_Systemes_de_recherche_dinformations_Arabes.pdf

AUSSENAC-GILLES Nathalie, HERNANDEZ Nathalie, BAZIZ Mustapha. Ontologies pour la recherche d'information, importance de la dimension terminologique. *In* EL HADI Widad Mustafa. *Terminologie et accès à l'information.* Paris : Hermès Science Publications, 2006. 262 p. (Traité des sciences et techniques de l'information). ISBN 2-7462-1295
Chapitre également disponible en ligne à l'adresse : https://www.irit.fr/publis/IC3/Aussenac-traiteWidad_2006.pdf

AUSSENAC-GILLES Nathalie. Le Web sémantique, quel renouvellement pour la recherche d'information ?. In : BOUGHANEM Mohand, SAVOY Jacques. *Recherche d'information : état des lieux et perspectives*. Cachan: Lavoisier, 2008. p. 97-132. ISBN 978-2746220058.
Chapitre également disponible en ligne à l'adresse : https://www.irit.fr/publis/IC3/aussenac-LivreBoughanem2007_18janv.pdf

BACHIMONT Bruno, *Ingénierie des connaissances et des contenus. Le numérique entre ontologies et documents.* Paris : Hermes science publications-Lavoisier, 2007. 279 p. Collection Science informatique et SHS. ISBN 978-2746213692. Également disponible en ligne à l'adresse : https://stph.scenaricommunity.org/nf29/res/2007Bachimont_IngenierieDesConnaissancesEtDesContenus.pdf

BACHIMONT Bruno *et al*..Enjeux et technologies : des données au sens, *Documentaliste-Sciences de l'Information*, 2011, Vol. 48, n°4, p. 24-41.
Également disponible en ligne à l'adresse : http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2011-4-page-24.html

BOURIGAULT Didier, AUSSENAC-GILLES Nathalie, CHARLET Jean. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'intelligence Artificielle*, 2002, Vol. 10, n° 10, p. 1-10

Également disponible en ligne à l'adresse :
https://pdfs.semanticscholar.org/c6f0/9760da950e577ebdb78eb1764b62c8166c03.pdf

BOURIGAULT Didier, AUSSENAC-GILLES Nathalie. Construction d'ontologies à partir de textes. *Conférence TALN 2003*, Juin 2003 (Batz-sur-Mer, France) [en ligne]. p. 11-14. [Consulté le 25/11/2017]. Disponible à l'adresse : http://www.atala.org/taln_archives/TALN/TALN-2003/taln-2003-tutoriel-002.pdf

CHARLET Jean, REYNAUD Chantal, TEULIER Régine. Ingénierie des connaissances pour les systèmes d'information. In : CAUVET Corinne (éd.). *Ingénierie des systèmes d'informations.* Paris : Hermès, 2001. 353 p. ISBN 978-2-7462-0219-1.Chapitre également disponible en ligne à l'adresse : https://abiteboul.com/gemoReports/GemoReport-288.pdf

CHARLET Jean. *L'ingénierie des connaissances : développement et application pour la gestion de connaissances médicales.* Mémoire d'Habilitation à diriger des recherches. Université Pierre et Marie Curie, Paris, 2002. 143 p. [Consulté le 25/11/2017]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-00006920/document

CHAUMIER Jacques. Les ontologies. Antécédents, aspects techniques et limites. *Documentaliste-Sciences de l'Information*, 2007, Vol. 44, n°1, p. 81-83.
Également disponible en ligne à l'adresse : https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-81.htm

DIENG-KUNTZ Rose. *Le Web Sémantique pour la Gestion des Connaissances.* Diaporama. EGC 2005. Paris, 2005. [Consulté le 25/11/2017].
Disponible à l'adresse : www.egc.asso.fr/sdoc-60-egc05_conf_inv_Web_semantique.pdf

DRAME Khadim. *Contribution à la construction d'ontologies et à la recherche d'information : application au domaine médical* [en ligne]. Thèse pour obtenir le grade de Docteur en informatique et santé. Université de Bordeaux, 2014. 187 p. [Consulté le 20/02/2017]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-01166042/document

FOURCASSIER Eric. *Des ontologies pour les humanités : Les ontologies de domaine à l'épreuve des humanités numériques.* Mémoire pour l'obtention du Master esDOC. Université de Poitiers, 2016. 97 p.

FRONTERE Mikhail. *Assistance intelligente à la recherche d'information : élaboration d'un projet de moteur de recherche au service de la connaissance dans l'organisation* [en ligne]. Mémoire pour obtenir le Titre professionnel « Chef de projet en ingénierie documentaire ». INTD-CNAM, 2015. 170 p. [Consulté le 25/11/2017].
Disponible à l'adresse : https://memsic.ccsd.cnrs.fr/mem_01309438/document

GAGNON Olivier, *Indexation de documents Web à l'aide d'ontologies*, Mémoire pour obtenir le diplôme de Maîtrise Es Sciences Appliquées (Génie informatique). École polytechnique de Montréal, 2013. [Consulté le 25/11/2017]. 98 p.
Disponible à l'adresse : https://publications.polymtl.ca/1131/1/2013_OlivierGagnon.pdf

GANDON Fabien. Ontologies informatiques. *Interstices.info*. [En ligne]. 22 mai 2006. [consulté le 25/11/2017].
Disponible à l'adresse : https://interstices.info/jcms/c_17672/ontologies-informatiques

GANDON Fabien, DIENG-KUNTZ Rose. Ontologie pour un système multi-agents dédié à une mémoire d'entreprise. *IC'2001, Ingénierie des Connaissances, plateforme AFIA'2001,* Juin 2001, (Grenoble, France) [en ligne]. [Consulté le 25/11/2017].
Disponible à l'adresse : https://hal.inria.fr/hal-01145808

GANDON Fabien, FARON-ZUCKER Catherine, CORBY Olivier. *Le Web sémantique : comment lier les données et les schémas sur le Web ?*. Paris : Dunod, 2012. 206 p. ISBN 978-2-10-057294-6

HERIGAULT Myriam. *Moteur de recherche d'entreprise : déploiement du moteur sémantique Exalead à la R&D de Diagnostica Stago* [en ligne]. Mémoire pour obtenir le Titre professionnel « Chef de projet en ingénierie documentaire ». INTD-CNAM, 2012. [Consulté le 25/11/2017].
Disponible à l'adresse : https://memsic.ccsd.cnrs.fr/mem_00803358/document

HERNANDEZ Nathalie *et al*. *RI et Ontologies – État de l'art 2008* [en ligne]. Rapport interne. IRIT, Université de Toulouse, 2008. 45 p. [Consulté le 25/11/2017].
Disponible à l'adresse : https://www.irit.fr/publis/SIG/2008_RA-14-FR_HHMR.pdf

LAGARDE ., RENAUDINEAU C.. *Développement d'un moteur de recherche sémantique : une contribution au projet Ethnosiris dédié à la préservation du patrimoine vendéen.* Mémoire de Master 1 en Informatique. Université de Nantes, 2009. [Consulté le 25/11/2017].
Disponible à l'adresse : http://ethnosiris.com/memoireTER2009.pdf

LUSTREMENT Amandine. *Thésaurus et Web sémantique : quelles problématiques de mise en œuvre ?.* Mémoire pour l'obtention du Master esDOC. Université de Poitiers, 2016. 89 p., p.11

MAFOKOUA TCHIGUI Ingrid Pamela. Recherche d'informations dans le Web sémantique. *Supinfo.com* [en ligne]. 30 octobre 2016. [Consulté le 25/11/2017].
Disponible à l'adresse : https://www.supinfo.com/articles/single/3373-recherche-informations-clans-Web-semantique

NOY N.F., McGuinness D.L., *Développement d!une ontologie 101 : Guide pour la création de votre première ontologie*, Université de Stanford, 2005. 26p. Traduit de l'anglais par Anila Angjeli, BnF, Bureau de normalisation documentaire. [Consulté le 20/06/2017].
Disponible à l'adresse : http://www-limics.smbh.univ-paris13.fr/GBPOnto/data/documents/2010/6babahamedcontribution.pdf

PAQUETTE Gilbert. *Introduction aux technologies sémantiques* [en ligne]. Cours Technologies sémantiques pour la gestion des connaissances. Université TELUQ (Québec), 2015. 25 p. [Consulté le 25/11/2017]. Disponible à l'adresse : http://inf6070.teluq.ca/teluqDownload.php?file=2013/07/INF6070_M1_a5_ApplicationTechnologiesSemantiquesGC.pdf

PAQUETTE Gilbert. *Applications des technologies sémantiques à la gestion des connaissances* [en ligne]. Cours Technologies sémantiques pour la gestion des connaissances. Université TELUQ (Québec), 2015. 27 p. [Consulté le 25/11/2017].
Disponible à l'adresse : http://inf6070.teluq.ca/teluqDownload.php?file=2013/07/INF6070_M1_a5_ApplicationTechnologiesSemantiquesGC.pdf

PICARD Anne-Claire (Le). *Le cycle de vie d'une ontologie : évaluation de l'ontologie du domaine de la Toxicologie Nucléaire.* [en ligne]. Mémoire pour obtenir le Titre professionnel « Chef de projet en ingénierie documentaire ». INTD-CNAM, 2014. 174 p. [Consulté le 25/11/2017].
Disponible à l'adresse : https://memsic.ccsd.cnrs.fr/mem_01128938

REYMONET Axel, THOMAS Jérôme, AUSSENAC-GILLES Nathalie. Modélisation de Ressources Termino-Ontologiques en OWL. *Journées Francophones d'Ingénierie des Connaissances,* Juillet 2007 (Grenoble, France). Cépaduès Éditions, p.169-180, 2007.
Également disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-00365888

REYMONET Axel, THOMAS Jérôme, AUSSENAC-GILLES Nathalie. Ontologies et Recherche d'Information : une application au diagnostic automobile. *21èmes Journées Francophones*

d'Ingénierie des Connaissances, Juin 2010 (Nîmes, France) [en ligne].École des Mines d'Alès. p. 283-294. [Consulté le 25/11/2017].
Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-00487737/document

SY Mohameth-François. *Utilisation d'ontologies comme support à la recherche et à la navigation dans une collection de documents* [en ligne]. Thèse pour l'obtention du grade de Docteur en Informatique. Université de Montpellier II, 2012. 135 p. [Consulté le 25/11/2017].
Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-00822516

TRIOU Frédéric, PICAROUGNE Fabien, BRIAND Henri. *Apport du Web sémantique dans la réalisation d'un moteur de recherche géo-localisé à usage des entreprises*, EGC 2007. Paris, 2007. 12 p. [Consulté le 25/11/2017].
Disponible à l'adresse : http://editions-rnti.fr/render_pdf.php?p=1001307&p1

VUILLEQUEZ Jean-Yves. *Le moteur de recherche d'entreprise : quels enjeux organisationnels et technologiques ?* [en ligne]. Mémoire pour obtenir le Titre professionnel « Chef de projet en ingénierie documentaire ». INTD - CNAM, 2013. 125 p. [Consulté le 25/11/2017].
Disponible à l'adresse : https://memsic.ccsd.cnrs.fr/mem_00945629/document

ZACKLAD Manuel. Évaluation des systèmes d'organisation des connaissances. *Les Cahiers du numérique*, 2010, Vol.6, n°3, p. 133-166. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-les-cahiers-du-numerique-2010-3-page-133.htm

ZACKLAD Manuel, GIBOIN. Systèmes d'organisation des connaissances hétérogènes pour les applications documentaires. *Document numérique*, 2010, Vol. 13, n°2, p. 7-12. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-document-numerique-2010-2-page-7.htm

ZIDI Amir, *Recherche d'information dirigée par les interfaces utilisateurs : approche basée sur l'utilisation des ontologies de domaine* [en ligne]. Thèse pour l'obtention du grade de Docteur en Sciences et Technologies, mention Informatique. Université de Valenciennes et du Hainaut Cambrésis, 2015. 118 p. [Consulté le 25/11/2017].
Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-01356087/document

Table des annexes

Annexe n° 1 : *E-mail* de prise de contact envoyé à notre échantillon

E-mails envoyés entre les 14 et 20 mai 2018 (sauf à Marie-Paule Cochet, Nadia Fafi et Henri-Maxime Suchier travaillant au sein du service BDC à Ouest-France)

Une partie de l'*e-mail* est personnalisée selon l'interlocuteur et le projet sur lequel nous avions fait des recherches : ici, nous trouvons le message envoyé à Jean Charlet.

Objet

Mémoire étudiant : Demande d'informations concernant l'élaboration d'ontologies

Bonjour,

Je suis étudiante en master 2 EsDOC (Documentation Bibliothèques Veille) à l'Université de Poitiers, j'effectue actuellement mon stage de fin d'études à Ouest-France au sein du service Banque de Contenus, dont Michel Le Nouy est responsable. J'ai suivi des cours sur les technologies du Web sémantique avec Thomas Francart, qui m'a suggéré de faire appel à vous.

Dans le cadre de ce stage, je travaille sur la construction d'une ontologie qui modélisera le domaine de la vie des commerces. En parallèle, je centre mon mémoire étudiant sur la construction d'ontologies comme systèmes d'organisation des connaissances, et plus particulièrement sur les différences qui peuvent apparaître entre formalisme et pratiques professionnelles lors de cette phase de construction. Après avoir rédigé une première partie davantage théorique présentant un état de l'art de mon sujet, je mettrai en place une expérimentation qui permettra d'apporter une réponse à mes hypothèses de départ.

En effet, en échangeant avec des documentalistes et informaticiens de mon service qui travaillent sur une ontologie de haut niveau en collaboration avec des chercheurs de l'INRIA, j'ai constaté qu'il y avait parfois des différentiels entre l'ontologie qui était proposée par des enseignants-chercheurs et celle qui était effectivement mise en place dans un cadre davantage professionnel. La question de l'élaboration d'une ontologie adaptée aux usages des publics cibles est également centrale dans ma réflexion.

Je me demande si les objectifs des chercheurs et « professionnels » concernant la construction d'une ontologie sont similaires - même si j'ai bien conscience que dans beaucoup de projets, ces fonctions se confondent. Par exemple, une ontologie pensée par des documentalistes d'une

entreprise pourrait être davantage centrée sur les usages futurs de l'ontologie, alors que des chercheurs pourraient viser davantage l'exhaustivité. Les niveaux de granularité ou de couverture du domaine à modéliser pourraient alors varier. L'ontologie peut passer par plusieurs versions, sans que l'une d'entre elles soit forcément « moins juste » que les autres.

Ainsi, dans le cadre de projets actuels ou passés, avez-vous été confronté à des questions/problématiques similaires ? En ce moment, j'ai constaté que vous travailliez notamment sur une modélisation des maladies neurodégénératives pour le projet PARON.

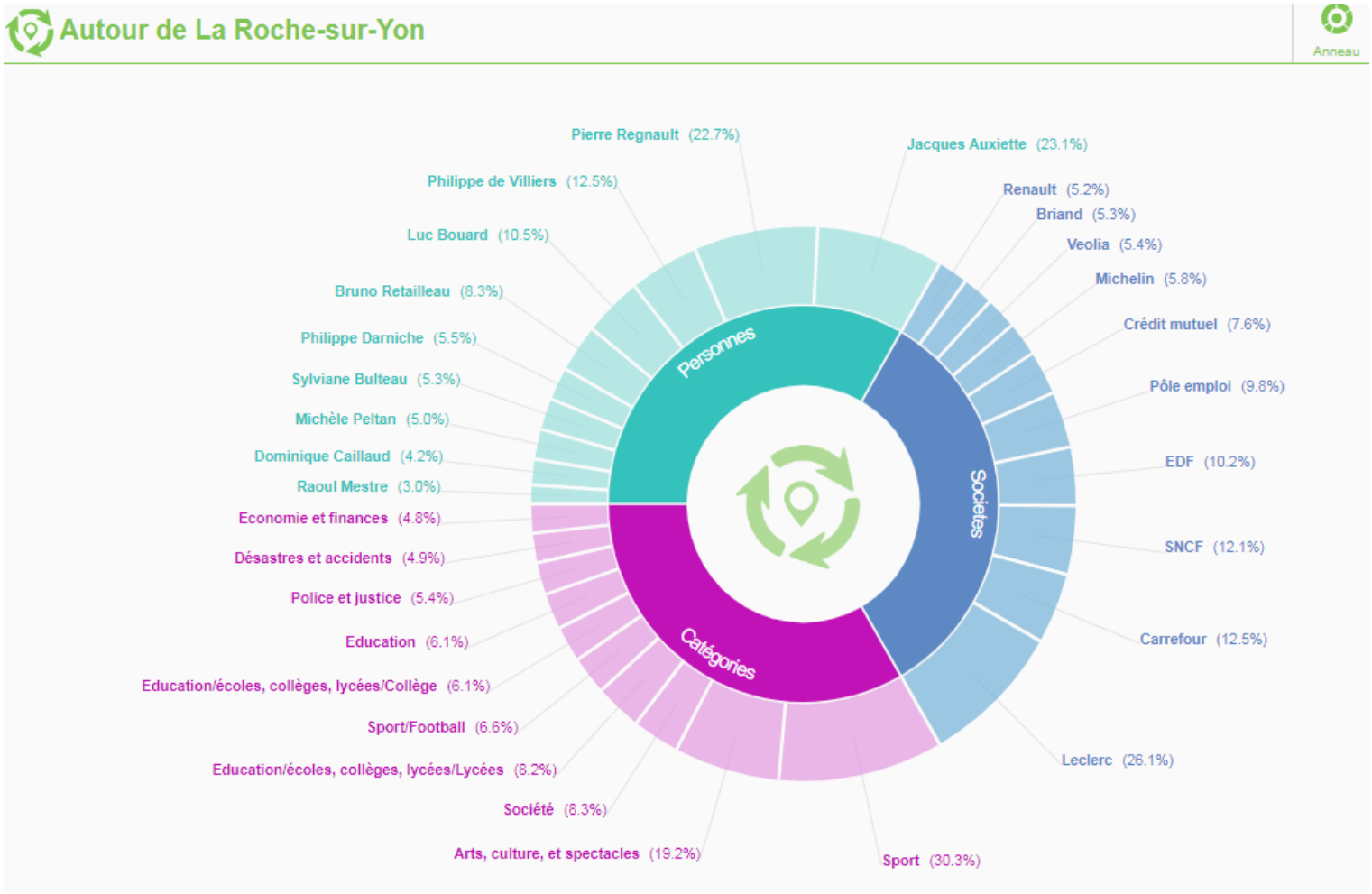
Par exemple, avez-vous choisi une méthodologie déjà employée au sein d'autres projets ou avez-vous mis en place une autre méthodologie ? Avez-vous organisé une collaboration entre chercheurs et professionnels (documentalistes, informaticiens, professeurs...) lors de la phase de modélisation de l'ontologie ? Comment a été pensée l'articulation entre les usages du public qui utilisera l'application dans laquelle l'ontologie sera implantée et la création de l'ontologie ?

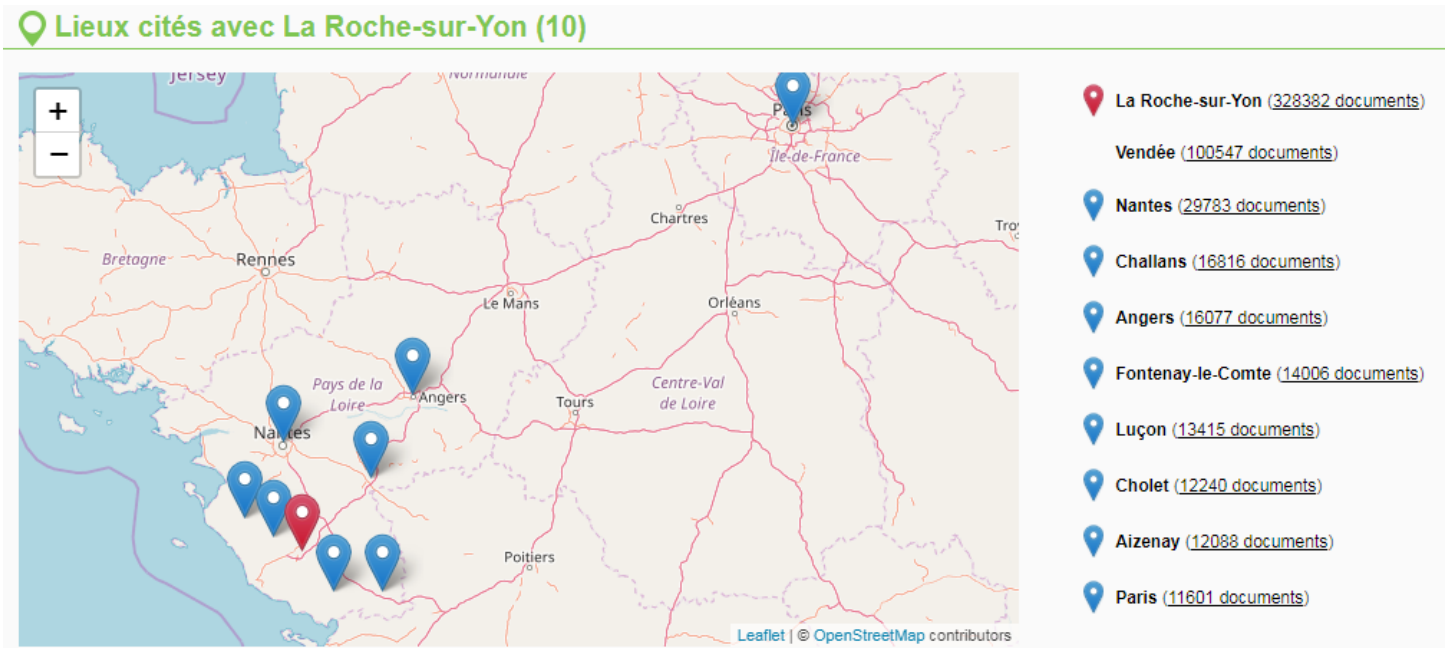
D'avance, je vous remercie de votre réponse et reste disponible pour davantage de précisions,

Cordialement,

Annexe n° 2 : Visualisation de l'entité « La Roche-sur-Yon » sur Troove

Le 09/07/2018





Annexe n° 3 : Première version de la grille d'entretien (au 02/05/2018)

1- Formation et parcours

- Études suivies, diplômes obtenus
- Résumé du parcours professionnel
- Genre d'établissement où la personne a travaillé et travaille aujourd'hui, type de services, niveau de responsabilité, ancienneté

2- Le projet (faisant appel à une ontologie)

- Rapide présentation du projet, grandes étapes, durée
- Le projet aujourd'hui : état d'avancement, finalité, avenir
- Rôle au sein de ce projet : fonction, niveau de responsabilité
- Pour vous, une ontologie, c'est... ?
- Personne/corps de métier à l'origine de la décision d'utiliser une ontologie pour ces tâches, à quel moment cela a été décidé ? Consensus ou non dans le choix d'utiliser une ontologie ?

3- La construction et le peuplement de l'ontologie

- Rôle dans la construction et le peuplement de l'ontologie : construction ? Pilotage du projet ? Conseil ? Exécution ?
- Pour vous, que signifie une méthodologie de construction d'ontologie ? Qu'est-ce que cette expression recoupe ?
- Méthode utilisée pour construire l'ontologie: appui sur des méthodes de référence ? Méthode propre au projet ? Pas de méthode spécifique ?
- Méthode utilisée pour peupler l'ontologie
- Approche privilégiée : exhaustivité, simplicité, complétude, minimalisme… ?
- Articulation avec les usages/attentes du public à qui le projet est destiné : comment ces usages ont été mis en lumière ? Qui a fait remonter ces attentes ? Choix de créer une

ontologie répondant seulement à cette fonction/ces usages ou volonté d'élargir l'ontologie pour des besoins ultérieurs ?

- Quelles relations/concepts/instances, etc. ont été choisis pour aider à la recherche d'information ?
- Outils utilisés : logiciels ? outils de gestion de projet ? *mindmap* ?, etc.
- Les grandes étapes de ce projet

4- Relation entre les différents professionnels ?

- Collaboration avec d'autres professionnels : avec des personnes de quelles professions la personne a été mise en contact et a travaillé ? quel niveau de collaboration (fort, liens ponctuels ?...), quel genre d'échanges et canal ? (conseils au téléphone, réunions projet en présentiel…). Comment le travail s'est déroulé ?
- Facilité ou non de la collaboration avec ces personnes : quelle facilité ? Quels apports ? Quelles difficultés ? Quels points bloquants ?

Annexe n° 4 : Grille d'entretien finale

1-Formation et parcours

- Études suivies, diplômes obtenus
- Résumé du parcours professionnel
- Genre d'établissement où la personne a travaillé et travaille aujourd'hui, type de services, niveau de responsabilité, ancienneté

2- Le projet faisant appel à une ontologie

- Rapide présentation du projet, grandes étapes, durée
- Le projet aujourd'hui : état d'avancement, finalité, avenir
- Rôle au sein de ce projet : fonction, niveau de responsabilité
- Objectifs/utilité de l'ontologie dans le cadre de ce projet
- L'ontologie au cœur du projet pour aider à la recherche d'information : rôle, utilité, quelles fonctions d'aide à la recherche d'information
- Personne/corps de métier à l'origine de la décision d'utiliser une ontologie pour ces tâches, à quel moment cela a été décidé ? Consensus ou non dans le choix d'utiliser une ontologie ?

3- La construction et le peuplement de l'ontologie

- Rôle dans la construction et le peuplement de l'ontologie : construction ? Pilotage du projet ? Conseil ? Exécution ?
- Les grandes étapes du projet, méthodologie de gestion de projet ?
- Outils utilisés : logiciels ? *mindmap* ?, logiciel d'élaboration d'ontologie ?, etc.
- Comment l'ontologie a été construite ? (Appui sur des méthodes de référence ? Méthode propre au projet ? Méthodologie type arrêtée pour la construction des ontologies ? Pas de méthode spécifique ?)
- Si une méthodologie préexistante a été utilisée ou si l'ontologie prend appui sur un outil déjà existant (ontologie, thésaurus, ontologie de haut niveau…) : les niveaux de

granularité, de détail, de couverture du domaine, de formalisation existants ou préconisés ont-ils été revus ou restent-ils constants ?

- Pour vous, que signifie une méthodologie de construction d'ontologie ? Qu'est-ce que cette expression recoupe ?
- Itérations ou non lors de la construction de l'ontologie ?
- Avez-vous utilisé d'autres ressources/référentiels pour vous aider à construire l'ontologie ? Si oui, comment les avez-vous re-exploités ?
- Méthode utilisée pour peupler l'ontologie
- Quels points de vigilance : sur quels points les porteurs du projet ont davantage porté attention ? (conformité avec les usages, facilité de prise en main ?)
- Approche privilégiée pour la construction et le peuplement de l'ontologie : exhaustivité, simplicité, complétude, minimalisme… ? Quel niveau de détail a été choisi ?
- Articulation avec les usages/attentes du public à qui le projet est destiné : comment ces usages ont été mis en lumière ? Qui a fait remonter ces attentes ? Choix de créer une ontologie répondant seulement à cette fonction/ces usages ou volonté d'élargir l'ontologie pour des besoins ultérieurs ?
- L'interface de l'application a-t-elle déjà été pensée ? Si oui, comment a été pensée l'articulation entre l'interface et les usages futurs ?

4- Relation entre les différents professionnels

- Collaboration avec d'autres professionnels : avec des personnes de quelles professions la personne a été mise en contact et a travaillé ?, quel niveau de collaboration (fort, liens ponctuels ?...), quel genre d'échanges et canal ? (conseils au téléphone, réunions projet en présentiel…). Comment le travail s'est déroulé ?
- Facilité ou non de la collaboration avec ces personnes : Quelle facilité ? Quels apports ? Quelles difficultés ? Quels points bloquants ?
- **Pour les professionnels** : avez-vous décidé de ne pas tenir compte de certaines remarques venant des chercheurs ou d'un autre corps de métier ? Quels choix avez-vous privilégiés ? Pourquoi ?
- **Pour les enseignants-chercheurs** : avez-vous été d'accord avec tous les choix des professionnels ? Aviez-vous préconisé d'autres choix ? Ont-ils été suivis ?

Annexe n° 5 : Entretien avec Marie-Paule Cochet (projet ontologie Socle)

5 juin 2018, 10h, durée : 1h

[...][127]

On va se centrer sur l'ontologie de haut niveau « Socle ». Est-ce que tu pourrais me présenter rapidement le projet et me dire quand il a commencé ?

On pensait aux ontologies depuis un bout de temps, on avait eu des formations avec Michel Le Nouy il y a deux/trois ans je pense, notamment avec Thomas Francart. Et on a commencé à essayer d'en mettre en œuvre avec le stage de Nadia l'année dernière, c'était vraiment axé sur la mer. On a vu les balbutiements et le potentiel qu'il y avait. Par rapport à ce qu'on avait déjà en production au niveau de l'annuaire (personnes, lieux et sociétés) et par rapport au projet de recherche qui s'est mis en place avec ICODA, on a dit qu'il fallait reprendre le travail de Nadia. On voyait bien la démarche, on a validé la démarche. On a validé comment, à partir de nos contenus, on pouvait commencer à utiliser une base de connaissances pour pouvoir extraire des informations un peu plus abstraites que ce qu'on fait avec un moteur de recherche. Donc la démarche, on l'a validée. Alors l'objectif, c'était de tout remettre à plat, de repartir un peu de zéro techniquement au niveau de l'ontologie, tout en ayant investi sur tout le travail qui avait déjà été fait. Mais on voulait avoir le socle : à la base, qu'est-ce qu'on a dans nos contenus ? Qu'est-ce qu'on modélise ? Sur ce socle-là, on devra pouvoir après, en fonction des différents domaines thématiques, affiner certains domaines, par exemple le maritime. On a le socle, on pourra rebrancher le maritime dessus et avoir quelque chose de cohérent. Donc après, n'avoir plus une vision thématique, mais quelque chose qui est généraliste en utilisant les concepts de base du Socle.

Et qui est à l'origine du projet ? de la décision d'utiliser une ontologie ?

Alors ICODA a démarré vite sur une ontologie, axée au départ sur le maritime, après ça a évolué vers la démission des élus municipaux. Et donc ils ont commencé à créer une ontologie et nous en parallèle, on a commencé à réfléchir aussi, sur les lieux notamment parce qu'on avait déjà des annuaires et des besoins qui sont assez présents, autour des communautés de communes notamment. Donc on avait besoin de redéfinir ces concepts-là. Je pense qu'ICODA est allé plus vite que nous et on aurait pu dire « on prend le travail d'ICODA » mais le travail d'ICODA est un travail

[127] Pour tous les entretiens, les parties concernant le parcours universitaire et professionnel ne sont pas retranscrites en annexe. Les éléments principaux sont synthétisés dans la 2e partie.

prospectif, plus sur le long terme et nous, on a besoin d'avoir quelque chose qui tienne en tant que production avec des échéances plus courtes. Vu nos contenus, très généralistes, on est obligés d'avoir un socle, des concepts communs solides, et après on peut enrichir. Je ne peux pas te dire la personne précise qui est à l'origine de ça. Mais en tout cas, c'est clair qu'on a plusieurs sujets : les *data journalists* ont évoqué plusieurs sujets autour de l'exploitation des contenus via la base de connaissances qui sont autour des manifestations, de la politique locale, les commerces, etc. Mais on ne peut pas travailler ces sujets en parallèle. Il y a des fondements en commun puisque c'est le même contenu donc on est obligés d'avoir cette ontologie de base. En plus, comme on a 90% d'informations vraiment locales dans nos contenus, donc beaucoup d'informations sur la géographie, les acteurs locaux, les entreprises locales...Sur presque tous les articles, on a un de ces éléments qui apparaît, donc il faut quelque chose de solide sur ça. Après, seulement, on pourra ajouter dessus, détailler l'ontologie.

Et tu dis que vous avez été en relation avec les *data journalists* pour ce projet ?

Oui alors Ouest-France met en place depuis l'année dernière une équipe un peu transversale au sein de la rédaction, le service *data journalism*. Ils sont cinq, il y a un journaliste, un infographiste, un Web designer, un informaticien et...je ne me rappelle plus. Ils analysent beaucoup l'*open data* avec une application dans l'ouest de la France pour écrire des articles. On a démarché, notamment au départ Nadia et Henri-Maxime, les journalistes pour leur montrer qu'il y a l'*open data* mais qu'il y a aussi en interne des contenus du groupe qui peuvent être supports et peuvent être utilisés pour ce genre de projet, donc ça peut être intéressant également. On a présenté ça à deux personnes de cette équipe-là. Tout de suite, ils ont bien vu l'intérêt qu'il pourrait y avoir et ils ont listé des thématiques qui pourraient les intéresser.

Et de quelle manière s'est déroulée cette collaboration avec les *data journalist*s ?

Alors, on avait fait deux réunions avec ce service, une où le service documentation était présent et l'autre pas. Lors d'une réunion, on a listé sept ou huit sujets qui pouvaient être intéressants. Par ailleurs, on avait demandé au service documentation de travailler avec les *data journalists*. Donc on a essayé de montrer qu'au sein du service SIB[128], on travaille tous ensemble, on a besoin de tout le monde, on a besoin des journalistes, on a besoin des documentalistes, on a besoin des informaticiens. Il y a toute une démarche et à chaque étape, on se demande de qui on a besoin, au départ des documentalistes pour faire des corpus, pareil pour travailler sur les catégories, on a besoin des *data scientists* pour mettre en place les algorithmes d'apprentissage. On a formalisé

[128] Service Informatique Banque de Contenus

une fiche avec la démarche formalisée après le stage de Nadia. Et avec les commerces, on va enrichir sûrement la démarche et on va essayer d'avancer plus loin que ce qu'on avait fait pendant le stage de Nadia. L'objectif, à terme, notamment avec le projet des commerces, j'espère qu'on pourra vraiment embarquer les infographistes et les Web designer pour concrétiser les usages qu'on imagine avec ces données-là.

Est-ce que tu te souviens de quelques thématiques qui avaient été proposées par les *data journalists* ?

Alors on eu un sujet centré sur les élus locaux : élections, démissions…[...] En mars 2015, on avait à jour la liste de tous les élus municipaux. Mais par contre, depuis 2014, il y a des démissions, de nouveaux élus mais la liste n'a pas été mise à jour. On a l'info dans nos contenus, donc on a les moyens à partir de nos contenus, en détectant ces éléments-là, avec l'aide d'algorithmes, d'avoir une base de connaissances sur les élus municipaux qui soit vraiment à jour. Comme sujet, on avait aussi mai 1968, les commerces, les poilus, les initiatives locales…[...]. Ça c'est issu d'un *brainstorming* commun entre les *data journalists* et SIB. Du coup après ces réunions, on n'a pas fait de point régulier, c'est plutôt en fonction des sujets qu'il faut les relancer. C'est vrai que sur les commerces, j'ai relancé et je n'ai pas encore eu de retour sur les usages qu'il pourrait y avoir.

Et maintenant, est-ce que tu peux me parler de l'utilité qu'aura l'ontologie Socle pour la recherche d'information ?

Une fois que l'ontologie est peuplée, c'est-à-dire qu'il y a toutes les instances, par exemple sur les élus, on mettra toutes les communautés de commune, tous les élus, etc. Ça va servir à aller interroger les contenus plus facilement que par mots-clés, de façon plus pertinente puisqu'on peut aller chercher tout ce qu'ont dit les élus municipaux de Rennes métropole sur la réforme des rythmes scolaires les deux dernières années par exemple...sans avoir à dire « qu'a dit Paul Martin ? qu'a dit Jean Dupont ? ». Donc le premier usage, c'est vraiment ça, c'est de pouvoir faciliter la fouille des contenus, faire ressortir des contenus, d'éclairer complètement les contenus parmi les 10000 articles qu'on reçoit chaque jour. Ce n'est pas facile, juste avec des mots-clés.

Un autre usage : la base de connaissances est peuplée donc ça va permettre d'identifier plus précisément ce dont on parle et de qui en parle, ça va permettre d'extraire les entités des articles avec plus de précision puisque ça permet de désambiguïser. Si par exemple, dans un article, on parle d'Orange, « est-ce que c'est la ville ? Est-ce que c'est l'entreprise… ». Si dans l'annuaire, on a son PDG, et qu'on retrouve son nom dans un article où on trouve le terme « orange », on sait qu'on parle d'Orange, l'entreprise, avec 99% de chance d'avoir raison. On utilise les relations dans

l'ontologie aussi : si dans un contenu, on parle de deux ou trois communes d'Ille-et-Vilaine, et qu'il y en a une qui a un nom présent pour une commune en Ille-et-Vilaine et dans la Manche, si on nomme d'autres communes d'Ille-et-Vilaine dans l'article, il y a de grandes chances pour qu'on parle de la commune de l'Ille-et-Vilaine dans l'article.

À l'inverse, on a l'ontologie, on a la structure et, à partir de nos contenus, on veut détecter de nouvelles entités nommées. Donc, sur les personnes et sur les relations (par exemple, « est président de »), à partir du modèle, on analyse un peu les contenus pour voir si on retrouve un peu ce modèle. Et même si on n'a pas les instances dans l'annuaire, on suggère que des entités sont pertinentes. Par exemple, dans la base de connaissances, on n'a pas encore le PDG d'Orange et si, dans un des articles, on voit « Jean Dupont, PDG d'Orange, dit… », on peut suggérer qu'il est PDG d'Orange, après c'est soumis à la validation d'êtres humains, notamment des documentalistes. Donc le modèle d'ontologie, c'est un cercle vertueux : on alimente automatiquement l'ontologie avec des algorithmes de détection, d'apprentissage. Et comme l'ontologie est alimentée, on arrive à mieux annoter les contenus, mieux extraire la connaissance et donc mieux exploiter les contenus. Ce qu'on sait faire aujourd'hui au niveau des entités, c'est reconnaître quand on parle d'une personne, d'un lieu, d'une société...demain, il faudra qu'on puisse reconnaître qu'on parle...d'une association, d'une technologie, d'un type de commerce. Si on parle sur le domaine des commerces, si on donne des concepts à l'apprentissage, l'algorithme va pouvoir notamment reconnaître qu'il y a un nouveau type de commerce dans les contenus,...où sont vendues les cigarettes électroniques par exemple. Le système a appris à reconnaître dans les contenus ce formalisme-là donc le système va proposer « est-ce que "magasin de cigarettes électroniques" ne serait pas un nouveau type de commerce ? »

Si on revient sur l'ontologie Socle, quel rôle as-tu au sein de ce projet ?

Mon rôle, c'est vraiment que ce soit réalisable. Donc je fais le lien avec tous les outils qui sont déjà en place et je vois vers où on veut aller, je vérifie que l'ontologie Socle soit réaliste. C'est pour ça qu'on a d'autres contraintes, d'autres objectifs qu'ICODA et il faut les prendre en compte. Donc il faut valider qu'on respecte bien les contraintes de production...que le modèle soit viable tout en étant évolutif donc compatible avec tous les sujets qu'on a en parallèle, le maritime, le commerce...Il faut qu'on puisse le mettre en production et qu'on ne fasse pas une usine à gaz ! C'est vrai que je m'intéresse plutôt à la partie métier sur laquelle a été greffée rapidement par ICODA une partie technique qu'Henri-Maxime maîtrise plus que moi. Le socle au sein de Ouest-France, il faut qu'on l'ait vraiment validé à 100%, qu'il soit cohérent avec ce que promeut ICODA

mais on ne peut pas prendre l'intégralité de ce qu'ils ont décidé parce que c'est un peu trop tôt pour nous. Et ce socle doit être réalisable, c'est-à-dire qu'on doit pouvoir l'intégrer à la BDC[129].

Quand tu dis « c'est un peu trop tôt pour nous », tu as des exemples entre l'ontologie d'ICODA et celle mise en place dans le service ?

Oui, il y a des aspects à la fois métier dans tout ce qui était « évènements » notamment, il y avait des détails très fins, je ne sais pas si on va aller jusque-là. Tout ce qui est « typologie des organisations » et « formes d'organisation », il y a beaucoup de détails sur les SCOP, les SARL, etc. Ils sont en avance de phase sur nous, donc je pense que c'est trop tôt et que ça ne change rien si on ne prend qu'un sous-ensemble de ce qu'ils ont fait. C'est vrai qu'ils ont envie d'aller vite parce qu'ils ont d'autres sujets derrière. Pour moi, ça va trop vite par rapport à ça. Donc on intégrera sûrement ces autres concepts mais il faut d'abord qu'on arrive à être à 100% convaincus de notre premier périmètre. Un des objectifs de l'ontologie, enfin l'idée que j'en avais avant de mettre les mains dans le cambouis on va dire, c'était que l'ontologie pouvait être un outil pour converser avec les journalistes, voire les documentalistes. Donc je voulais vraiment implanter des notions métier qui ressemblaient au terrain. Je vois qu'on diverge un petit peu en ce moment sur ça par contre. En fait, je pense que je ne passe pas assez de temps sur cette ontologie pour pouvoir creuser tout ça, j'aimerais passer plus de temps dessus, ça serait indispensable. Si on a un socle qui n'est pas stable, tout le reste au-dessus va s'effondrer...Une des raisons du socle aussi, c'est que pour chaque thématique, il ne faut pas qu'on se repose les mêmes questions de modélisation...sur les lieux, les sociétés, toutes les personnes. Il faut qu'on l'ait défini une fois pour toute. Et quand on travaillera sur ces thématiques, il faut qu'on soit bien raccords dessus.

Et quand tu disais que tu voulais que les concepts soient facilement compréhensibles par les journalistes, tu veux parler du label ?

Oui, le label en fait partie. Et on construit la base de connaissances en fonction d'usages qu'on a identifiés en partenariat plus ou moins avec les *data journalists*. Du coup, oui parce que sur ce projet, on est plus demandeurs qu'eux, pour qu'ils nous proposent des usages. On essaie de définir ça et on ne va pas être exhaustifs sur les usages. Et à la limite, à partir de la base de connaissances qu'on a construite, que les journalistes s'inventent des usages, ça serait super ! Mais ils ne peuvent le faire que si on a des concepts qui leur parlent. Par exemple, sur le sport, il ne faut pas qu'on les perde en rajoutant « des groupes de joueurs », « des groupes de supporters », etc. Alors c'est peut-être utopique, mais avoir une base de connaissances

[129] Banque de Contenus

entièrement lisible par les journalistes, ça permettrait qu'ils se baladent là-dedans et qu'ils utilisent les contenus de façon décuplée. Là, le premier travail fait avec ICODA, ce n'est pas forcément gagné ! Pour moi oui, on arrive sur une ontologie un peu technique. Il faudrait qu'on puisse avoir une vue de cette ontologie avec seulement les classes métier, à partir desquelles les journalistes puissent interroger la base.

Est-ce que vous avez déjà pensé à l'interface de l'application quand l'ontologie sera implantée dedans ?

Au tout début du projet, sur une des premières versions de Troove, il y avait de l'auto-complétion sur les éléments de l'annuaire quand on cherchait des mots-clés. Donc on pourrait imaginer avoir ça, de dire « je cherche une EPCI qui s'appelle Rennes Métropole » et à partir de cette EPCI, rebondir sur les communes, les conseillers municipaux. Mais cela fonctionnerait seulement si on a un modèle métier, s'il faut passer par les classes techniques : agent, *datatype, statement*...ça devient plus compliqué. Ça c'est ce que je m'imagine déjà dans un premier temps pour fouiller les contenus via la base de connaissances. À l'autre bout, il y a des applications type DataMaritime qui cachent complètement l'ontologie mais qui l'utilisent. À partir d'autres personnes, tu peux rebondir sur d'autres personnes, sur d'autres sociétés...Aujourd'hui, ça se fait via les occurrences dans les contenus mais l'objectif, c'est que les liens entre les entités se fassent via la base de connaissances. L'objectif, c'est de tenir compte des liens entre les entités : par exemple, deux entreprises sont implantées à Saint-Nazaire donc elles pourraient être reliées. Il y a donc une notion de voisinage, qui fait qu'on va visualiser la proximité pour ces entités.

On va parler davantage des étapes pour construire l'ontologie : quelles sont les étapes pour la construction ?

Je pense que la démarche est vraiment la même : c'est de partir des contenus, de faire des corpus sur la thématique qui nous intéresse, et ensuite, c'est de faire un zoom sur des contenus vraiment typiques et d'essayer d'analyser ce dont on parle. Ce qui était préconisé dans la démarche, c'était de faire une carte mentale avec toutes les instances de ce dont on parle, donc faire ça sur deux/trois exemples. Ainsi, des concepts et relations commencent à apparaître. Donc à partir de la carte mentale et des instances, on essaie de définir les concepts, les relations et de modéliser l'ontologie à partir de là. Après, on n'a pas bon du premier coup donc on se pose des questions, on regarde aussi des ontologies qui peuvent exister comme FOAF, schema.org...tout en essayant de bien rester dans nos contenus et dans les usages qu'on veut avoir.

Et quels ont été vos points de vigilance pour construire cette ontologie ?

L'ontologie doit être exhaustive...oui par rapport aux lieux !. Mais ce n'est pas le principal, je dirais plutôt que l'ontologie devait être évolutive, oui par rapport aux lieux, si on veut rajouter un autre type de lieu, on sait comment s'y prendre sans casser l'ontologie. On doit reprendre l'existant, ça c'est une contrainte que n'a pas forcément ICODA. On a un existant qui est la version de la banque de contenus en production donc il faut que ça rentre dedans. Puis, il faut que ça reste simple même si l'existant n'est pas toujours simple, il y a des problèmes. Il faut que ça reste le plus simple parce que si on commence à faire un socle compliqué, on va droit dans le mur. Peut-être que je freine beaucoup par rapport aux propositions d'ICODA puisqu'au début, Michel[130] disait qu'il fallait qu'on parte sur la même chose qu'ICODA mais il y a des choses qu'on va avoir du mal à faire valider. Il y a des concepts transparents pour les utilisateurs finaux, comme « agent », qui ne seront pas dans l'implémentation, mais il faut quand même qu'on reste sur quelque chose de lisible, même pour nous, au niveau de l'équipe technique, pour qu'on ne se repose pas les mêmes questions « pourquoi on a mis ça là ? », etc.

Comment tu décrirais le niveau de granularité de cette ontologie Socle telle qu'elle est aujourd'hui ? Est-ce que ce niveau est le même sur toutes les branches ?

Non, il est différent sur toutes les branches, du coup dans le Socle, il va y avoir des concepts assez généraux. On sait qu'on va en avoir besoin et on sait qu'il va falloir détailler. Dans le Socle, c'est généraliste, l'objectif c'est d'avoir en parallèle un projet qui réfléchit à ce sujet-là et qui aboutira à une ontologie externe, ou soit on ajoutera des concepts dans le Socle. Je pense que ce sera le cas quand on travaillera sur l'ontologie sur le domaine des entreprises. La granularité ne sera pas la même : on détaillera la branche des sociétés et des formes d'organisation dans ce cas-là. Par contre, la granularité au niveau des lieux est plus fine, d'une part, parce qu'on est plus avancés sur les lieux dans bdc 3. En parallèle, il y a un projet référentiel géographique au sein du groupe qui doit être déployé. Donc là, on a besoin d'avoir quelque chose de plus exhaustif et de plus détaillé. On est plus en avance sur le travail donc c'est plus détaillé. Parce que c'est tiré par ICODA qui a un cas d'usage maintenant qui est la démission des conseillers municipaux, il y a un déséquilibre et il faut voir ce qu'on va valider au niveau du Socle interne. On ne coupe pas une branche comme ça, par exemple, on ne va pas dire, « on coupe la branche "Fonction politique" parce que c'est un autre projet »...Je pense quand même qu'il faut qu'on ait quelque chose d'homogène : s'il y a beaucoup de détails sur une branche, il faudrait peut-être externaliser la branche. Par exemple, la branche "évènement" commence déjà à être bien détaillée, plus que certaines branches de

[130] Michel Le Nouy, responsable du Service Informatique Banque de Contenus (SIB) ou service BDC

l'ontologie donc il faudrait peut-être la mettre dans une autre ontologie centrée sur ce domaine en particulier.

Comment s'est déroulée la collaboration avec les chercheurs d'ICODA ?

Il y a eu pas mal de réunions de vive voix. Parce qu'ICODA est au moins sur cinq centres de recherche. Donc il y en a au moins deux à Paris, un à Montpellier, un à Rennes...On a fait plusieurs réunions à l'INRIA, de vive voix avec les chercheurs rennais, et par visioconférence avec des chercheurs d'autres sites. On a fait des réunions à Paris aussi lors des plénières de l'INRIA, où tout le monde est réuni sur place et où là, on travaille vraiment les différents aspects du sujet, dont l'ontologie. [...]

Oui, après on fait des rendez-vous par téléphone maintenant mais tout le monde n'avance pas au même rythme.

Et y a-t-il eu des points questionnants lors de cette collaboration ?

Oui, aujourd'hui, par exemple, on n'a pas le moyen de répondre à la question « quels sont les conseillers municipaux qui ont démissionné depuis 2014 en France ? ». Ce que veut mettre en place ICODA tout de suite - ils sont sûrement raison sur leur axe - c'est la notation de datation pour répondre notamment à cette question. Mais nous tout ce qui nous intéresse, c'est « le frais » et je trouve qu'on a passé beaucoup beaucoup de temps sur des aspects techniques, enfin c'est aussi ce pourquoi les chercheurs sont là. Et puis, oui, même quand on a avancé sur l'ontologie Socle très tirée vers la politique locale et les conseillers municipaux, j'avais l'impression qu'on dénaturait certains autres concepts comme les communes. Le sujet d'ICODA, ce sont les conseillers municipaux mais ce n'est pas notre sujet. Du coup, on a des objectifs un peu différents. Pour nous, on devra pouvoir traiter tous les sujets avec cette ontologie Socle.

Annexe n° 6 : Entretien avec Henri-Maxime Suchier (projet ontologie Socle)

5 juin 2018, 15h, durée : 30 min

[...]

Quand est-ce que le projet de l'ontologie Socle a commencé ?

Moi j'avais commencé à poser une première pierre assez rudimentaire de cette ontologie-là, il y a un an pour faire une sorte de démonstrateur pour montrer qu'on pouvait effectivement traduire l'annuaire qui existe pour le moment dans la BDC au format sémantique. Donc moi j'avais posé une première ontologie assez simpliste qui parlait des entités, des sociétés, des personnes et des lieux. Et également de l'endroit où ils pouvaient être cités dans les contenus. Ça c'était un petit peu avant le démarrage effectif d'ICODA et une fois qu'ICODA a vraiment commencé, c'était en juin les premières grosses réunions... à ce moment-là, il était question de produire une ontologie et de modéliser ça au format RDF, donc c'est là que ça a commencé.

Est-ce que tu as un poste défini sur ce projet ?...Est-ce qu'il y a quelque chose de formalisé ?

Il n'y a pas grand-chose de formalisé, moi on va dire que je vais plus m'occuper de tout ce qui est technique : quels outils informatiques on met en place pour porter cette ontologie, pour l'implémenter. Puis j'interviens aussi sur la partie modélisation, au même titre que toi, Nadia ou Marie-Paule pourraient intervenir en interaction avec les gens de l'équipe ICODA. Toute l'infrastructure, quelle base de données on utilise pour modéliser ça, quel script on utilise pour copier depuis l'annuaire existant vers la modélisation en RDF, ce genre de chose…

Et comment avez-vous fait pour construire cette ontologie Socle ? Peux-tu présenter les grandes étapes de la construction ?

Je pense qu'à un moment donné avec les gens du projet ICODA, on s'est dit « il faut qu'on fasse cette ontologie » et on est tous plus ou moins partis chacun de notre côté pour commencer à modéliser ça. Donc, à un moment, on s'est retrouvés et chacun de notre côté, on avait travaillé sur une ontologie. Et puis on a essayé de réconcilier ça : les gens de Montpellier avaient quelque chose de beaucoup plus fin et beaucoup plus détaillé. C'est toujours la même discussion qu'on a avec eux à chaque fois. Eux ils vont aller chercher très loin dans les détails sur des trucs très précis. Et on se rend compte que nous, on n'a non seulement pas besoin d'un tel niveau de détail et qu'en plus, ça pourrait polluer au final le modèle de données : ça serait trop précis et on n'arriverait pas

à savoir précisément dans quelle boîte mettre typiquement une entité ou ce genre de chose. On avait vraiment une problématique qu'on a essayé de mettre en place et je sais que sur certaines questions, typiquement Marie-Paule est allée impliquer Nathalie de la documentation pour modéliser ça. Je pense que l'idée, c'était vraiment d'avoir l'avis de tout le monde sur les usages au final. Moi ce que j'ai fait avec l'ontologie rudimentaire de mon côté, ce ne prenait pas en compte les besoins des uns et des autres, c'était vraiment une preuve technique.

Tu dis qu'elle était « rudimentaire »...mais tu avais déjà mis des concepts ? Combien ?

Oui, je pense que j'avais déjà la notion de poste occupé...j'avais repris telle qu'elle la modélisation proposée par l'INSEE concernant les lieux. Tu sais, il y a une ontologie de l'INSEE qui existe pour modéliser commune, département et région, j'avais repris ça tel quel. J'avais environ une dizaine de concepts, très très larges.

Sur quel outil tu as fait cette première ébauche d'ontologie ?

Sur du papier, je n'avais pas utilisé d'outil. J'ai fait ça à la main.

Et tu disais qu'après, vous aviez impliqué les documentalistes dans la réflexion pour connaître les usages ? Comment s'est passée cette collaboration ?

Je pense qu'à l'époque, on n'avait pas encore une idée très claire de comment on manipulait ça, quels usages vraiment avoir derrière...Pour nous l'idée c'était de dire, c'est un peu le compromis qu'on a à chaque fois, entre une ontologie qui doit pouvoir résoudre des problèmes techniques par une machine et une ontologie qui doit pouvoir être compréhensible, être manipulable par un être humain non-informaticien. Du coup, le fait d'impliquer des gens qui ne sont pas des informaticiens, mais qui sont des utilisateurs finaux des outils, c'est une façon d'avoir au final un outil qui soit adapté aux utilisateurs.

Sur la partie technique de l'ontologie à Ouest-France, tu travailles tout seul ?

Sur la partie vraiment technique, oui je suis tout seul. Peut-être qu'à terme, je serais épaulé par des gens de la BDC qui m'aideront à faire en sorte que cette ontologie soit supportée par la BDC. Mais pour l'instant c'est vraiment moi qui fait ça.

Au niveau technique, tu as pu échanger avec les chercheurs de l'INRIA également ?

Euh..très partiellement au final, sur les choix d'outillage technique, je connaissais déjà les outils qui pouvaient être utilisés dans un cadre industriel grâce à mon expérience chez Cojecto. C'est peut-être une approche naïve de ma part mais je me dis que les chercheurs n'ont pas vraiment des connaissances des outils technico-techniques utilisés dans un milieu industriel. Mais c'est sûrement une approche naïve de ma part en fait. Il y a ça et le fait que l'occasion ne s'est tout simplement pas présentée. Avec les gens d'ICODA, on travaille vraiment sur la conception d'une ontologie, on n'est pas trop sur des considérations vraiment techniques de comment est-ce qu'on gère les données. Peut-être que ces questions-là vont venir après, oui elles vont venir sûrement. Parce que moi je pense naïvement qu'ils ne sont pas compétents là-dedans... mais en fait je pense qu'ils le sont vraiment. Mais je pense qu'il y a des considérations qui sont vraiment techniques, je pense que ce sont des gens qui sont quand même compétents et avec lesquels on doit pouvoir échanger. Après, j'ai pas forcément de considérations vraiment techniques pour l'instant...je pense que le jour où l'ontologie sera implantée, elle sera utilisée, on sera confrontés à des problèmes qui feront qu'il faudra qu'on puisse échanger avec des gens qui s'y connaissent.

Est-ce que vous aviez décidé d'une date à laquelle l'ontologie sera utilisée dans le service ?

Michel Le Nouy parle d'avant l'été, courant du mois de juillet. Mais la question c'est, pour l'instant, les usages ne sont pas forcément bien définis. Il va y avoir Thomas[131] qui va s'appuyer là-dessus pour que l'algorithme Recco puisse faire de la désambiguïsation sur les entités. Derrière, j'imagine que l'on va pouvoir faire en sorte que les moteurs de recherche exploitent ce graph-là pour faire des recherches plus pertinentes. C'est tout un tas d'usages qui sont vraiment très intéressants, qui sont possibles mais...après moi je ne peux pas être tout seul pour mettre en place cela. On va viser la fin de l'été et on aura une ontologie qui marchera. J'imagine que derrière, les usages en découleront.

Quand tu dis « les usages » là, ce ne seront pas les usages du public potentiel, ce seront des usages que la machine va faire de l'ontologie ?

Les deux au final, ça va être comment on utilise cette ontologie-là et ce graph de données pour améliorer la recherche ou les performances en quelque sorte, même si ce n'est pas forcément le domaine de considération d'une base de données graph. On arrivera à des résultats de recherche qui seront meilleurs tout simplement.

[131] Thomas Girault, *data scientist*, spécialisé dans les questions du TAL

Est-ce que pour cette ontologie Socle, vous aviez pensé à réutiliser d'autres ontologies qui existaient déjà ?

On a regardé typiquement ce qui se faisait du côté de l'INSEE, la modélisation en RDF de la base des lieux. Après, il y a toujours tout ce qui concerne les organisations, ce sont des choses qu'on va retrouver dans FOAF, schema.org...Donc c'est toujours intéressant d'aller regarder ce qui se passe à chaque fois où on a des problèmes, pour voir comment eux ont inversé le problème. Et puis il y a des trucs classiques du genre RDFs pour parler d'un label, c'est la base. Après sur les ontologies de manière générique, il y a Schema, INSEE, FOAF...et comme on est sur quelque chose de globalement générique, on n'a pas forcément quelque chose de très précis à ce niveau-là.

Et est-ce que tu t'es posé des questions de granularité sur cette ontologie ?

Moi, à titre personnel, oui et au titre de l'équipe, oui aussi. On s'est posés ces questions-là. Oui sur comment on doit organiser les concepts sous « organisation » par exemple. Mais je pense qu'un des avantages de travailler avec RDF, c'est que tu peux faire un choix à un moment donné et le remettre en question après, c'est possible en fait. Parce que tu as un modèle que tu peux faire évoluer, ce qui n'est pas forcément le cas avec d'autres paradigmes de gestion de données. Tu peux décider à un moment de remettre en question ton niveau de granularité. Si ce n'est pas assez détaillé, je le redétaille et je suis capable de redisperser toutes mes entités par rapport à ce nouveau niveau de granularité. Sur cette ontologie-là, on se pose peut-être trop de questions sur ce niveau de granularité. On devrait se dire qu'on fait comme ça puis on sait qu'après, on peut faire évoluer ça d'une manière ou d'une autre, rebasculer, changer ça, faire évoluer un peu le schéma.

Quand tu disais paradigme de gestion de données, ça veut dire...?

Je ne sais pas si ça existe vraiment, en fait c'est un graph de données, c'est une façon de représenter, de stocker des données sur ordinateur. Mais tu as d'autres façons de gérer ça. On parle généralement de bases de données relationnelles qui sont des façons de gérer des données, tu vas avoir besoin de définir des tables, des relations entre les tables. C'est vraiment très très précis et ces modèles de données-là, pour les faire évoluer, ça peut être simple parfois mais aussi très compliqué d'autres fois. Donc c'est beaucoup plus figé que du RDF qui évolue assez facilement.

Et est-ce que tu aurais pu utiliser d'autres choses que le RDF ?

Je pense que oui, je pense qu'on aurait pu utiliser des trucs qui sont émergents mais qui tournent toujours autour de la notion de graph, ce sont des technologies hybrides. Puis je pense que des passerelles sont possibles entre du RDF et de l'hybride comme ça. Mais encore une fois, c'est une question d'usage derrière. Peut-être que quand les usages vont s'affiner, on aura pu se rendre compte que le RDF était vraiment plus adapté pour notre cas.

Tu penses qu'on pourra changer si ce n'est pas satisfaisant ?

Oui, on peut toujours traduire d'un modèle vers un autre. Quand on se dit qu'il y a l'ontologie ICODA et les autres, on peut finalement se dire qu'ICODA fait ce qu'il veut et nous on fait comme ça, on peut toujours prévoir des passerelles entre les deux. Au final, c'est pareil...toutes les traductions sont possibles, après il y en a qui sont plus coûteuses que d'autres, mais tu peux toujours passer de l'une à l'autre. Dans la mesure où tu stockes la même quantité d'informations dans un modèle ou dans l'autre, les traductions sont toujours possibles. Il faut bien sûr s'assurer de ne pas perdre d'informations parce qu'au bout d'un moment, le retour en arrière n'est pas possible.

Quand est-ce que vous avez décidé d'utiliser COGUI[132] au lieu de Protégé ? Quand s'est faite la bascule ?

C'était juste au moment de ton arrivée en fait[133]. C'est l'outil dont nous avaient parlé les gens de Montpellier en janvier. Avant, il n'était pas encore utilisable. Puis en avril, COGUI est arrivé à un niveau de maturité suffisant en avril, Michel Chein nous en a fait la présentation. Puis on a commencé à l'utiliser la semaine suivante et pour l'ontologie Socle.

Il y a beaucoup de choses qui changent sur COGUI, à part la visualisation de l'ontologie qui est simplifiée ?

Non, je pense qu'il doit y avoir beaucoup de choses qui changent pour une utilisation poussée. Mais je n'ai pas fait assez le tour des deux pour vraiment faire la différence entre ces logiciels. Notre niveau d'utilisation qui est la conception d'ontologies est relativement simple, quand même malgré tout. Les deux se valent. Ce qui est appréciable sur COGUI, c'est effectivement le fait de pouvoir utiliser un graph, c'est beaucoup plus compliqué sur Protégé.

---

[132] Éditeur d'ontologies développé par le LIRMM : http://www.lirmm.fr/cogui/

[133] Donc mi-avril 2018

OK, est-ce que tu peux m'en dire plus sur la collaboration entre les chercheurs ?

La collaboration, c'est vraiment les échanges qu'on peut avoir via skype, où on parle. Finalement, eux avancent de leur côté, nous un petit peu du nôtre. Et par moment, on fait des réunions. C'est là où on essaie de se resynchroniser et de dire sur telle question « je ne comprends pas ça, etc. » Eux vont nous poser des questions comme « quel sera votre usage de telle partie de l'ontologie ? », eux ça leur permet d'avancer. La collaboration est presque plus à un niveau où, nous, on est client d'eux, dans le sens où, eux, font ce travail de conception et nous derrière on récupère, on fait un peu évoluer l'ontologie pour nos usages à nous. Alors, au début, ce n'était pas tout à fait ça parce qu'on a fait une première version de l'ontologie de notre côté et eux, du leur. Et on a réussi à superposer ça, se dire « au final, on a à peu près les mêmes choses, les grands concepts dans les grandes lignes ». Globalement, c'était la même chose, excepté les détails. Et depuis, j'ai l'impression que c'est quand même plus eux qui sont moteurs je pense. Après, on vient avec des questions à leur poser : « Pourquoi ci ? Pourquoi ça ? Pourquoi avoir fait ce choix-là ? ».

Et qu'est-ce que ça t'apporte cette collaboration avec des chercheurs sur l'ontologie ?

Comme à chaque fois que j'ai travaillé avec des gens qui sont experts dans un domaine, au final, tu tires profit de leur expérience et de leurs connaissances. Là où c'est un peu frustrant, c'est de les avoir à distance finalement. Ça serait bien d'avoir la possibilité de les voir plus au fil de l'eau et de pouvoir les voir dès qu'on a une question, etc. Ça serait plus bénéfique. Mais quand on a une question, une incompréhension sur l'ontologie, on note ça sur un papier et puis on essaie de le formaliser dans un *e-mail.* Ils nous répondent, ils n'ont pas forcément compris la question, on n'a pas forcément compris la réponse donc y a toujours un peu de latence au niveau de la compréhension d'un côté comme de l'autre. Mais je pense que globalement, on tire profit de leur expérience. Et assez souvent j'ai l'impression que c'est comme pour se dire « Bon, c'est bon, on ne fait pas n'importe quoi parce que nos choix sont confortés par ce qu'eux font ». Et des fois, les chercheurs ont des points de vue qui entrent en contradiction entre eux. Donc, ça montre bien qu'il n'y a pas juste une façon de faire. Il n'y a pas une seule modélisation possible mais plusieurs points de vue qui vont à chaque fois correspondre à un besoin ou une compréhension du problème.

Annexe n° 7 : Entretien avec Nadia Fafi (projet Datamaritime)

31 mai 2018, 17h, durée : 37 min

[...]

Tout d'abord on va commencer par la genèse du projet de l'ontologie du domaine maritime : qui a eu cette idée et quand le projet a été mis en place ?

Moi j'ai été prise en stage pour la constitution de l'ontologie mais c'était la première du service, il n'y en avait pas eu avant. Et je sais que Michel Le Nouy avait rencontré des interlocuteurs qui l'avait emmené sur ces pistes-là, de construire des ontologies. Je suis arrivée pile au moment au début du projet ICODA. Il y avait aussi le fait de construire tout le projet DataMaritime pour le cluster, qui était notre potentiel client à l'époque.

D'accord, est-ce que tu pourrais expliciter le projet ICODA et ensuite le cluster, qu'est-ce que sont ces projets concrètement ?

Ce sont deux projets différents. C'est un peu représentatif du fonctionnement du service : il y a une partie recherche où on essaie de développer des outils qui vont nous permettre d'avancer sur plein de problèmes rencontrés dans la presse. Donc là, ICODA c'est vraiment un projet de recherche où on a un partenariat avec des chercheurs de l'IRISSA et de l'INRIA. Et il me semble que Michel a eu des rendez-vous avec des personnes des Décodeurs du Monde et du journal Le Monde. Là, on est vraiment sur une partie recherche, on va essayer de travailler sur les outils qu'on peut développer pour la vérification des sources, contre les *fake news*, pouvoir travailler sur les ontologies. On va pouvoir aborder plein de sujets différents qui seront utiles à la presse. Mais de plus en plus, on s'approche quand même d'une utilité, du fait de pouvoir utiliser ce qui ressort de ces recherches.

Le deuxième projet qui était simultané c'était le projet pour le cluster. C'était un projet qu'on avait avec la rédactrice en chef du Marin, Alexandra Turcat, qui était notre principale interlocutrice et qui faisait la médiation entre le service et le cluster maritime. L'idée, c'était de valoriser les contenus du Marin et aussi les contenus de Ouest-France en entier. Puis de pouvoir constituer une interface qui permette d'accéder tous les jours aux contenus qui concernaient le domaine maritime.

Quand tu dis le cluster, ça veut dire catégorisation ?

Le cluster, c'est le nom de l'entreprise. Oui, oui effectivement, c'était vraiment catégoriser seulement les contenus sur le domaine maritime. On dit « cluster » car c'est une entreprise qui va relier plein d'entreprises du domaine maritime et va pouvoir les relier. Donc le cluster va être un point de rendez-vous où on va pouvoir communiquer sur ce qui se passe dans le domaine maritime en prenant en compte tout ce qui vient de ces différentes entreprises, qu'elles soient centrées sur le transport maritime, sur les chantiers….mais toujours sur le maritime. Et le projet s'appelle DataMaritime, l'idée c'était de pouvoir valoriser les données qui sont dans les articles. Effectivement, nous on construit une interface avec des articles du groupe sur le domaine du maritime. Sauf qu'il y a toute une partie datavisualisation qui permet de bien valoriser toutes les données qui sont dans les contenus du Marin.

Tu sais quand ce projet a été mis en production ou a été livré ?

Le deuxième jour de mon stage, l'ancienne ergonome avait montré les premières maquettes de l'interface. À la fin de mon stage, on était un peu dans le rush, il fallait présenter le projet donc vers le mois d'octobre, tout est passé en production.

Et où en est rendu le projet maintenant ? Il va y avoir des évolutions ?

Là c'était un prototype, le cluster n'était pas forcé de l'utiliser. Là, Datamaritime pourrait être présenté à la rédaction pour voir si les journalistes seraient éventuellement intéressés par l'utilisation du site. Sinon, on pourrait se projeter sur la construction d'autres sites. On pourrait par exemple construire un DataCulture pour valoriser toutes nos données sur le monde culturel...On se projette sur des variantes peut-être.

Tu devais construire une ontologie qui s'inscrivait dans ce projet, c'est ça ?

Il y avait l'ontologie sur le domaine maritime et la catégorisation des contenus pour donner des points d'accès sur l'interface. L'idée, ce n'était pas de mettre un bouton « Articles » sur l'interface, mais de bien pouvoir les segmenter pour qu'on puisse les valoriser, afin de pouvoir y accéder plus facilement. La catégorisation était là pour ça. Il fallait confronter à chaque fois les besoins du cluster, les idées de la rédactrice en chef du Marin, et la réalité des contenus du Marin. On a essayé de faire un compromis et ça s'est bien passé.

Pour l'ontologie, l'idée était d'optimiser l'organisation des données sur le domaine maritime.

Peux-tu réexpliquer ce qu'est le référentiel ?

L'annuaire nous permet déjà de faire ressortir des données des contenus. Pour l'instant, on n'a que des données qui sont : les noms des sociétés qui apparaissent dans les articles, les noms des personnes et les noms de lieux donc les communes. Pour l'instant, on a seulement la possibilité d'intégrer ces trois types de données-là, en tout cas de les valoriser en les annotant.

Tu as dit que tu avais dû jongler entre les contenus, la réalité...tu avais dû prendre en compte différentes types de contraintes...Tu peux expliquer ça ?

Il y a des thématiques récurrentes quand on parle du domaine maritime, un peu comme quand on parle du domaine de la santé, de l'agriculture. Si on te demande de catégoriser les contenus ou de créer une base pour organiser des données à partir des thématiques récurrentes, parfois ça peut bloquer dans la mesure où tous nos contenus ne correspondent pas toujours à ces thématiques récurrentes. On va pouvoir faire ressurgir des thématiques qui sont moins récurrentes ailleurs mais on va peut-être aussi manquer de contenus sur d'autres. Il y avait aussi des catégories transversales : formation, environnement, etc. C'était pas évident de pouvoir travailler dessus parce que l'algorithme ne va pas comprendre puisque par exemple, environnement, c'est une catégorie qui peut surgir partout. Notre rôle en tant qu'architectes de l'information c'est de bien regarder dans les contenus pour voir si on a bien des articles sur toutes les thématiques proposées : aquaculture, énergies renouvelables, bio-technologies, tourisme, etc. Aquaculture, c'est un bon exemple : on n'a pas énormément de contenus, on va donc manquer de contenus. Il va falloir faire des compromis et dire « Ok, là j'ai des contenus sur l'aquaculture mais peut-être pas assez pour en faire une catégorie donc je propose de placer la catégorie aquaculture dans la catégorie pêche ». Il fallait créer une architecture pour pouvoir classer tous les contenus.

Ça a été un travail préparatoire à la construction de l'ontologie ?

C'était simultané parce que quand tu dois constituer une ontologie sur le domaine maritime, par exemple, (que ce soit celui-ci ou sur un autre domaine), il y a toujours une partie où tu es censée développer une expertise de ce domaine-là pour pouvoir savoir quelles sont les données à valoriser, quels concepts sont indéniablement importants. Il fallait développer une expertise en s'appropriant les contenus, il fallait passer du temps pour essayer de les catégoriser, beaucoup de lectures sur les contenus du domaine maritime. Ce travail-là de catégorisation des contenus me permettait quand même d'alimenter une expertise sur ce domaine pour pouvoir juger de quels concepts allaient nous être utiles derrière.

Et après, quelles ont été les étapes suivantes pour arriver à la version finale de ton ontologie ?

Avec Henri-Maxime et Marie-Paule, on est passés par deux étapes pour l'ontologie finale. D'abord, on avait essayé de faire une espèce de carte mentale, un peu brouillon qui devait juste nous servir à noter tous les concepts qu'on avait l'impression de voir émerger et toutes les relations qu'on pouvait projeter sur ces concepts-là. Ça a été un premier travail important puisque cela nous a permis de voir que l'on confondait certains éléments avec des concepts et que c'était simplement des thématiques transversales qui pouvaient apparaître dans n'importe quel type de contenus, et qui n'étaient pas seulement propres au domaine maritime. Essayer de créer et de visualiser les relations qu'il y avait entre les premiers concepts qui émergeaient nous a permis d'écarter les choses qu'on ne voulait pas dans l'ontologie. Ensuite, on a essayé de créer un schéma vraiment solide avec les concepts qu'on gardait, les relations qui nous semblaient importantes, d'essayer de se projeter sur les différents types de données qu'on allait pouvoir trouver. Là, on a décidé sur ce schéma-là d'aller au plus large parce qu'on savait qu'il y avait des données qu'on n'allait pas utiliser sur ce projet-là. Mais, on se disait qu'il fallait mieux construire un modèle très très large qui allait pouvoir accueillir beaucoup de types de données et d'instances derrière, au cas où ce serait nécessaire plus tard.

Quels outils vous avez utilisé alors ? Pour faire la carte mentale ou pour la construction de l'ontologie…?

J'ai utilisé des outils Google…(rire). En fait, j'ai rajouté des *plugins* juste pour avoir des outils qui permettaient de faire des schémas heuristiques. Je voulais travailler collaborativement. C'était un peu l'urgence, j'étais stagiaire. C'est après mon stage qu'on a commencé à chercher d'autres outils qui nous permettaient de travailler de façon collaborative. Pour la construction de l'ontologie, on a utilisé le logiciel Protégé. Je ne connaissais pas du tout comme c'était ma première ontologie. On a découvert petit à petit comment on allait l'utiliser.

Parce que quand tu dis « on a découvert, on a fait.. », comment tu as fait concrètement pour travailler avec Marie-Paule et Henri-Maxime ? De quelle manière vous avez travaillé ?

C'était ça qui était assez incroyable pendant ce stage ! Pour Marie-Paule et Henri-Maxime, c'était aussi leur première ontologie sauf que Henri-Maxime avait beaucoup de connaissances techniques sur le sujet, ce qui lui permettait quand même d'avoir des réponses sur ce qui allait

fonctionner ou pas techniquement sur le modèle qu'on était en train de constituer. On découvrait ensemble comment on allait constituer notre première ontologie.

Donc, c'est au fur et à mesure que vous avez constitué une manière de faire ?

Oui, la méthode on ne l'a pas forcément anticipée. Quand je suis arrivée en stage, j'avais eu juste un cours de trois heures sur les ontologies...et ce n'était pas au centre de mon sujet de mémoire. Et comme c'était la première ontologie du service, j'avoue qu'au début c'était plutôt Henri-Maxime qui nous guidait, qui nous disait « Là, on va faire les concepts ». Il nous a expliqué les différences qu'on voulait faire entre les différents types d'instances, etc. Moi, je découvrais vraiment avec lui. Il avait vraiment un bagage technique plus fort. Pour le reste, quand Henri-Maxime, ne savait pas comme nous, on faisait des recherches chacun de notre côté et on essayait de proposer des choses. Donc c'était vraiment la découverte, il n'y avait pas de méthode anticipée.

Et je pense que vous avez été en relation avec des chercheurs pour faire cette ontologie aussi...

Alors, l'aide des chercheurs est venue après. Pendant la construction de l'ontologie, on se débrouillait avec nos contenus et nos outils. Par exemple, avec Marie-Paule, on prenait un article, on allait à la pêche aux données dedans pour essayer de trouver tous les types de données qu'on pouvait trouver dans un article fort sur le domaine maritime. Ensuite, on essayait de confronter nos contenus au modèle qu'on avait constitué pour voir si c'était bon. À ce moment-là, on ne faisait pas forcément appel aux chercheurs, c'est venu après. Ensuite, j'ai reçu des documents des chercheurs. Ils m'envoyaient des recommandations à partir des petites failles, des lacunes qu'eux avaient constatées dans l'ontologie. C'est venu après en fait, ce n'était pas simultané.

Comme c'est venu après, est-ce que tu as pu faire des modifications ?

Oui, j'ai pu faire certaines modifications sur la forme de l'ontologie. J'ai fait ce que j'ai pu pour essayer de l'améliorer. Mais il y avait vraiment des choses à reconstituer dès le départ. Il y a certaines autres lacunes qu'on n'a pas pu reprendre à ce moment-là. Leurs recommandations étaient utiles, pas forcément pour cette ontologie qui a été un peu expérimentale pour nous, mais pour les ontologies à venir. Parce qu'on a les recommandations, on sait ce qu'il n'allait pas sur la première et on va pouvoir s'améliorer sur les prochaines.

Est-ce que tu peux me dire les principales lacunes ?

Il y avait des petits désaccords concernant les labels. Il y avait désaccord sur le fait qu'un nom de concept soit plus ou moins explicite pour les autres. Et au départ, on n'avait pas mis des noms de concepts qui étaient explicites. Donc on a essayé de rendre les choses plus claires et de vulgariser les noms des concepts.

Et sur l'architecture générale de l'ontologie, tu as eu des remarques ou c'était plutôt sur la forme ?

C'était vraiment sur la forme.

Sur quel point tu as porté ton attention quand tu as fait cette ontologie ? Quels points étaient déterminants quand tu l'as construite ?...Parce que tu as dit qu'il y avait plusieurs contraintes à prendre en compte.

C'est vrai qu'à ce moment-là, on avait les recommandations de la rédactrice en chef et pour moi, ce qui était vraiment important, et ce qui m'a beaucoup aidée, c'est que je ne m'éloignais jamais des contenus, c'était un élément qu'il ne fallait pas que je perde de vue. Mes données, c'est de là qu'elles vont venir donc je ne m'éloignais pas des contenus, j'allais vraiment voir quel type de données j'avais à chaque fois dans les articles. Et ensuite, j'avais par contre l'aide assez récurrente d'Alexandra Turcat, la rédactrice en chef, parce qu'elle regardait ce qu'on avait fait, elle pouvait nous reprendre sur certains aspects, nous dire qu'elle n'était pas tellement d'accord avec un concept, qu'elle l'aurait formulé autrement ou qu'elle l'aurait représenté en relation avec un autre, etc. Ça c'était son expertise à elle du domaine maritime, mais aussi du contenu du Marin qui nous permettait d'avancer très très vite et de régler justement des problèmes d'architecture qu'on aurait pu avoir par la suite. Donc, pour moi, mes référents, c'était vraiment mes contenus et l'experte, la rédactrice en chef.

Tu as su dès le départ pourquoi tu faisais cette ontologie ? Quels étaient les usages...à quoi elle allait servir ?

Non, je n'avais pas d'impératif au niveau des usagers parce que je n'étais pas en relation avec eux. Je n'avais aucun moyen de communiquer avec les potentiels usagers. Pour moi, la médiatrice entre eux et nous, c'était vraiment la rédactrice en chef. Donc je ne savais pas sur quel niveau de compétence on allait être, etc.

Parce que qui auraient été les futurs usagers ?

À ce moment-là, on parlait plus des personnes du cluster maritime, enfin de toutes les entreprises qui sont reliées au cluster donc ça fait énormément de personnes. C'est vrai que je savais qu'il fallait valoriser les données qui venaient des contenus mais je n'avais pas de contrainte explicite de la part d'usagers. J'avais des contraintes de la part de la rédactrice en chef, qui m'orientait déjà bien. Par contre, les contraintes avec lesquelles j'ai dû faire, c'est que je travaillais dans un service informatique donc je devais partir des contenus, je devais essayer de construire quelque chose qui devait bien valoriser nos données mais, je devais toujours faire face à une contrainte technique qui va être celle d'un collègue qui me dit que « non, là ça représenterait un volume trop important de données », qu'il n'est pas forcément pour les données en triplets... Donc je devais avancer en ayant en tête les contraintes techniques que mes collègues informaticiens pouvaient avoir autour de moi et je ne prenais pas tellement en compte les besoins des usagers, pour être honnête, car je ne communiquais pas avec.

Et, alors, tu l'as vu comme un point bloquant d'être avec beaucoup d'informaticiens dans ton service ?

Non, ça faisait avancer le projet, ils sont vraiment essentiels. De toute façon, sans eux, ton ontologie ne sera pas utilisée, elle restera sur une carte mentale ou dans Protégé. Non, bien sûr mais...c'est vrai que ce n'était pas toujours évident de savoir qu'en tant que documentaliste, tu développes des outils, un modèle mais que tu dépends directement par contre de l'accord final d'une majorité d'informaticiens.

Comment s'est déroulée la collaboration avec la rédactrice en chef ?

Au départ, je ne connaissais pas grand-chose au domaine maritime et au fur et à mesure, je développais une expertise, je commençais à développer un modèle, j'avais des doutes de temps en temps parce que je découvrais en même temps le domaine maritime, toute son économie...Du coup, Marie-Paule m'avait proposé de rencontrer Alexandra Turcat afin de pouvoir confronter mon travail avec ses volontés à elle, ses exigences. Ça s'est super bien passé, Alexandra m'a donné de très bons conseils dès le premier rendez-vous. Donc après, j'ai pris l'initiative de continuer à lui demander des rendez-vous pour qu'elle puisse valider le travail qui était fait au fur et à mesure.

D'accord, maintenant on va plutôt parler du peuplement de l'ontologie parce que j'ai vu que tu avais mis pas mal d'instances dedans. Comment tu t'y es prise ?

Alors les instances venaient de deux sources différentes à ce moment-là. La première source, c'était nos contenus. Une fois qu'on avait catégorisé nos contenus, on avait des points d'accès qui nous permettaient par exemple d'accéder à tous les contenus publiés sur la pêche et Thomas Girault est allé extraire automatiquement toutes les entités qui émergeaient de ces articles-là : les noms de personnes, d'entreprises, de lieux, etc. Il me fournissait des listes sur des tableaux avec toutes les données qu'il avait pu extraire. On avait déjà certaines données dans notre annuaire donc elles étaient déjà annotées. Pour certaines, je les avais déjà dans l'ontologie. Si une donnée était essentielle et n'apparaissait pas dans l'ontologie, là je pouvais l'ajouter. La deuxième source, c'était l'annuaire du cluster: j'avais accès à cet annuaire dans lequel il y avait beaucoup de noms de personnes qui étaient reliées à des sociétés et je pouvais essayer de vérifier si la donnée était toujours fiable. Par exemple, si la personne travaillait encore dans la même société et voir si la donnée apparaissait bien dans nos contenus. Je mettais les instances dans l'ontologie, et en même temps dans le back-office pour qu'on puisse les annoter et avoir de belles datavisualisations.

Et quand je te dis « méthodologie de construction d'ontologies », à quoi penses-tu ? Qu'est-ce que ça recoupe pour toi ?

....Maintenant qu'on est sur le deuxième modèle pour constituer une ontologie, je me rends beaucoup plus compte de la nécessité d'avoir une méthode contrairement à la première ontologie où on découvrait. Là, j'ai l'impression qu'on a un semblant de méthode qui évolue au fil du temps. Moi quand tu dis « méthode », j'ai des étapes en tête, des étapes de travail. Donc j'imagine que mes collègues informaticiens ne donneront pas forcément les mêmes réponses parce qu'on va être sur des étapes de travail différentes.

Est-ce que tu peux me parler plus des étapes de travail ?

Pour moi, j'ai beaucoup plus à cœur la question des usages maintenant. Certes, parmi les étapes, la plus importante, ça reste celle que j'avais déjà sur la première ontologie, c'est d'aller voir quel type de données on a, quels concepts vont apparaître au sein de nos contenus. Mais maintenant j'ai aussi beaucoup plus en tête quand on fait un *brainstorming* avec Marie-Paule et Henri-Maxime, l'usage final pour essayer de voir si on a trop affiné ou pas, si les choses sont explicites, si elles sont compréhensibles pour différents métiers, différents niveaux de compétence, etc.

Tu parlais de granularité, est-ce que tu as eu à faire des choix concernant le niveau de couverture du domaine, le niveau de détail…?

Oui, et souvent sur la granularité, je me rends compte que quand on échange, j'ai parfois l'impression qu'on est allés un peu trop loin...mais je me rends compte après que c'est parce que c'est des besoins techniques. C'est pour ça que moi au départ, j'y vois moins d'intérêt, je pense aux usages métier. Souvent, Henri-Maxime va nous expliquer que c'est moins pour l'humain que pour la machine qu'on a fait certains choix. Et là, ça me rassure !

Donc quand tu allais plus dans le détail, c'était plutôt pour la machine ?

Oui parce que par exemple, si on parle de gouvernance de la donnée, etc., je me disais « je ne vais jamais réussir à dissocier tous ces types de données, c'est très fin et sur les dates, je ne vais jamais pouvoir toujours trouver des dates pour alimenter l'ontologie ». Mais pour que nos données restent fiables, ces dates sont quand même très importantes, toutes les durées dans les *data-properties*. Tout le travail qui a été fait, c'est quand même très essentiel pour que la machine puisse comprendre quelles données sont fiables à ce moment. Sur le modèle qu'on constitue en ce moment[134], il y a des choses que je n'aurais pas forcément utilisées parce que par exemple, le code NAF m'aurait suffi. Mais je sais qu'on doit ajouter plein d'autres propriétés pour que la machine puisse dissocier les choses aussi et qu'il y ait moins d'ambiguïté.

En quoi ton ontologie devait aider à la recherche d'information dans l'interface ?

Je ne faisais pas de lien direct avec la recherche d'information parce que la recherche d'information, je l'ai beaucoup plus en tête quand je travaille sur l'étape catégorisation des contenus. Ça c'est vraiment pour moi la mission qui va permettre d'optimiser la recherche d'information. J'avoue que je l'ai moins en tête sur les ontologies parce que les ontologies pour moi, c'est surtout la valorisation de nos données par la suite.

[134] L'ontologie Socle

Annexe n° 8 : Entretien avec Michel Chein (projet ontologie Socle/ICODA)

7 juin 2018, 15h, durée : 55 min

[...]

Et si on se centre davantage sur ICODA, est-ce que vous pouvez me présenter brièvement le projet ?

Moi, la partie qui m'intéresse, c'est la partie sur le *data journalism,* donc donner des outils informatiques pour aider le travail des journalistes. Il y a d'autres aspects mais les aspects sur lesquels on va travailler dans l'équipe c'est de faire en sorte qu'on puisse par des requêtes attaquer tout un tas de bases différentes, que ce soit des bases de données bien structurées comme celles de l'INSEE ou des choses sur le Web sémantique comme Wikidata, Wikipédia, et puis les contenus propres à des journaux donc des articles, les bases Ouest-France, la Banque de contenus. On souhaite que l'usager puisse interroger ces bases de manière transparente. Il faut alors s'occuper de la validité des sources, voir si ces sources sont pertinentes. [...] On travaille beaucoup sur le temps maintenant dans l'équipe, être capable de gérer le temps, de pouvoir gérer des questions temporelles. On a choisi d'être dans un format Web sémantique et d'utiliser du RDF, le RDF a tout un tas de contraintes avec des relations binaires, ce qui veut dire que quand on veut gérer le temps, on est obligés de faire des opérations de réification pour qu'on puisse associer à un triplet un certain nombre d'informations temporelles, de pertinence...Ça prend plus de temps de travailler avec des triplets, c'est une contrainte forte qui va nécessiter de faire des développements nouveaux, là c'est plus la partie recherche. Puis l'idée c'est de pouvoir faire des raisonnements un peu complexes derrière, donc le modèle doit pouvoir permettre de faire ces raisonnements complexes.

L'ontologie Socle, sur laquelle on est en train de travailler dans le service Banque de Contenus à Ouest-France, vous l'appelez aussi « ontologie Socle » de votre côté ?

Ah non, celle-ci on l'a appelée « Ontologie version 1 »...exclusivement celle concernant la démission des conseillers municipaux. On doit avoir terminé le prototype avant septembre-octobre. Donc on a limité énormément le nombre de données, quand je dis « nous », c'est l'équipe ICODA, ce n'est pas forcément nous à Montpellier, au niveau de l'équipe GraphIK. Lors de la dernière réunion plénière, on a décidé de limiter aux démissions de conseillers municipaux. Donc l'ontologie Ont-1, c'est le niveau le plus simple, le niveau utilisateur. Ont-2, ça sera le niveau où on a réifié pour prendre en compte le temps, la provenance et la pertinence des sources, etc.

Et il y aura un niveau 3 sans doute qui sera lié à des problèmes d'efficacité : tout ça est compliqué donc il faut en même temps que le délai de réponse à des requêtes soit rapide sinon les gens ne s'en serviront pas. Donc à la fin, il faudra rajouter dans l'ontologie des notions purement techniques, histoires de ne pas passer trop de temps sur le traitement d'une requête. Voilà les trois niveaux : donc le premier niveau avec lequel on a travaillé en collaboration avec des gens de Ouest-France. Donc on a mis dedans ce qui concerne essentiellement les conseillers municipaux, naturellement, on a étendu à « départements, « structures administratives », etc. Et on a étendu, puisqu'on avait parlé à un moment donné des applications sur le domaine du commerce, du maritime...Donc on a ajouté quelques petites informations plus générales, par exemple, les sociétés, les entreprises en regardant ce qui se faisait au niveau de l'INSEE, au niveau du Centre National de la Fonction Publique Territoriale (CNFPT) sur les structures politiques, communales, etc. Ces sources donnent des bonnes indications puis on a trouvé quelques informations sur ce qui était déjà dans la base de connaissances de Ouest-France sur les conseillers municipaux par exemple.

Qui a eu l'idée d'élaborer une ontologie pour ce projet au départ ?

On avait décidé dès le début du projet d'attaquer un ensemble de bases de données de nature différente et ça nécessitait d'avoir un vocabulaire commun. Qui dit vocabulaire commun, dit ontologie dans le jargon actuel. Le point de départ, c'était vraiment celui-là. Chaque base de données a son propre vocabulaire, s'exprime différemment des autres, a ses propres métadonnées. Comme on veut attaquer tout ça d'une manière complètement transparente pour les utilisateurs, il faut bâtir un langage commun pour décrire toutes ces ressources différentes.

Et qui a décidé de recentrer le sujet sur les conseillers municipaux ?

Là, c'est à la demande d'un journaliste de Ouest-France, Erwan Alix qui était présent à la dernière réunion à Rennes à Ouest-France. Puis après, c'est Guillaume Gravier, le responsable du projet, qui a trouvé que c'était bien de se limiter à un sujet et tout le monde était d'accord. Ça intéresse un journaliste, politiquement c'est bien puisqu'il y aura bientôt les élections municipales...Ils voulaient savoir quel conseiller municipal démissionnait, donc avoir des villes, des régions, avoir les dates, les raisons des démissions, la nuance politique du conseiller aussi. Là j'utilise le jargon du ministère de l'Intérieur qui parle de nuance politique. Il faut bien parler comme les gens qui vont ensuite faire les statistiques sur les résultats des élections. Vraiment, on s'est appuyés sur une demande réelle d'un journaliste. Pendant cette réunion, Erwan avait dit « j'ai vu qu'il a eu plein de démissions des conseillers municipaux, ça m'intéresserait de faire un article dessus donc

j'ai besoin d'acquérir des données sur ces démissions ». Donc il y a une petite base qui a été faite par Ouest-France déjà mais c'est une base locale. On veut attaquer INSEE, si on veut attaquer d'autres bases, d'autres informations sur les conseillers...on est obligés d'avoir un vocabulaire commun pour ne pas avoir à attaquer chaque base séparément. En effet, certaines bases pourraient ne pas comprendre des requêtes.

Vous avez encore des relations avec Erwan Alix ?

Pour l'instant, aucune, on passe par l'intermédiaire des gens de Ouest-France avec Henri-Maxime, avec Marie-Paule...On n'a pas de lien direct.

Est-ce que vous avez eu une méthodologie arrêtée pour construire cette ontologie ?

Ça, j'ai envie de dire, c'est un peu du bricolage ! Comment on a fait ? Donc, une partie a été faite par des gens de Ouest-France, on en avait parlé lors de la réunion. C'est vraiment du bricolage : c'est-à-dire qu'on utilise des sources qui concernent des notions importantes pour les démissions. Par exemple, on va se demander quels sont les concepts les plus importants : ville, village...Donc on va regarder les sources qui existent, qui parlent de ces choses-là. On s'est essentiellement basés sur des sources bien structurées, l'INSEE par exemple, on s'est beaucoup servis du vocabulaire de l'INSEE. Il y a des ontologies qui existent comme elle du CIDOC-CRM ou FRBR qui sont des ontologies utilisées beaucoup en documentation. C'est des normes maintenant qui sont utilisées. Puis après, c'est du bon sens en étant toujours guidés par l'objectif final qui est celui de poser des questions concernant la démission de « qui ? quand ? où ? pourquoi ? quel est son profil ? » On a les bases, on sait quel genre de requêtes on veut être capable de faire. Du coup, petit à petit on remplit comme ça...mais c'est vraiment un processus itératif. Au début les gens pensaient qu'il pouvait y avoir des étapes séquentielles, ça c'est pas vrai du tout, on s'est rendu compte que ce n'était pas raisonnable, c'est pas possible...C'est la méthode itérative qui, le plus rapidement possible, prend en compte les problèmes qu'on veut résoudre. Donc la grande difficulté, c'est d'avoir de bons échantillons. En gros, il y a quatre grands groupes de personnes qui sont concernés : il y a les utilisateurs finaux, en l'occurrence des journalistes. Après, il y a les spécialistes, les experts du domaine donc ce sont des gens qui ont déjà eu à faire de la modélisation, un peu de représentation des connaissances. Pour prendre un parallèle plus simple avec le monde des bibliothèques, l'utilisateur c'est le documentaliste, l'expert c'est une personne qui met au point les outils, en l'occurrence en France, ça serait les gens de la BNF, de l'INIST ou des gens de l'ABES pour l'enseignement supérieur. Après, il y a des gens qu'on appelle souvent « ingénieurs de la connaissance », pour moi ce sont des gens qui essaient de modéliser les

connaissances qui seraient utiles pour les utilisateurs médiatisés par « les experts ». Puis à la fin, il y a les informaticiens, qui eux se préoccupent des différents modèles techniques qui existent pour faire tourner tout ça, comment on bricole ?, etc.

Alors cette démarche peut faire penser qu'on peut avoir une démarche complètement séquentielle, donc commencer par demander l'avis des utilisateurs, tout ça fait par des experts. Puis on représente les connaissances des experts et on implémente...mais ça, ça ne marche pas. On s'est rendu compte de ça.

Et comment vous en êtes-vous rendu compte ?

On s'est rendu compte de ça parce que quand on s'intéresse à des domaines qui ne sont pas des domaines techniques, pour lesquels les choses sont très normalisées et très formalisées. Si on s'intéresse à la mécanique par exemple, il y a des gens dans l'équipe qui se centrent sur un système expert pour des gens qui font des pièces de mécanique. C'est déjà bien mathématisé donc on peut aller très vite. Quand on s'intéresse à un domaine qui est plutôt du domaine de la langue naturelle puisqu'on a plutôt des ressources qui sont des textes écrits..Savoir quels seront les concepts pertinents, les classes fondamentales et les propriétés fondamentales entre ces classes...sans essayer de faire tourner un système à partir des questions que vont poser les utilisateurs, ce n'est pas très pertinent. Il faut faire tellement de choix, il faut éliminer tellement de choses, le domaine est tellement compliqué qu'il va falloir une approche itérative. Donc l'approche qu'on a choisie pour faire l'ontologie sur la démission des conseillers municipaux me semble très raisonnable, on a choisi des questions très limitées parce que sinon, on ne s'en sort pas. Il faut pouvoir faire une évaluation sur la construction itérative du prototype très vite.

Et comment et à quel moment décidez-vous de faire des itérations ? Comment ça fonctionne ?

À l'heure actuelle, si quelqu'un travaille sur un sujet, il nous montre ce qu'il a fait dessus puis on lui dit ce qu'on en pense, puis il reprend le sujet puis on compare avec les questions que pourrait poser le journaliste. Maintenant, il va falloir à un moment donné que Ouest-France, enfin les informaticiens de Ouest-France servent d'intermédiaires et puissent discuter avec les journalistes pour aussi poser des questions. Par exemple, dans un conseil municipal, il y a plusieurs commissions et il faudrait qu'on sache si on s'intéresse aussi aux démissions des commissions, si ça intéresse les journalistes. [...] Mais ça c'est une idée qu'on a et qu'il faut à tout prix qu'on confronte à ce que veut le journaliste. Si lui, il considère que c'est inintéressant, que c'est trop pointu, que ce n'est pas la peine, on laissera tomber. On a besoin d'avoir son avis vite parce que ça complique énormément les choses.

Et il y a eu d'autres points questionnants lors de la construction de cette ontologie ?

Oui, par exemple, il y a plein de problèmes qui se posent et qui ne sont toujours pas réglés. Par exemple, un conseil municipal, c'est quoi ? Est-ce que c'est un ensemble de personnes ? Si on essaie de tout dissocier, non un conseil municipal, ce n'est pas un ensemble de personnes. Et à un conseil municipal, on associe un ensemble de personnes. Pourquoi ? L'ensemble de personnes, on va le gérer comme un ensemble mathématique. Alors qu'un conseil municipal, on peut lui attacher ses fonctions, son rôle...et donc ne pas simplement considérer que c'est un ensemble de personnes. L'ensemble de personnes, c'est seulement un attribut, une facette d'un conseil municipal. Voilà le genre de discussion typique. Et pareil, qu'est-ce qu'on fait avec Rennes ? C'est une ville, c'est une commune ? Donc là, on nous demande un peu de trancher différemment par rapport aux discussions qu'on avait eues avant. Pourquoi ? Parce qu'au début, on avait mis ça comme une commune et après on a eu des discussions avec Marie-Paule nous disant que la notion de lieu était totalement fondamentale pour parler de la notion de « quartier » ou de « lieu touristique » si un jour on veut étendre un peu l'ontologie. [...] Il faut faire de nombreux choix. Donc tout cela, ce sont des choses sur lesquelles il a fallu trancher. Dès que quelqu'un modifie, on le voit parce qu'on travaille sur un outil coopératif, une forge, sur lequel il y a tout un tas de répertoires emboîtés les uns dans les autres. Nous, dès qu'on fait une modification, on la met dedans. Donc les gens savent quelles sont les modifications dès qu'ils se connectent.

À part les usages, avez-vous eu d'autres points qui ont dirigé la construction de l'ontologie ?..par exemple, la simplicité, la complétude, l'exactitude…

Simple, oui ça c'est un critère, il faut que ce soit simple. Ça veut dire par exemple qu'on essaie de limiter l'héritage multiple, c'est-à-dire une classe qui est sous-classe de plusieurs classes. Ça c'est un peu compliqué...mais si on ne veut pas ça, ça veut dire qu'il faut créer beaucoup de relations. Donc il y a une sorte de compromis à faire : dans ce qu'on propose, il y a peu d'héritage multiple mais on rajoute beaucoup de relations à chaque fois.

Et donc, c'est plus simple pour la machine ou pour les utilisateurs ?

C'est complexe pour les utilisateurs. Comme cette ontologie est destinée aux utilisateurs, c'est avec cette ontologie qu'ils interrogeront toutes les bases, qu'ils feront leurs requêtes. Ils utiliseront le vocabulaire de l'ontologie donc il faut à tout prix qu'il soit le plus simple possible...enfin le plus compréhensible possible. Pour l'instant, il est faible, on s'est limités pour l'instant à « conseiller municipal » et à quelques notions, il n'y a même pas une centaine de concepts et soixante relations je crois. Mais ça va vite grossir dès qu'on va l'appliquer à autre

chose. Si on voulait prendre simplement toutes les démissions d'organismes politiques, déjà on passerait à 200/300 concepts…

Et comment avez-vous peuplé l'ontologie ?

Je ne peux pas vous en dire beaucoup plus car la démarche est itérative avec des évaluations à faire le plus souvent possible. Donc qui dit évaluation dit des bons échantillons, ça veut dire des échantillons des utilisateurs, des questions, des évaluateurs...de tout ce dont on a parlé. Par exemple, pour les bibliothèques, il a fallu passer beaucoup de temps avec « des bibliothécaires experts » donc des conservateurs du patrimoine qui forment les documentalistes, qui sont par exemple professeurs à l'ENSSIB. Ils ont été obligés d'évaluer les résultats fournis par le système et ça, ça prend un temps absolument considérable. Si on n'a pas prévu ces phases dès le début, ça se casse la figure. L'idée, c'est d'avoir prévu dès le début d'évaluer le plus souvent possible. Et avec toujours dans la boucle les utilisateurs finaux. Même s'ils ne comprennent pas tout, c'est hyper important de toujours revenir à ce qu'ils souhaitent, ce qu'ils en pensent. Il ne faut pas attendre dans son coin, faire un prototype et finir en disant « vous êtes contents ? ».

Est-ce que vous avez déjà pensé l'interface pour les usagers ?

C'est un truc qu'on ne fait pas du tout. Il y a des gens qui sont spécialistes. Par exemple, pour le projet sur les bibliothèques, c'est des gens de l'ABES qui ont des gens qui font une interface et des graphistes, pour respecter les normes, les habitudes, des utilisateurs. Et dans le cadre d'ICODA, il y a des équipes qui ne font que ça. Ce sont vraiment des domaines à part entière.

J'ai cru comprendre que dans notre service[135], on travaillait sur une ontologie de haut niveau, qui n'était pas seulement centrée sur les démissions des élus municipaux, du coup on essaie d'élargir l'ontologie sur le commerce, sur d'autres domaines, etc.. Du coup l'ontologie à Ouest-France n'a pas forcément les mêmes objectifs. Comment vivez-vous cette différence, cette divergence de points de vue ?

Normalement, il ne devrait pas y avoir de problème puisque les niveaux hauts devraient se ressembler. Et dès qu'il y a des parties communes, il faudrait se mettre d'accord. Puisqu'il y a plein de choses qui sont pareilles. Et sur le niveau de l'ontologie, faudrait se mettre d'accord. Et ensuite, ça ne serait que fusionner ou rajouter des branches. Ça ne devrait pas être très difficile à faire, ça peut prendre un peu de temps...Généralement, le haut des ontologies est très cohérent parce que

[135] Service Banque de Contenus

les gens s'en servent pour faire les mêmes ontologies standards. Par exemple, ils utilisent FOAF pour les personnes ou DublinCore, enfin les trucs de base. Tout le monde utilise un peu ça donc ce n'est pas très très compliqué de se mettre d'accord sur le haut niveau. Nous on a utilisé FOAF pour les personnes dans cette ontologie, on a pioché, on essaie de suivre les standards tant qu'on peut, c'est plus agréable pour tout le monde.

Est-ce que vous avez mis en place de bonnes pratiques dans votre groupe projet pour construire l'ontologie ?

Oui, oui bien sûr. Il y avait une petite note qu'avait faite Marie-Laure Mugnier[136] dessus. On a un outil qui s'appelle COGUI et nos bonnes pratiques découlent de l'utilisation de cet outil. Cet outil est basé sur un travail théorique, on a écrit un livre avec Marie-Laure dessus, sur la représentation de connaissances et de raisonnements par les graphs. Ça a été le point de départ, déjà là, il n'y avait pas de « norme » au niveau des concepts, des noms donnés aux objets. Ça, on l'a implémenté dans un outil qui a renforcé ça avec une syntaxe. Là, nos bonnes pratiques sont complètement issues d'un modèle théorique qui a été implémenté dans un outil et qui est utilisé. C'est pour ça qu'on a mis qu'une relation doit commencer par une minuscule, les noms de classes commencent par une majuscule, il y a plein de gens qui font ça. Mais il n'y en a pas plus de pratiques que ça.

Et quand vous dites « il y a plein de gens qui font ça », est-ce que vous vous rappelez de la première fois où vous avez entendu parler de ces bonnes pratiques ?

....Alors, c'est difficile à dire. Moi, ça va faire trente ans que je fais des systèmes experts ! (rire) J'ai commencé à travailler avec des médecins et j'ai l'impression que dans les différentes communautés avec lesquelles j'ai travaillé, avec des médecins, des agronomes...Enfin j'ai travaillé dans beaucoup de domaines très différents les uns des autres. Maintenant, il y a des normes communes qui font partie du folklore..De toute façon, c'est tellement facile en informatique de passer d'une syntaxe à une autre, on sait faire des traductions. Il y a eu plein de choses qui ont été écrites sur « comment construire une ontologie », des méthodologies, des choses comme ça mais...pfff, je vais dire c'est assez creux. (rire)

[136] Enseignante-chercheuse au Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM)

Et pourquoi, selon vous, c'est « assez creux » ?

C'est souvent que du bon sens, c'est pas beaucoup plus que du bon sens. Et un des derniers trucs que j'ai vu, ce sont des gens qui se sont servis de Protégé tout le temps et donc en fin de compte, c'est l'équivalent d'un manuel d'utilisation de Protégé. Pour nous qui partons de notre modèle théorique et de notre outil, on pourrait faire ça aussi et ça serait un manuel d'utilisation de COGUI. Des fois, des gens ont construit des outils et après, quand ils font « une méthodologie de construction d'ontologie », ça revient à une description de leur outil.

Et je ne sais pas si vous avez un peu lu sur des méthodologies de construction dites « officielles » comme ARCHONTE, Terminae…?

Ça ne me dit strictement rien. Ou alors si j'ai lu dessus, c'était il y a longtemps, ou alors j'ai tout oublié parce que ça ne m'apportait rien. Quand on connaît le modèle mathématique auquel on veut arriver, ça fonctionne ! C'est toujours pareil. Nous on sait sur quoi on veut tomber, on veut tomber sur de la logique. Par exemple, on peut tomber sur de la logique du premier ordre avec des négations particulières, pas trop compliquées. Donc sachant ça, le passage de connaissances verbalisées vers le modèle mathématique issu de la logique du premier, c'est presque du bon sens. Entre ces deux niveaux, il y a des outils pour manipuler des ontologies et si on connaît relativement bien le modèle mathématique qui est mis en œuvre et la logique du premier ordre, le reste c'est bon sens.

Donc selon vous, il faudrait quand même avoir des capacités informatiques et mathématiques assez poussées pour réfléchir sur les ontologies ?

À mon avis oui, ça c'est clair ! On ne peut pas construire une ontologie sans connaître un minimum de logique du premier ordre. Parce que si des gens [sans connaissance] construisent quelque chose, ils ne sauront pas comment cela sera utilisé ensuite. Donc c'est un peu délicat...C'est un peu comme faire des opérations sur les nombres sans connaître l'addition ou la soustraction. (rire) Ok, il peut y avoir une machine qui fait, mais il faut bien savoir à quoi ça correspond. Dès qu'on dit « ontologie », c'est pour faire du raisonnement, donc il faut connaître un minimum de choses sur les raisonnements, qui seront faits par des machines.

Et pour construire ces ontologies, vous avez toujours travaillé avec des informaticiens ou des spécialistes des mathématiques alors ?

Non pas toujours, par exemple, pour le projet sur les bibliothèques, les bibliothécaires apportaient toute la partie sur la verbalisation, la conceptualisation et ils apportaient des connaissances, ça c'est fondamental. La représentation, la modélisation, c'est nous qui l'avons fait...en discutant avec eux ! Souvent, par exemple, avec les médecins, un des trucs catastrophiques, c'est qu'ils ont du mal à se rendre compte avec leur mode de raisonnement hypothético-déductif de ce qu'est l'implication : « si alors ». Quand on leur dit « si a, alors b », ils nous disent « ah non, je ne vous ai jamais dit ça ». Ils ne comprennent pas ça. Alors il faudra mettre des exceptions, etc. sinon on va arriver à des catastrophes. Les gens qui construisent des ontologies sont obligés d'avoir des connaissances en logique du premier ordre donc. Sinon, ils font une simple taxinomie ! Une ontologie, c'est un ensemble de classes, de propriétés entre ces classes, un ensemble d'individus pour remplir les classes et puis c'est des règles, des contraintes. S'il n'y a pas de règle, c'est forcément des ontologies pauvres, il faut des règles. Tout ça s'exprime en logique.

Et au final les bibliothécaires ou les médecins ont réussi à comprendre ce que vous leur avez dit sur la logique ?

C'est pourquoi, parfois, il faut des intermédiaires, je vous ai parlé des quatre niveaux. Parfois, il faut cet intermédiaire, cet ingénieur de la connaissance, même si je n'aime pas du tout ce terme, il est un peu ancien. Il faut des personnes capables de faire de lien entre les raisonnements implémentés dans la machine et ce qu'ont dit les experts, les gens du domaine. Mais on fait tout ici dans notre équipe. Notre démarche est de faire la théorie, les outils et d'aller jusqu'aux applications réelles. Par exemple, quand on a travaillé avec des agronomes sur le comté, il y a des gens dans l'équipe qui connaissent l'informatique. Ça c'est super important ! Avec Ouest-France aussi, on travaille aussi avec des gens qui connaissent l'informatique. Une même personne peut intervenir ou jouer deux rôles mais il y a vraiment quatre rôles différents : les utilisateurs, les acteurs du domaine, des personnes capables de faire le lien entre la théorie du domaine et sa modélisation mathématique puis enfin les gens qui connaissent bien le modèle mathématique et qui sont capables d'informatiser. Mais voilà, souvent les gens ont plusieurs casquettes mais tous ces rôles sont joués même s'il n'y a que deux personnes dans l'équipe. [...]

Du coup, j'ai abordé tout ce que je voulais. Est-ce que vous avez quelque chose à rajouter ?

Les points sur lesquels je voulais insister c'est : le côté itératif, le côté évaluation à chaque instant, à chaque étape puis la difficulté de l'évaluation. S'il y a quelque chose à développer au niveau de

la méthodologie, à mon avis c'est sur l'évaluation. Trouver les bons échantillons et ne pas attendre la phase finale pour faire l'évaluation. Il faut faire de l'évaluation permanente en ayant de bons échantillons. Par exemple, il faut savoir qu'est-ce qu'on mesure, quelles sont les requêtes test ? qui les construit ? etc.

Dès qu'on a commencé à travailler sur les systèmes à base de connaissances dans les années 1970, début des années 1980, il y a eu tout un tas de livres écrits sur ce sujet. Et il y avait une approche très séquentielle, étapes par étapes. Dès que c'est un domaine un peu complexe, qui n'est pas très bien formalisé, modélisé, ça se casse la figure. Quand on utilise des ressources en langue naturelle, c'est absolument fondamental d'avoir une approche itérative.

Annexe n° 9 : Entretien avec Jean Charlet (projet LERUDI)

4 juin 2018, 15h, 33 min

[...]

On va se centrer plus particulièrement sur le projet LERUDI, quand a commencé ce projet et quel est son état d'avancement ? J'ai vu par exemple qu'un article de 2017 parlait de son évaluation.

Il a commencé début 2009. Alors, ce projet n'a fait que la preuve de concepts. Maintenant, il serait prêt à fonctionner. Donc ça, ça avait été fait dans le cadre du DMP, le Dossier Médical Personnel qui est devenu le Dossier Médical Partagé, une espèce de serpent de mer, il ne se met jamais en place mais en fait il existe. Je crois qu'il y a 700 000 personnes qui ont un DMP maintenant en France.

Et dans cet article, j'ai vu que l'ontologie servait à catégoriser les contenus et à mesurer la proximité entre deux termes. Est-ce que vous pouvez m'en dire plus sur l'utilité de l'ontologie dans ce projet LERUDI ?

Les ontologies en santé vont servir à plusieurs choses : il y a celles qui servent à des systèmes d'aide à la décision et celles qui servent à des systèmes de recherche d'information et LERUDI est dans le deuxième cas. Donc dans LERUDI, l'ontologie est au cœur d'un système de traitement automatique des langues qui sert à indexer les contenus des documents du dossier patient. Quand on fait du TAL pour repérer les contenus, on a deux types d'approche qu'on appelle les approches numériques qui sont très très à la mode ces temps-ci. Les approches symboliques qui cherchent à repérer des contenus par rapport à des modèles faits à priori, en plus ces modèles peuvent être des thésaurus, des classifications. Et si on fait les choses de manière un peu plus carrée, un peu plus formelle, ce sont des ontologies. C'est l'ontologie qui sert dans le système de TAL pour indexer des documents.

Qui a eu l'idée d'implanter une ontologie dans cette application ?

C'est moi. Ça a été acté assez rapidement parce que j'ai été dans des réunions de travail avec l'ASIP Santé[137] qui était le financeur. Et quand il s'est agi de décider de l'approche, on était tous tombés d'accord dans les groupes de travail sur le fait qu'on allait faire du symbolique et il y a eu quelques discussions scientifiques pour savoir si on réutilisait des référentiels qui étaient déjà disponibles

[137] Agence des Systèmes d'Information Partagés de Santé

ou si on reconstruisait une ontologie pour le projet et c'est moi qui a eu l'idée de reconstruire une ontologie.

Et pourquoi avoir fait ce choix pour le projet ?

Dans de nombreuses applications, on s'aperçoit que si on n'a pas d'ontologie, le système de recherche fonctionne moins bien. En gros, dans tout ce type de systèmes, on sait qu'une ontologie marche mieux. Après, il faut mettre en balance le coût du développement de l'ontologie, et ça peut souvent être un problème. Ça peut aussi être un problème de compétences parfois parce que les gens n'ont pas les compétences. Et donc un problème de coût parce que dans tous les cas de figure, on a des gens qui travaillent des mois et des mois sur l'ontologie.

Et vous avez dit aussi que vous aviez choisi de ne pas réutiliser un autre référentiel, un autre thésaurus...de partir de zéro pour cette ontologie ? Pourquoi ce choix ?

Ce n'est pas tout à fait « partir de zéro » même si on décide de ne pas réutiliser une autre ontologie. Il n'empêche que quand on construit une nouvelle ontologie, on regarde toujours ce qui a déjà été fait dans le domaine. Donc, d'une certaine manière, on réutilise les thésaurus et les classifications du domaine pour aller piocher dedans ce qui nous intéresse.

Comment avez-vous fait pour construire cette ontologie ? Êtes-vous partis de textes comme dans beaucoup d'ontologies médicales que vous avez dirigées ?

On est tout d'abord partis de textes, mais rapidement on s'est aperçus qu'on manquait de textes. Parce qu'on avait des articles scientifiques mais on n'avait pas beaucoup de documents sur les patients. À l'époque, les services d'urgence avaient très peu d'applications qui permettaient de générer des textes qu'on aurait pu réutiliser pour faire de la fouille de textes. On n'avait pas assez de matériaux, de documents écrits pendant l'activité. Les articles scientifiques représentent le domaine de la recherche sur les urgences, alors que les articles sur l'activité représentant le domaine tel qu'il se pratique. On avait besoin plutôt de la partie « représenter le domaine tel qu'il se pratique ». On a quand même pu travailler avec les quelques concepts qui nous été proposés par la fouille de textes sur les articles scientifiques. Puis ensuite on a réessayé de travailler avec les référentiels qui étaient existants. Et un médecin urgentiste en particulier a été très très présent pour ce travail. Quand on a de bons corpus pour travailler, on sollicite les médecins mais pas tant que ça, on peut pas mal travailler sans eux. Dans ce cas-là, le médecin-urgentiste, qui est un des signataires des articles, a été très très souvent sollicité.

Oui, lors du précédent entretien téléphonique, vous m'aviez dit que sur chaque projet, il y avait au moins un médecin qui était très impliqué dans le projet et qui avait été formé par vous sur les ontologies pour vous aider…

En l'occurrence oui et dans ce projet encore plus que d'habitude.

J'ai vu que sur votre ontologie de la pneumologie, vous vous étiez inspirés de la méthodologie ARCHONTE et que vous l'aviez retravaillée ?

Alors la méthodologie ARCHONTE...c'est une méthodologie qui a été mise au point avec un collègue Bruno Bachimont, co-signataire d'un certain nombre d'articles, qui est à la fois informaticien et philosophe. Oui c'est un peu comme ça qu'on travaille, ARCHONTE nous a surtout aidés à mettre au point la raison pour laquelle on organise la hiérarchie des concepts. Une partie de cette méthodologie est visible dans le cadre d'un plugin qui avait été développé pour l'éditeur d'ontologies Protégé.

Et si on repasse sur l'ontologie développée dans le cadre du projet LERUDI, quels outils ont été utilisés ? Comme des mindmap ou Protégé..

On n'a pas utilisé des *mindmap* non, Protégé, évidemment, parce qu'à la fin, on construit toujours l'ontologie avec Protégé. Des outils de fouille de textes aussi pour fouiller des textes d'articles scientifiques. En 2009, ça devait encore être un outil qui est difficilement disponible maintenant, qui s'appelle SYNTEX-UPERY. Des fichiers Excel parce que souvent, on range des tas de choses dans les fichiers Excel avant de les mettre dans Protégé.

Quand je vous dis « méthodologie de construction d'ontologie », qu'est-ce que ça évoque pour vous ? Est-ce que par exemple dedans, vous mettez le peuplement de l'ontologie ?

Dans les ontologies en médecine, y n'y a pas trop cette question de peuplement d'ontologies parce que dans le modèle, on fait que des classes. Par exemple « l'appendicite de M. Dupont », l'appendicite c'est la classe. Le peuplement c'est quand on met des instances. Nos ontologies à nous n'ont pas d'instances tant qu'elles ne servent pas dans un système. Et encore dans les systèmes dans lesquels elles servent, encore faut-il que ce soit des systèmes qui fonctionnent avec des classes et des instances, ce qui n'est pas toujours le cas.

Par exemple, est-ce que dans la méthodologie de construction d'ontologie, vous incluez l'étape d'évaluation aussi ?

Oui parce que dans toutes ces étapes, il y a toujours une évaluation à la fin. C'est ce qui est important dans cette méthodologie. Le plus important et ce qui était original, qui l'est moins maintenant, parce que les gens suivent cette contrainte, c'est le fait de structurer l'arbre des classes et concepts en réfléchissant aux concepts différentiels. Bien vérifier que les concepts qui sont sous un même niveau, sous un même père partagent un même point de vue.

Pour l'ontologie de projet LERUDI, avez-vous porté votre attention sur des points précis ? Par exemple, est-ce qu'il fallait que l'ontologie soit exhaustive, simple à utiliser, minimaliste…?

Ça dépendait, « simple à utiliser », non. Il y avait une branche qui était décrite dans un article, on savait qu'il fallait qu'on soit exhaustifs sur les médicaments, il fallait faire attention à ce que la branche des médicaments soit complète. Sur les maladies, on a fait attention avec le médecin urgentiste avec lequel on a travaillé. Effectivement, la partie complétude était très importante sur les médicaments. Sur le reste, la méthodologie d'évaluation a vraiment été importante, on a vraiment travaillé avec les médecins. Dès l'instant que les médecins étaient impliqués dans le processus de construction, que le système était évalué... on voyait rapidement que certains documents étaient mal indexés et on vérifiait avec les industriels si c'était la faute du système, si le système de reconnaissance fonctionnait mal...ou si c'était la faute de l'ontologie à qui il manquait des classes. Ce qui manque dans les ontologies de recherche d'information, c'est que souvent on loupe certains concepts parce qu'on n'a pas les bons termes, les bons synonymes associés aux concepts. On a fait des évaluations et on a enrichi l'ontologie quand on repérait des manques lors de la campagne d'évaluation.

J'ai aussi lu dans un de vos articles que le niveau de granularité dépendait aussi beaucoup du professionnel à qui l'ontologie était destinée. Par exemple, pour un spécialiste qui travaillait sur les nerfs, il fallait vraiment porter attention aux nerfs alors que sur une ontologie qui était destinée à un médecin généraliste, on développerait peut-être moins cette branche…

Oui, quand on fait des ontologies en santé, un médecin, quelle que soit sa spécialité, regarde le patient dans son entier. Donc on peut dire qu'il y a un modèle qui tient à peu près compte de tout ce qui concerne le patient. Bien sûr, pour un neurologue, on va s'intéresser au cerveau et aux nerfs. Les urgences ont ça de spécifique qu'elles s'occupent de tout le corps humain à un certain niveau de granularité, à un niveau moins élevé que dans certaines situations. Ça dépend un peu mais il y a des problèmes de nutrition qui vont être moins importants dans le domaine des

urgences alors que la partie cardiaque va être très développée : c'est une des raisons de prise en charge en urgence. Donc on va être plus ou moins granulaires selon les domaines. Et dans les urgences, on est granulaires dans un certain domaine qui correspond à des urgences potentielles.

Et pour connaître les usages des spécialistes à qui l'ontologie sera destinée, comment faites-vous ?

Alors des spécialistes, il n'y en a pas tant que ça. Par exemple, dans le cadre du projet LERUDI, il y a eu souvent des réunions de montage de projet avec les médecins urgentistes. Les médecins sont souvent très très occupés par leur activité. Si on arrive à dégager à temps, voire même qu'on décide dans le financement qu'un médecin soit payé pour travailler sur l'ontologie pendant un quart de son temps...c'est déjà très bien pour nous. On essaie de mettre en tête de pont un ou deux médecins qui vont vraiment être là, comme je l'ai dit auparavant.

Est-ce qu'il y a quelquefois des point questionnants concernant la collaboration avec des spécialistes, même si cette collaboration est brève ?

On n'a pas eu de point bloquant sur l'Ontolurgences pour LERUDI parce que le médecin qui a été impliqué était déjà intéressé par les choses, avait déjà fait un master en informatique médicale. Dans d'autres domaines, effectivement, il y a une espèce d'acculturation, pas de point bloquant, simplement on doit prévoir un temps de formation avec les médecins pour leur montrer ce qu'on fait. Par exemple, souvent, j'arrive à les faire assister à des enseignements que je donne en master pour qu'ils n'aient pas d'appréhension sur les ontologies.

Est-ce que vous avez déjà collaboré avec des documentalistes où des personnes d'autres corps de métier lors de la construction d'ontologies ?

Oui, dans le cadre de l'ontologie ToxNuc sur la toxicité nucléaire sur laquelle on a travaillé, il s'est trouvé que c'était une ontologie qui avait été construite avec des chercheurs en toxicité nucléaire. Il fallait la refaire, la réorganiser, vérifier les sources. Et on a travaillé avec une documentaliste qui s'appelle Anne-Claire Le Picard, elle a beaucoup aidé. Donc l'ontologie était déjà plus ou moins construite. Il y a eu un gros travail sur la qualité des sources qui a été fait grâce à elle.

Ok, et dans quels cas vous choisissez de réutiliser un thésaurus ou une autre RTO et dans quels autres cas, vous choisissez de développer et d'intégrer vos propres concepts ?

En pratique, on est dans une optique où la plupart du temps, on se ressert plus ou moins des thésaurus mais on va quand même refaire le modèle. En médecine, on sait qu'on a des référentiels incontournables dès qu'on commence à travailler. Donc il y a deux grandes classifications utilisées : la CIM 10 (Classification Internationale des Maladies) et la CCAM (Classification Commune des Actes Médicaux). De toute façon, on sait qu'on va regarder ce qu'il y a là-dedans par rapport à la spécialité qui nous intéresse quand on va faire notre ontologie. C'est principalement ces deux-là.

Est-ce que par rapport aux années 2000, vous avez l'impression que les méthodologies de construction sont plus itératives, qu'elles sont moins linéaires ? Ou sont-elles à peu près semblables ?

Pour moi, c'est à peu près semblable. Pour nous, je sais que pour peu qu'on ait un expert disponible, on va de plus en plus vite en fait. Parce qu'on caractérise mieux ce qu'on veut, on réutilise toujours le même haut pour les ontologies. On forme mieux les gens, les gens sont plus habitués parce qu'ils ont entendu parler de la question des ontologies. Globalement, on a tendance à aller plus vite qu'il y a quinze ans.

Au début, les spécialistes écrivent une note avec les usages qu'ils voudraient voir développer ou cela se fait de manière orale ?

Non, ça dépasse le problème de l'ontologie. L'ontologie va servir dans un système d'aide à la décision dans un système de recherche d'information. Ce système, ça ne dépend pas de nous. Dans certains cas, ça dépend de nous et à ce moment-là, on fait des spécifications de ce que doit faire le système. De ce que devra faire le système, découlera l'organisation de l'ontologie.

Et une dernière question...Comment faites-vous pour aligner des ontologies ou d'autres ressources quand vous en avez besoin ? Par exemple, pour reprendre une classification existante pour refaire une nouvelle ontologie…

Alors il y a deux choses dans cette question, il y a le fait de reprendre des morceaux. Il peut nous arriver de reprendre des classifications..ou directement des ontologies ou des thésaurus en SKOS. On fait un peu de bricolage informatique pour récupérer des concepts dans le langage OWL. Et ensuite on travaille sur ce fichier qui nous donne les classes qui correspondent aux besoins des urgentistes.

Annexe n° 10 : Entretien avec Florence Amardeilh (projet Open Food System)

11/06 10h 1h

[...]

On va parler plus particulièrement du projet *Nos recettes* de Open Food System (OFS). Sur quelles étapes de ce projet avez-vous collaboré ?

J'ai collaboré dès le départ, dès la soumission du projet car en tant que directrice R&D, ma fonction à Mondeca était de pouvoir mettre en place des projets collaboratifs avec des entreprises, des PME, des TPE...et de travailler sur des projets innovants. Donc j'ai participé à la constitution du dossier de l'appel à projet en collaboration avec SEB qui était le coordinateur du projet OFS. J'ai aussi travaillé avec d'autres partenaires académiques, dont l'Université Paris 13 avec le LIMICS[138] qui est le partenaire avec qui j'ai collaboré sur la tâche de création de l'ontologie. Le projet a duré quatre ans.

Qui a eu l'idée en premier l'idée d'utiliser une ontologie pour ce projet ?

En fait, on était déjà en discussion avec SEB, j'avais fait des sortes de *meetup* où on réunissait des gens intéressés par les ontologies, pour leur présenter un peu ce qu'était une ontologie. Les gens de SEB ont assisté à ces réunions et ont tout de suite pensé qu'effectivement c'était très valorisant pour le projet par rapport à ce qu'ils voulaient faire avec ces recettes numériques. Et dès le départ on a conçu le projet ensemble comme ça, autour des ontologies au service des recettes.

Et quelle est l'utilité de l'ontologie au sein du projet *Nos recettes* ?

L'utilité de l'ontologie, c'est de pouvoir structurer les recettes, de manière plus lisible pour qu'on puisse voir comment est constituée une recette, pour pouvoir distinguer les notions d'entités, d'ingrédients. Par exemple, « qu'est-ce qu'un ingrédient ? », « pâte brisée », c'est un ingrédient dans la recette de la quiche, mais c'est aussi une recette à part entière qui elle-même peut se décliner. Et après dans les instructions, il faut aussi bien repérer les actions, les ingrédients en jeu dans l'action et le matériel utilisé. Donc ça nous permettait de pouvoir décrire très très finement les concepts et ensuite de pouvoir les annoter avec des outils de Traitement Automatique du Langage pour pouvoir au final les structurer et pouvoir travailler dessus pour rendre différents services : soit de recommandation, soit d'amélioration au niveau du moteur de recherche, soit

[138] Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-Santé

même de pouvoir, avec les appareils de SEB, dire « telle recette peut être liée à tel matériel de cuisine, comme les Cookeo » par exemple...On peut remplacer telle instruction par telle autre et il y a du raisonnement derrière qui devait être joué pour pouvoir vraiment améliorer les recherches et les adaptations sur les recettes.

Et vous avez dit qu'elle servait « à l'amélioration du moteur de recherche », pouvez-vous expliquer cela ?

Comme on a une description assez fine de tous les aliments qui participent à la constitution des ingrédients de recette, justement on arrive à les repérer plus facilement et donc on peut vraiment, au niveau du moteur de recherche, soit faire de l'extension sémantique : on recherche un aliment « Patate » et du coup, on va trouver toutes les recettes qui contiennent « Pomme de terre » et vice versa. Donc on peut travailler également sur les mots, si on recherche « enfourner », je sais que derrière, il y a « four » donc ça peut me permettre de rebondir sur toutes les recettes avec un four. Donc on peut effectivement avoir une recherche plus précise, c'est pas juste par un mot-clé qui est lancé comme ça. « Moule » par exemple, qu'est-ce que c'est ? la moule, l'aliment ou le moule, l'ustensile de cuisine, on peut arriver à désambiguïser pas mal de mots, pas mal de termes employés en cuisine. On peut donc affiner les résultats en fonction de ce que les gens cherchent aussi. Et on peut faire des calculs aussi, par exemple avec « est-ce que les gens veulent des recettes ? avec ou sans fromage ? » parce que peut-être qu'ils sont allergiques. Et on peut donc aller chercher dans la constitution de certains sous-ingrédients s'il y a effectivement du fromage utilisé, même si l'aliment n'apparaît pas tel quel dans la recette.

Et à quel stade se situe maintenant le projet *Nos recettes* ? Je n'ai pas trouvé d'information récente sur le site…

Oui, le projet de recherche en tant que tel est arrêté aujourd'hui depuis juin 2016. Maintenant SEB continue à travailler sur ces aspects-là avec un partenaire qui est Orange et avec Mondeca. Maintenant, ça s'appelle *Cooking Recipe*. Ils sont en train d'industrialiser tous les travaux de recherche qu'on a faits petit à petit. Ils le font de manière itérative dont c'est basé sur l'ontologie qu'on a créée. C'est en train d'être révisé régulièrement...C'est un gros enjeu d'arriver à avoir cette structuration et de pouvoir être interopérable, de pouvoir proposer des API.

Maintenant, une question assez large : comment avez-vous fait pour créer cette ontologie ?

Alors on a travaillé par modules, parce que c'était vraiment une très très grosse ontologie, c'est sur tout le champ de la cuisine, pas seulement sur les recettes. Donc on a décidé de la compartimenter en modules, on a six-sept modules, il y en a un qu'on n'a pas trop développé qui est celui de la personne. On a le modèle « cuisine » qui regroupe la vue haute sur une recette : « qu'est-ce qu'une recette ? ». On a aussi les aliments, le beurre, la viande, l'huile. On a « les préparations », comme le pain, les confitures, les sauces...On a le module « nutrition », il fallait qu'on puisse calculer pour chaque recette les valeurs nutritionnelles. On avait tout le module « matériel » puis le module « d'actions » puis un dernier module pour tout ce qui est « représentation organoleptique », les sens, les saveurs, les goûts, le craquant, le gratinage...Et le cœur, c'est vraiment la cuisine, les aliments, les actions, les matériels.

Qui a eu l'idée de faire plusieurs modules comme ça ?

C'est Sylvie[139], la chercheuse du côté LIMICS et moi-même qui avons préconisé cette approche-là parce que l'ontologie est assez importante. Ça nous semblait être une bonne approche vu la taille que ça allait être.

Et concrètement, pour constituer cette ontologie constituée de différents modules, quelle a été votre méthode ? Avez-vous utilisé une méthodologie que vous connaissiez déjà par exemple ?

Oui on a utilisé la méthodologie NEON qui a été faite par des Espagnols il y a une dizaine d'années de ça, une méthodologie qui explique la manière de s'y prendre pour créer une ontologie. Sachant qu'en fait c'est déjà beaucoup : a) aller voir ce qui existe, voir s'il n'y a pas certains modules desquels on peut s'inspirer et on voit ensuite si on peut les intéropérabiliser avec nous. La chercheuse Sylvie avait déjà travaillé sur un projet qui s'appelait Taaable donc elle avait déjà un embryon d'ontologie sur ce projet-là. Mais dans tous les cas, ce qu'on avait récupéré ne collait pas tout à fait avec les exigences de finesse qu'on devait prendre en compte pour le projet. Donc après, à partir de cette première idée, on a travaillé avec différents chercheurs des autres domaines, des nutritionnistes, des anthropologues, des chefs cuisinier de la structure Paul Bocuse, des gens de SEB...on a travaillé avec tout ce monde là pour que dans chaque module, on arrive à caractériser le domaine. On a aussi regardé beaucoup de livres de cuisine, des livres de référence dans la formation des cuisiniers…

[139] Sylvie Desprès, enseignante-chercheuse

Pour collaborer avec différents partenaires comme vous avez dit, comment avez-vous pris en compte ce qu'ils avaient à vous dire ?

Ça s'est fait sous forme de réunions donc on est allées les voir à plusieurs reprises. C'est plutôt Sylvie qui a fait l'interface avec eux, avec les nutritionnistes ou les chefs cuisinier. Moi je venais régulièrement pour les synthèses. Sylvie les voyait plus. En fait, il y avait un *work package*, un lot de tâches qui était spécifiquement dédié à ça et dont Sylvie était la référente. C'était moi qui faisais la partie opérationnalisation, donc qui mettais en jeu l'application, je devais bien utiliser ce qui avait été modélisé. Il y avait effectivement des phases de conception faites exclusivement avec Sylvie et des phases de validation qui ont été faites avec toute l'équipe.

Est-ce que vous pouvez me présenter les étapes de la méthodologie NEON ?

Alors il faut déjà regarder dans ce qui se fait déjà. Ensuite il faut pouvoir adapter à notre cas particulier. Il faut développer...Et après il faut pouvoir potentiellement étendre et aligner avec les ontologies d'autres domaines. Et il y a une dernière phase qui doit être de tester, valider l'ensemble.

Qui a eu l'idée d'utiliser cette méthodologie ?

C'est Sylvie. Sachant que c'est une méthodologie qui est assez commune en conception d'ontologies.

Vous avez dit que l'embryon d'ontologie Taaable ne répondait pas à vos exigences de finesse donc comment l'avez-vous adaptée pour faire l'ontologie de *Nos recettes* ?

Alors, on a regardé comment elle était structurée, on a confronté par rapport à ce que nous disaient les experts métier, nutritionnistes et autre. On voyait qu'il y avait plein d'informations qui manquaient. Par exemple, les aspects nutritionnels étaient très peu abordés alors qu'il fallait qu'on rentre dans plus de détails pour pouvoir justement calculer des valeurs nutritionnelles. Les objectifs du projet Taaable n'étaient pas les mêmes que pour le projet OFS. Ça participait aux différences. Et le projet Taaable était juste un projet universitaire, académique, les exigences industrielles portées par SEB n'étaient pas les mêmes. On s'est rendu compte que cette ontologie ne répondrait pas aux objectifs de SEB.

Donc vos objectifs différaient puisque l'ontologie devait être dans une application industrialisée…?

Tout à fait oui. On avait un fort besoin, après ça a aussi créé des complications on va dire parce...il y avait plusieurs *work packages*. Avec Sylvie, on s'est centrées sur celle de la construction de l'ontologie et de la chaîne d'annotation sémantique pour annoter les recettes et les rendre plus riches en informations. Et il y avait un autre *work package* qui était celui de l'opérationnalisation donc tout ce qui est logiciels. C'est le lot le plus technique. Et on avait un autre lot qui était sur les interfaces utilisateurs après les applications des prototypes qui allaient être mises en place. Et le problème dans cette organisation-là, c'est qu'on avait trois prototypes à réaliser, c'était *drivé* par le lot UX. Ces trois prototypes étaient indépendants les uns des autres, les choses présentées n'étaient pas forcément les mêmes. Mais nous, au niveau de l'ontologie, c'était bien la même qu'il fallait faire évoluer pour qu'elle réponde aux différents besoins. Et à chaque fois, on avait très peu de temps lorsque l'info redescendait vers nous au niveau des besoins d'usages. Donc il fallait adapter l'ontologie, la faire évoluer, pour ensuite la donner au lot d'implémentation logicielle pour qu'ils puissent créer le prototype. Donc on avait une dépendance assez forte avec les autres *work packages* pour être sûr à la fin qu'il y ait un prototype qui sorte et qui marche et qui soit industrialisable par SEB. Donc les exigences étaient très très élevées et nous on a été très très contraints parce que normalement pour qu'on élabore une ontologie, surtout dans ce genre d'application, elle doit être désignée pour un usage bien précis. Et cet usage on l'attendait pendant longtemps. Pour ne pas attendre, on essayait de voir avec les experts, avec les gens de SEB à quoi pourrait servir l'ontologie. On essayait d'anticiper les usages qui devaient redescendre après vers nous par les membres de l'UX...mais voilà il y avait toujours des décalages et c'était vraiment pas facile de travailler dans ce contexte-là avec autant de contraintes.

Quand vous parlez d' « exigence industrielle », c'est plutôt l'interdépendance avec d'autres *work packages* ?

Alors oui effectivement il y a ça, puis il y a vraiment aussi le contexte industriel de SEB. C'est-à-dire que l'ontologie puisse répondre à certains besoins, en termes de volume, en termes de rapidité, s'inscrire dans une chaîne qui est dépendante d'autres acteurs : après il y a le moteur de recherche, les interfaces avec le moteur de recherche, les interfaces avec le moteur de recommandations pour que ça puisse enrichir le moteur interne qui était du XLM et pas du RDF pour les recettes, donc on a eu pas mal de contraintes. Tous les systèmes de raisonnement sur l'ontologie peuvent se faire dans un temps limité, c'est plus performant, ça demande moins de temps. Donc il y a la phase de construction de l'ontologie mais une fois qu'elle est construite cette

ontologie, il faut qu'elle serve à quelque chose. Elle va servir dans la chaîne de traitement linguistique pour annoter les recettes, ensuite elle va être exportée et exploitée dans les autres briques de l'application. Lorsqu'on va devoir l'interroger, elle va répondre du tac au tac. Il y avait cette exigence très très forte de rapidité sachant que l'ontologie est très volumineuse et très complète. [...] Dans une architecture qui utilise une ontologie, qui l'exploite notamment pour faire de l'annotation sémantique, donc c'est la chaîne d'annotation sémantique, on va avoir l'ontologie qui va être stabilisée à un moment x, on va s'en servir pour alimenter les outils de TAL comme thésaurus, comme dictionnaires. Par exemple, tous les aliments, tous les matériels, toutes les actions vont être exportés sous forme de vocabulaire et vont venir constituer des dictionnaires sur lesquels vont se baser les règles d'annotation linguistiques. Donc l'annotation va pouvoir avoir lieu en se basant là-dessus et il va en ressortir un résultat qui n'est jamais optimum à 100 % parce que le langage est flou et ambigu...et c'est très bien comme ça ! Et ensuite il y a un *mapping* entre ce qui a été extrait par le moteur de TAL et ce qu'on va vraiment utiliser pour annoter la recette.

Est-ce que vous avez engagé une réflexion sur le vocabulaire utilisé dans l'ontologie ?

À chaque fois, on a essayé d'être le plus précis possible pour les termes qu'on utilisait dans l'ontologie. Et lorsqu'il y avait besoin de définir un ensemble de synonymes, effectivement, on voulait que des gens n'ayant pas les mêmes vues sur la recette puissent quand même comprendre la recette. Entre un nutritionniste, un chef-cuisinier et un utilisateur lambda, ils ne vont pas forcément utiliser le même vocabulaire. Par exemple, pour les ustensiles, les chefs vont sans doute utiliser du vocabulaire plus compliqué que nous, on ne va pas comprendre, pareil pour les pièces de viande. Du coup, on met en place des synonymes qui permettent de donner plusieurs points de vue sur la même entité.

Vous avez dit plusieurs fois que c'était une ontologie très volumineuse, combien y-a-t-il de concepts ?

Il doit y en avoir entre 6000 et 7000. Après ça dépend des modules parce qu'il y a des modules avec beaucoup moins de classes.

Et le fait de ne pas toujours connaître les usages avant d'avancer sur la construction, est-ce que ça a entraîné d'autres points questionnants ?

Alors on a essayé de ne pas se brider lors de cette phase mais ça nous a quand même posé de gros problèmes d'organisation. On n'aurait pas fait l'ontologie de la même manière. Parce que là, du

coup, on l'a fait étapes par étapes. Disons que l'impact était sur les raisonnements attendus et comme on ne savait pas quels étaient les raisonnements attendus par l'ontologie, on ne savait pas comment modéliser l'ontologie pour répondre à ces raisonnements. Donc on avançait sur l'ontologie, on se disait « les aliments, il faut bien les modéliser, etc. » On avançait comme on pouvait et par exemple, après on se rendait compte que la priorité n'était plus du tout mise sur les aliments et plus sur le matériel. Et là branle-bas de combat, il fallait mobiliser les experts. Du coup c'était un peu compliqué de suivre le train et de rattraper les wagons en fonction de ce qu'on nous demandait, au moment où on nous le demandait. Du coup, il y a des choix qui ont été faits, il fallait revenir sur l'ontologie, casser des arbres que l'on avait faits car le point de vue changeait complètement, il fallait refaire.

Et est-ce que vous vous êtes un peu penchée sur l'évaluation aussi ?

Oui, alors c'était plutôt Sylvie qui s'occupait de l'évaluation avec les experts, on faisait des réunions avec les experts pour valider qu'on avait bien un consensus sur les termes. Sylvie, avec son équipe de laboratoire, a implémenté une interface pour que les experts puissent parcourir l'ontologie et valider des briques de l'ontologie. Donc son rôle, c'était vraiment de valider la connaissance qui était modélisée dans l'ontologie. Et ensuite, à Mondeca, on a ensuite fait la validation technique, opérationnelle, on se demandait « est-ce qu'on obtient bien les bons raisonnements attendus ? », « est-ce qu'il faut reprendre la structure des classes et des propriétés parce qu'on n'obtient pas les bons raisonnements ? ». Et oui, on vérifiait si on obtenait bien les bonnes annotations en sortie de la chaîne.

Et comment vous mettiez en commun les résultats de ces deux validations ?

On avait des réunions très régulièrement au niveau de ce *work package*, avec les gens qui intervenaient au niveau de cette chaîne d'annotation sémantique. Ensemble, on présentait les résultats qu'on avait obtenus et l'intégration qui été faite au niveau de la chaîne et les résultats finaux de la chaîne lorsqu'elle marchait. C'était tous les deux mois ces réunions pour valider ces résultats et corriger le tir si besoin. Il y avait à la fois Sylvie du côté académique pur, d'autres partenaires…

Pour cette tâche de construction d'ontologie, quels outils avez-vous utilisés ?

On a utilisé Protégé, le plus utilisé aujourd'hui, gratuit et ouvert à la communauté. On faisait aussi beaucoup de *mindmap* pour pouvoir visualiser des bouts d'ontologies qui nous intéressaient, ceux sur lesquels portaient les raisonnements justement pour pouvoir les partager et les comprendre.

Est-ce que vous avez privilégié une approche pour la construction de l'ontologie ?

Il fallait qu'elle soit la plus complète possible, la plus...utile possible, il y a un vrai service rendu. Il fallait qu'elle soit la plus...opérationnalisable possible et la plus facile à manipuler, d'où le fait de partir en modules. Et facile à maintenir aussi. Après, on n'a pas eu de mots d'ordre spécifique, si on voulait vraiment convaincre SEB que c'était efficace et la meilleure solution pour eux, parce que c'était quand même un projet de recherche...il faut convaincre les gens du business, les partenaires qui portent le truc. Donc, voilà, il fallait qu'on leur montre tout l'intérêt d'utiliser une ontologie, c'est facile à manipuler, etc.

Est-ce que vous avez peuplé cette ontologie ?

Alors il y a deux choses. Dans l'ontologie telle quelle, il n'y a que des classes, des relations et la modélisation. Ensuite à côté, on a la base de connaissances des recettes où effectivement on a joué l'ontologie sur un corpus de 55 000 recettes, donc c'est à peu près une base de connaissances de 55 000 recettes. Donc il fallait à chaque fois que les recettes passent par la chaîne et utilisent l'ontologie. C'est pour ça que les volumes étaient assez conséquents parce que sur une recette, on pouvait se retrouver avec 150 propriétés quand même. Et ça, il fallait que ce soit fait en un minimum de temps. Et donc à la fin, on pouvait s'amuser avec ce corpus de recettes annotées pour pouvoir manipuler, tester et voir si ça fonctionnait quand on faisait de la recherche d'information, du raisonnement…

Ce peuplement automatique a fonctionné tel quel ou il a fallu avoir une validation humaine derrière ?

Alors c'est impossible de valider 55000 recettes. Nous on avait 40 recettes test qui représentaient les différentes caractéristiques de recettes, des entrées, plats, desserts...Donc on a fait des tests de validation sur ce corpus. Par contre, avec les 55 000 autres de manière automatique, on a validé d'autres points concernant l'architecture : la validité, la robustesse de la chaîne, la stabilité, la performance des traitements...là, ce sont des calculs automatiques.

Comment s'est déroulée la collaboration avec les autres personnes avec lesquelles vous avez construit l'ontologie, notamment Sylvie ?

Ça a été fluide entre nous, on se connaît très bien, on sait quelles sont les exigences sur une onto, ça s'est très bien passé. Après, parfois, elle avait le côté modélisation connaissances pures, théoriques. Moi, j'avais plus le côté « oui, mais non, d'un côté pratique, ça ne va pas pouvoir se faire comme ça car il y a telle contrainte technique… ». On en rediscutait et on arrivait toujours à un compromis. Par contre, c'est plus avec les gens de chez SEB, que c'était des fois plus compliqué. Il fallait adopter le même vocabulaire qu'eux, ils ne savaient pas forcément ce qu'était une ontologie et à quoi ça servait, donc il fallait un peu les éduquer, en même temps avancer. Donc là, c'était un peu plus compliqué. Au sein de notre *work package*, avec les différents intervenants de la chaîne de traitement sémantique, que ce soit avec la personne qui travaillait sur le TAL ou celle qui se centrait sur le moteur de recommandations, ça se passait très très bien. On se réunissait très régulièrement.

Et vous dites que parfois vous avez dit que vous avez dû former des personnes de chez SEB qui ne connaissaient pas les ontologies, de quelle manière et quand les avez-vous formées ?

Alors on les a formées au début, avant de commencer la modélisation, on a assuré quelques formations avec Sylvie sur « qu'est-ce qu'une ontologie ? », « à quoi ça sert ? »...Sachant qu'à chaque fin d'année, on faisait « une grande messe » où on réunissait tous les partenaires des différents lots. On reprenait toujours une présentation avec « où en est l'ontologie aujourd'hui ? », « qu'est-ce que ça apporte dans le prototype aujourd'hui ? ». Donc je pense qu'on a quand même réussi car le projet continue aujourd'hui !

Oui...je voulais revenir sur les méthodologies de construction dites « officielles » : est-ce que vous en aviez entendu parler avant ?

Oui, tout à fait, je les avais utilisées dans le cadre de ma thèse. Effectivement, je connaissais bien les méthodos, plus pour construire des ressources termino-ontologiques, ce qui était tout à fait notre cas puisque notre ressource permettait d'avoir à la fois un côté ontologique pur, conceptuel et une ressource plus terminologique pour alimenter les outils de TAL. Donc après je ne suis pas sûre qu'on l'a suivie à la lettre, mais après grosso modo, ces étapes-là de production d'ontologies sont assez classiques, il n'y pas 36 manières non plus de construire une méthodologie. Ce sont des choix de modélisation par contre qui peuvent être faits : telle relation entre entités, est-ce que justement je la modélise par une relation ou est-ce que ça va plutôt être un attribut avec une

data-propertie ? Ça c'était plutôt une question de choix de modélisation et c'était vraiment guidé par nos usages, selon le raisonnement qu'on va mettre en place après.

Et quand vous dites « on n'est pas sûrs d'avoir suivi cette méthodologie à la lettre », est-ce que vous avez des exemples ?

Alors je n'ai pas d'exemple comme ça, ça fait un peu longtemps ! (rire). Voilà, il y a de grandes étapes, on va d'abord s'intéresser au vocabulaire, structurer hiérarchiquement ce vocabulaire, se demander s'il y a des concepts qui émergent et comment je peux les représenter...Pour tout vous dire, je ne me fie pas vraiment à la méthodo comme elle a été formalisée, je ne vais pas me dire « à quelle étape je suis de cette méthodo là ? ». On est plus dans le pragmatisme. Il faut que ce soit pratique, il y a des aspects plus théoriques que je vais laisser aux chercheurs. Moi ce qui m'intéresse, c'est mon application, comment je peux la faire fonctionner, comment ça peut répondre aux exigences...Après, je construis ça de manière itérative. Donc on ne peut pas dire qu'on a suivi la méthodologie à la lettre ! (rire)...Disons que je pense que ça guide beaucoup quand on démarre la modélisation d'ontologies, quand on n'en a jamais trop fait avant, ça permet d'avoir un guide des principes de modélisation qui sont effectivement très bien à avoir en tête. Et après, plus on a de l'expérience, et moins on a besoin de regarder les guides. C'est vrai qu'aujourd'hui, je ne regarde plus trop les guides...En fait chacun dérive sa propre méthodo de sa propre expérience, selon le contexte aussi. Il y a la théorie et la pratique ! [...]

Voilà, je pense que j'ai fini les questions de mon côté. Est-ce que vous avez autre chose à ajouter ?

...Oui c'était vraiment compliqué d'avancer sur la méthodologie avec la gestion du projet qu'il y avait. Ce n'était pas la gestion de l'ontologie en elle-même qui était compliquée, elle était riche, exigeante, elle n'était pas simple vu l'ampleur du projet...Mais ce n'est pas ça la partie la plus lourde. C'était vraiment la gestion du projet et la dépendance à d'autres partenaires, notamment l'UX qui était vraiment très contraignant pour le développement de l'ontologie. [...]

Du coup, au départ j'avais une vision assez naïve de la question, je me disais qu'il y avait les chercheurs d'un côté et les professionnels de l'autre. Mais c'est vrai que vous avez fait une thèse et là, vous étiez vraiment dans un contexte industriel où il fallait prendre en compte la pratique, je vois que tout ça est assez mouvant...

Oui, je suis un peu à cheval entre les deux mondes. C'est ça, souvent, dans le domaine du Web sémantique, un des problèmes aujourd'hui, c'est que ça reste beaucoup un sujet très théorique avec des chercheurs...ça a du mal à sortir sur des cadres applicatifs, après il y en a. Mondeca a maintenant presque vingt ans d'expérience et ils ne font pas que de la recherche, il y a deux tiers de projets clients. Il y a pas mal de gens qui s'y intéressent et qui font des applications concrètes. Mais c'est vrai qu'avec les ontologies, il faut y passer du temps, il faut convaincre, il faut évangéliser, ce n'est pas quelque chose d'aussi facile que sur d'autres technologies. Donc être à la croisée des deux mondes, c'est intéressant car on peut faire le pont entre les avancées théoriques et comment on les met en pratique pour répondre à de vrais besoins qui viennent des utilisateurs. Là pour SEB, on a vraiment été dans l'objectif d'une application industrielle derrière. [...] Voilà, il y a vraiment un point qui bloque aujourd'hui et qui gêne peut-être l'acceptation des ontologies, les chercheurs ont leur approche chercheur très théorique, mais quand on fait une application, des fois on ne peut pas être super exhaustifs ou super nickel sur tous les angles de la connaissance qu'on souhaite modéliser. À un moment, il faut arrondir les angles. Et oui, des fois les notions de vocabulaire ne sont pas les mêmes non plus, il faut ajuster. Quelquefois à Mondeca sur des projets, on va dépasser le cadre théorique pour vraiment faire quelque chose de très pratique. Et du coup, ça ne va pas plaire aux chercheurs parce qu'on n'a pas suivi les cadres théoriques formels. Il y a aussi un autre aspect sur le raisonnement qui est très important : souvent les chercheurs des laboratoires font des ontologies qui impliquent une certaine logique formelle dans un monde ouvert. Si je ne dis pas que quelque chose n'est pas possible, peut-être que ça l'est. Alors que nous, dans un cadre applicatif, on aime bien le monde fermé, c'est l'inverse, je peux faire des raisonnements plus facilement, interpréter et surtout conclure plus facilement. Mais voilà, ce n'est pas de la logique formelle...Ce sont deux mondes, monde ouvert et monde fermé, qui s'opposent un peu. Mais c'est en passe d'être réconcilié avec la nouvelle norme du W3C qui s'appelle SHACL (*Shapes Constraint Language*) qui permet de rajouter des contraintes à OWL et donc pourrait transformer le monde ouvert de OWL en monde fermé. Ça permettrait de faire le lien entre les applications et les ontologies formelles. Mais voilà, ça reste compliqué parce que les chercheurs veulent un monde parfait, bien modélisé, et les industriels veulent juste quelque chose qui marche.

Table des matières

Résumé et descripteurs

Résumé

La recherche d'information classique est freinée par de nombreuses limites. En effet, le langage est complexe et de nombreuses ambiguïtés dégradent la pertinence des résultats d'une requête. Les ontologies peuvent améliorer la recherche d'information. Modélisant les concepts d'un domaine et les relations les liant entre eux, elles fournissent un vocabulaire commun et formel. La construction de ces ressources est longue et complexe ; ainsi, des articles scientifiques présentant des méthodologies de conception d'ontologies abondent. Nous nous demandons si ces méthodes sont réellement utilisées dans des contextes industriels ou si elles restent dans la sphère de la recherche. Cela nous amène à questionner le lien entre les chercheurs et les personnes concevant des ontologies dans un cadre davantage professionnel. Des entretiens auprès de six concepteurs d'ontologies de divers domaines nous éclairent sur cette problématique. À l'issue de cette enquête, plusieurs points apparaissent clairement. Les méthodologies officielles sont peu utilisées, laissant place à des méthodes plus libres et itératives ; la prise en compte des usagers finaux de l'ontologie est cruciale ; les objectifs et pratiques des chercheurs et professionnels différent parfois, même si cette frontière entre professions tend à s'estomper.

The classic information retrieval has many limitations. Indeed, the language is complex and many ambiguities degrade the relevance of the results of a query. Ontologies are able to improve information retrieval. By modelling the concepts of a domain and the relationships between them, they provide a common and formal vocabulary. The construction of these resources is long and complex; thus, scientific papers presenting ontology design methodologies abound. We wonder whether these methods are really used in industrial contexts or whether they remain in the sphere of research. This leads us to question the link between researchers and people designing ontologies in a more professional setting. Interviews with six ontology designers from various fields shed light on this problem. At the end of this investigation, we were able to identify several important points. Official methodologies are little used, leaving room for freer and more iterative methods; taking into account the final users of ontology is crucial; the objectives and practices of researchers and professionals sometimes differ, even if this boundary between professions tends to blur.

Descripteurs

construction d'ontologie - recherche d'information - Web sémantique

- Système d'Organisation des Connaissances

ontology construction - information retrieval - Knowledge Organization System - semantic Web