Le cycle de vie d'une ontologie : évaluation de l'ontologie du domaine de la Toxicologie Nucléaire

Anne-Claire Le Picard

▶ **To cite this version:**

Anne-Claire Le Picard. Le cycle de vie d'une ontologie : évaluation de l'ontologie du domaine de la Toxicologie Nucléaire. domain_shs.info.docu. 2014. mem_01128938

HAL Id: mem_01128938

https://memsic.ccsd.cnrs.fr/mem_01128938

Submitted on 10 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

Ecole Management et Société-Département CITS

INTD

MEMOIRE pour obtenir le

Titre professionnel "Chef de projet en ingénierie documentaire" INTD

RNCP niveau I

Présenté et soutenu par

Anne-Claire Le Picard

le 26 novembre 2014

Le cycle de vie d'une ontologie :

évaluation de l'ontologie du domaine de la Toxicologie Nucléaire

Jury :
Ghislaine CHARTRON. Professeure titulaire de la chaire d'ingénierie documentaire du CNAM, directrice de l'Institut national des sciences et techniques de la documentation (co-directrice)
Marie-Thérèse MÉNAGER. Coordination de programmes Toxicologie, Commissariat à l'énergie atomique et aux énergies alternatives (co-encadrante)
Jean CHARLET. Chargé de mission recherche à l'Assistance Publique – Hôpitaux de Paris, Institut national de la santé et de la recherche médicale (co-directeur et co-encadrant)

Promotion 44

Remerciements

J'adresse mes remerciements à toutes les personnes qui m'ont aidée dans la réalisation de ce mémoire.

Je tiens tout d'abord à remercier Jean Charlet pour le suivi et l'intérêt porté à mon travail ainsi que pour ses encouragements. Je remercie également mes co-encadrants du LGI2P de Nîmes, Sylvie Ranwez et Gérard Dray, qui malgré la distance se sont montrés très disponibles pour répondre à mes nombreuses questions. Je remercie Marie-Thérèse Ménager, pour son enthousiasme et la rencontre permise avec Magali Le Discorde (BioDoc, CEA). Qu'elle soit elle aussi remerciée pour avoir bien voulu répondre à mes sollicitations. Je remercie Ghislaine Chartron, directrice de l'INTD, pour avoir accepté en dernière minute de diriger ce mémoire ainsi que pour ses orientations. Je souhaite également remercier Marie-Christine Jaulent et l'équipe du Limics, pour l'accueil chaleureux qu'ils m'ont accordé durant la mission de stage et pour m'avoir permis d'assister aux séminaires du laboratoire.

Je remercie les chercheurs et professionnels qui m'ont reçue. Mes remerciements vont à Loïc Lebigre de l'ADBS et de l'INTD. Du Limics, je remercie Xavier Aimé pour m'avoir exposé ses travaux. De l'université Paris Descartes, je remercie Éric Dagiral pour avoir partagé son regard de sociologue des techniques et de l'innovation. Merci à Diane Le Hénaff et Sophie Aubin de l'INRA qui ont partagé avec moi leur expérience et leurs avis concernant la mise en place des référentiels. Je remercie les consultants et formateurs Jean Delahousse et Thomas Francart à qui je suis très reconnaissante d'avoir encore facilité l'utilisation de l'outil *SKOS Play!.* Ma réflexion a grandement bénéficié des échanges et des conseils avisés de Sylvie Dalbin, que je remercie pour sa réactivité et pour son exigence stimulante.

Je profite également de cette page qui m'est laissée pour témoigner à nouveau de la satisfaction apportée par la belle humeur et la solidarité de la promotion 44 du titre 1, sans oublier l'équipe pédagogique et administrative de l'INTD.

J'ai une pensée pour mes anciens professeurs, mes anciens et présents collègues. Je pense à ceux, qui parmi eux, m'ont encouragée et autorisée à suivre cette année de formation à l'INTD.

Enfin, je remercie chaleureusement mes proches pour leur affection et pour leur soutien.

Notice

LE PICARD A.-C. Le cycle de vie d'une ontologie : évaluation de l'ontologie du domaine de la Toxicologie Nucléaire. Mémoire professionnel INTD, Titre I, Chef de projet en ingénierie documentaire. Conservatoire national des arts et métiers - Institut national des Sciences et Techniques de la Documentation, 2014, 172 p. Promotion 44

Résumé : Ce mémoire s'intéresse tout d'abord à l'émergence des ontologies aux côtés des autres systèmes d'organisation des connaissances (SOCs) avec lesquels elles sont mises en perspective : taxinomies, classifications, thésaurus. Les ontologies sous-tendent de nombreuses applications pour l'exploitation de l'information et sont au cœur des modèles du Web sémantique. L'analyse porte ensuite sur le contexte de création d'une ontologie de son domaine de recherche par le collectif français dédié à la toxicologie nucléaire. En plus des aspects méthodologiques, son étude permet aussi d'examiner les aspects organisationnels. Enfin, les méthodologies développées pour corriger l'ontologie et pour répertorier d'autres modèles concurrents ou complémentaires sont exposées. D'après ces observations, des orientations sont proposées pour cette ontologie ainsi que des préconisations quant à l'exploitation, le maintien et la valorisation d'un tel outil.

Descripteurs : Écotoxicologie ; France ; Gestion des connaissances ; Interopérabilité ; Langage documentaire ; Norme ; Ontologie ; Standard ouvert ; SKOS ; Toxicologie nucléaire ; Web sémantique ; Web de données ; Web 3.0

Abstract : In the first place, this essay deals with the emergence of KOSs (knowledge organisation systems) such as ontologies as compared with taxonomies, classifications and thesaurii, and it discusses the respective specificities. Many information retrieval systems run on ontologies which are particularly concerned with the semantic Web. Next, the analysis focuses on the creation context of domain ontology of the research field of the French nuclear toxicology community. Furthermore, this case study outlines both methodological and organisational aspects. And this essay eventually proceeds to methodologies developed in the course of this research in order to fix ontology bugs and to identify other competing or interesting resources to complete this ontology. Based upon these observations, three application examples are presented of how this ontology can serve the nuclear toxicology community, leading to perspectives on how to maintain it and the necessary measures to promote it.

Keywords : Controlled vocabulary ; Ecotoxicology ; France ; Knowledge management ; Interoperability ; Nuclear toxicolology ; Ontology ; Standard ; Semantic web ; SKOS ; Web of data ; Web 3.0.

Table des matières

Liste des figures

Liste des sigles et acronymes utilisés

Abes	Agence bibliographique de l'enseignement supérieur
ADBS	Association des professionnels de l'information et de la documentation
AIEA	*Atomic Energy Agency*
AIEA	*International Atomic Energy Agency*
ANR	Agence nationale de la recherche
BBC	*British Broadcasting Corporation*
BDID	Banque de données bibliographiques de la littérature en information documentation
BDSP	Banque de données en santé publique
BnF	Bibliothèque nationale de France
CC	*Colon Classification*
CEA	Commissariat à l'énergie atomique et aux énergies alternatives
Cern	Organisation européenne pour la recherche nucléaire, aussi appelée laboratoire européen pour la physique des particules. Il est couramment désigné sous l'acronyme de l'organe provisoire institué en 1952 : Conseil européen pour la recherche nucléaire.
CISMeF	Catalogage et indexation des sites médicaux de langue française
CMS	*Content Management System*
Dafoe	*Differential and formal ontologies editor for applications*
DC	*Dublin Core*
DSV	Direction des sciences du vivant (CEA)
EHESP	École des hautes études en santé publique (ex. ENSP)
EMA	École nationale supérieure des Mines d'Alès
EPA	*Environmental Protection Agency*

ERMS	*Electronic Ressources Management Systems*
ERP	*Enterprise Resource Planning*
ETDE	*Energy Technology Data Exchange*
GED	Gestion électronique de documents
GO	*Gene Ontology*
HeTOP	*Health Terminology / Ontology Portal*
I&D	Information et documentation
IA	Intelligence artificielle
IC	Ingénierie des connaissances
Inis	*International Nuclear Information System*
Inist	Institut national de l'information scientifique et technique
Inserm	Institut national de la santé et de la recherche médicale
IST	Information scientifique et technique
KOS	*Knowledge organisation system*
LCSH	*Library of Congress Subject Headings*
LED	*Linked Enterprise Data*
LGB	Logiciel de gestion de bibliographies
LGI2P	Laboratoire de génie informatique et d'ingénierie de production
LGRB	Logiciel de gestion de références bibliographiques
LIMICS	Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé
LISTA	*Library, Information Science & Technology Abstracts*
LOD	*Linked Open Data*
LOV	*Linked Open Vocabularies*

MCC	Ministère de Culture et de la Communication
MCF	Maître de conférences
MeSH	*Medical Subject Headings*
NAF	Nomenclature d'activités française
NCBO	*National Center for Biomedical Ontology*
NIH	*National Institute of Health*
NLP	*Natural Language Processing*
OWL	*Ontology Web Language*
OWL DL	*Ontology Web Language Description Logics*
PDF	Portable document format
PIST	Professionnel de l'information scientifique et technique
Rameau	Répertoire d'autorité matière encyclopédique et alphabétique unifié
RSE	Réseau social d'entreprise
RSN	Réseaux sociaux numériques
RSS	*Really Simple Syndication* ou *Rich Site Summary*
RTO	Ressources terminologiques ou ontologiques ; ressources termino-ontologiques
SBC	Systèmes à base de connaissances
SEO	*Search Engine Optimization*
SI	Système d'information
SIB	Sciences de l'information et des bibliothèques
Siaf	Service interministériel des archives de France
SIGB	Système intégré de gestion de bibliothèques
SKOS	*Simple Knowledge Organisation System*
SOC	Système d'organisation des connaissances

STIC	Sciences et technologies de l'information et de la communication
TAL	Traitement automatique des langues
TALN	Traitement automatique du langage naturel
TGIR	Très grande infrastructure de recherche
TSP	Thésaurus santé publique
UMR	Unité mixte de recherche
Urfist	Unité régionale de formation à l'information scientifique et technologique
URI	*Uniform Resource Identifier*
W3C	*World Wide Web Consortium*

Remarque préliminaire

Tous les liens d'accès à des ressources sur Internet étaient valides à la date du 16 novembre 2014.

Introduction

Présentation générale

Les ontologies sont des artefacts informatiques qui reposent sur la sémantique. C'est cette particularité qui fait leur puissance par rapport aux méthodes statistiques et linguistiques [55, DALBIN]. Elles complètent alors ces dernières dans le cadre de la recherche d'information dans des dispositifs numériques de type bases de connaissances, portails et autres plateformes collaboratives.

Les ontologies, telles qu'elles sont conçues majoritairement aujourd'hui reposent sur des standards et des formats ouverts, autrement appelés standards du Web sémantique [59, TRONCY] Ils rendent possible le Web de données ou *Linked Open Data* (LOD). Les ontologies sont alors les garantes de l'interopérabilité et de la pérennité de l'accès efficace à la connaissance ainsi formalisée. Ce sont là deux des fonctions auxquelles elles permettent de répondre. Pour une organisation, s'engager dans cette voie assure le décloisonnement de ses différentes bases de données au sein de son système d'information (SI). L'application de ces technologies Web utilisées seulement en interne pour rendre interopérables les données de l'organisation est appelée *Linked Enterprise Data* (LED). Depuis bien avant l'apparition de ces standards, les sociétés du secteur industriel s'emploient à rendre leurs différentes applications métiers interopérables entre elles par le recours aux ontologies. En effet, dans un souci d'efficacité, les grandes sociétés comme Thales par exemple s'efforcent d'harmoniser le traitement des données et des vocabulaires métiers. Ces derniers peuvent devenir un modèle pivot pour que des applications ne partageant pas les mêmes modèles de référence puissent communiquer entre elles [67, SCIANDRA ; 64 BARBAUX]. Par ailleurs, à l'échelle d'une administration comme celle du Ministère de Culture et de la Communication (MCC), ce sont les différentes bases de la BnF, de la Cité de la musique, des musées, du Service interministériel des archives de France (Siaf), de la Direction générale de la création artistique, de la Direction générale des patrimoines… qui progressivement vont s'interconnecter, s'enrichissant mutuellement et éviter ainsi la redondance en utilisant des modèles de référence communs[1].

Pour une organisation, quelle qu'elle soit, formaliser un référentiel selon les standards indique alors une volonté d'innovation, mais également une stratégie d'ouverture visant une plus grande visibilité. En effet, en cas de publication des données produites selon ces standards dans le cadre de l'activité de cette organisation, la réutilisation par d'autres acteurs en est favorisée. Pour une institution, dont les données sont publiques, c'est un impératif renforcé par la mouvance de l'*Open Data*[2]*.* Cela participe notamment à renforcer sa présence, à asseoir sa notoriété et son autorité dans son domaine. Des organismes, autres que ceux relevant de l'administration, de la culture, de

[1] La feuille de route stratégique Métadonnées culturelles et transition Web 3.0 a été présentée par B. Sajus lors du séminaire Web de données de l'INTD et est consultable en ligne à cette adresse : < http://fr.slideshare.net/culturefr/feuille-de-route-mtadonnes-et-30-camille-domange >

[2] Loi n° 1978-753 du 17/7/1978 dite loi CADA (Commission d'Accès aux Données Administratives) de 1978 relative à la liberté d'accès aux documents administratifs et à la réutilisation des informations publiques
http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068643&dateTexte=vig

l'enseignement et de la recherche publique, sont néanmoins concernés. Eux aussi développent ou utilisent des ontologies afin de bénéficier des données mises à disposition par les institutions[3]. Certains acteurs peuvent aussi veiller à l'interopérabilité de leur système avec celui des institutions avec lesquelles ils contractent. C'est le cas par exemple des éditeurs et des diffuseurs avec lesquels négocient les agences bibliographiques comme l'Abes et la BnF [58, ILLIEN]. Pour reprendre l'exemple du MCC, le rapport Lescure, paru en mai 2013, considère comme un enjeu national l'interconnexion entre les bases de données publiques et privées[4]. Par ailleurs, certains vendent des jeux de données issues de sources diverses, partiellement non structurées, auxquelles ils appliquent des traitements pour les rendre interopérables avec les systèmes de leurs clients[5]. Dans une autre mesure, ils peuvent mettre à disposition leurs modèles de données à défaut des données elles-mêmes. Et ceci, en documentant et en accompagnant les modèles d'une licence d'utilisation ouverte afin qu'ils soient réutilisés, enrichis ou adaptés à d'autres besoins[6]. Puisque ces données sont pour partie libres et gratuites, ce sont éventuellement des raisons économiques qui encouragent à adopter ces modèles pour les récupérer. Pour une organisation engagée dans un projet soutenu par de tels modèles, cela offre partiellement une indépendance vis-à-vis de solutions propriétaires [56, DELAHOUSSE].

Cependant, les ontologies sont des artefacts complexes. De plus, les méthodes et les outils visant à leur constitution sont encore à l'état de développement et font débat [25, 62, ZACKLAD]. Par ailleurs d'autres technologies bien que complémentaires pour améliorer l'accès à la connaissance, sinon à l'information, comme celles dites du *Big Data* ou mégadonnées selon la recommandation[7], leur font concurrence et pourraient ralentir l'exploration de ce champ. Conçues non pour elles-mêmes, mais pour appareiller des dispositifs numériques, c'est leur puissance technologique et leurs promesses qui sont principalement mises en avant. Que cela soit d'ailleurs dans la littérature académique ou celle professionnelle de l'information documentation, le professionnel peut donc être impressionné par cet aspect technique. Il y est en effet largement question de modèles conceptuels et de nouveaux langages informatiques à appréhender. Et cela peut parfois éclipser la question des besoins des différents utilisateurs et aussi celle de la finalité des outils. Ce sont pourtant, comme pour tout projet, les éléments importants à garder à l'esprit tout au long d'un processus de conception. A ce risque qui menace l'implémentation effective d'une ontologie dans un dispositif

---

[3] Le collectif associatif Regards Citoyens : http://www.regardscitoyens.org

[4] FRANCE. PRÉSIDENCE DE LA RÉPUBLIQUE; FRANCE. MINISTÈRE DE LA CULTURE ET DE LA COMMUNICATION. *Mission « Acte II de l'exception culturelle » : contribution aux politiques culturelles à l'ère numérique* [En ligne]. [s.l.] : [s.n.], [s.d.]. Disponible sur : <http://www.ladocumentationfrancaise.fr/rapports-publics/134000278

[5] Société Data Publica : http://www.data-publica.com

[6] *BioPortal* offre aux auteurs d'ontologies biomédicales de les référencer et de les documenter. http://bioportal.bioontology.org

[7] La délégation générale à la langue française et aux langues de France (DGLFLF) recommande d'employer cette expression à la place de Big data http://www.culture.fr/layout/set/print/franceterme/terme/COGE874

numérique et qui menace aussi la satisfaction des utilisateurs, s'ajoutent d'autres risques. Il s'agit de ceux liés à la dimension organisationnelle. Elle aussi est inhérente à tout projet, mais elle est particulièrement prégnante dans un projet lié à la gestion des connaissances. Celui-ci induit forcément la collaboration d'un nombre important d'individus aux compétences variées. Et que cela soit les divers individus que le dispositif doit servir tout aussi bien que les informaticiens aux profils différents et diversement amenés à collaborer pour développer l'ontologie. Ces individus sont, pour chacun, spécialistes de domaines liés mais par ailleurs divergents quant aux méthodes et aux outils employés. Par conséquent, représenter la connaissance qu'ils ont en partage n'est pas chose aisée. Aux côtés de ces spécialistes d'une discipline, les informaticiens venus de l'IA, les informaticiens linguistes et les ergonomes quand ces derniers sont sollicités, utilisent différentes méthodes, elles-mêmes encore en cours d'élaboration. Ce qui se traduit par des projets autrement plus complexes. Rendre encore plus opérationnelle la connaissance pour tous et pour chacun s'avère alors aussi stimulant qu'éprouvant, quand bien même chacun en perçoit les bénéfices. Des investissements aussi ambitieux intellectuellement et techniquement nécessitent alors d'envisager de se doter de moyens de pilotage pour garantir non seulement l'aboutissement des projets mais aussi la phase d'après projet.

Contexte de la mission de stage

C'est dans ce contexte que la direction du programme Transversal Toxicologie Nucléaire piloté par le Commissariat à l'énergie atomique et aux énergies alternatives (CEA) et ses partenaires du Laboratoire de Génie Informatique et d'Ingénierie de Production (LGI2P) de l'École Nationale Supérieure des Mines d'Alès (EMA) ont entamé en 2006 la construction d'une ontologie du domaine de la toxicologie nucléaire. En amont et en parallèle de cette ontologie nommée ToxNuc, d'autres dispositif numériques ont été développés.

Afin d'envisager comment poursuivre la construction de cette ontologie, une réflexion à son propos a fait l'objet d'une mission de stage. L'objectif de cette mission était de réaliser d'une part un travail de correction de l'ontologie informatique ToxNuc et d'autre part de confronter cette ontologie aux autres ressources de ce type apparues depuis 2006.

Problématique et questionnement

Ce mémoire s'inscrit dans la prolongation de l'étude de ce cas concret de construction d'une ontologie de domaine. Lors de ce stage, la réflexion a plus porté sur la dimension organisationnelle que sur la dimension technique. C'est cette réflexion qui est poursuivie dans le présent mémoire en proposant de répondre à la problématique suivante :

« dans quelle mesure les ontologies qui apportent fluidité et pertinence dans les systèmes d'information peuvent-elles être conçues, maintenues et pérennisées ? »

Une hypothèse est également formulée. Celle-ci est relative aux risques liés à toute gestion de projet. Il s'agit de démontrer que si la prise en compte de la dimension technique est nécessaire, la dimension organisationnelle doit l'être aussi largement[8]. Et ceci tout autant pour l'aboutissement des projets de conception que pour la pérennité des ontologies elles-mêmes après leur mise en production.

Méthodologie

Les réflexions présentées dans ce mémoire reposent sur une synthèse de lectures portant sur les ontologies informatiques. Elles se nourrissent également d'entretiens avec les différents contributeurs de l'ontologie ToxNuc pour en retracer l'historique et son contexte. Les informations recueillies sont confrontées aux recommandations et exemples fournis par la bibliographie constituée. De même, la rencontre avec divers professionnels a permis de comparer ce projet à d'autres (*benchmark*). Ces entretiens participent à mettre en perspective la mission de correction de l'ontologie ToxNuc et d'élargir l'horizon de ces réflexions.

Structure du mémoire

Ce mémoire s'articule en trois parties. Tout d'abord, c'est un travail de définition qui est réalisé. Il vise à mieux comprendre ce que recèle le terme « ontologie » qui apparaît de plus en plus dans le milieu des professionnels de documentation et des bibliothèques. A cette occasion, il est éclairant de revenir sur les caractéristiques et les usages des outils documentaires plus familiers de ces professionnels et auxquels les ontologies sont souvent comparées. Au-delà d'éclairer la vision de ces professionnels, quelle est la place aujourd'hui des ontologies au sein du Web et plus particulièrement du Web sémantique ? Répondre à cette question amène aussi à s'intéresser à d'autres acteurs étant aussi partie prenante avec les ontologies.

Dans un second temps, le mémoire est consacré à l'étude du projet d'ontologie de la toxicologie nucléaire ToxNuc. Cette étude s'attache à détailler les différentes étapes techniques menées jusqu'à la réalisation du stage et à en faire un premier bilan technique. Au travers de cet historique, un soin est apporté à décrire le contexte de ce projet. Le but étant d'analyser en quoi les différents éléments de celui-ci ainsi qu'une expérience récente ont fait évoluer les enjeux autour de cette ontologie.

Enfin, la démarche et les méthodes employées pour mener à bien la mission sont présentées dans la dernière partie. L'objectif de celle-ci est aussi de proposer différentes orientations pour l'ontologie ToxNuc. Pour chacune de ces orientations, il s'agit d'en exposer les implications principales afin d'alimenter la réflexion sur les suites à donner à cette ontologie. En effet, l'orientation de celle-ci dépend du besoin réel et des moyens disponibles. Des propositions sont alors dégagées pour la maintenance et la valorisation de ce type de ressources.

[8] DAGIRAL É., PEERBAYE A. « Les mains dans les bases de données : connaître et faire reconnaître le travail invisible ». *Revue d'anthropologie des connaissances*. 2012. Vol. 6, 1, n°1, p. 229.

Première partie

Les ontologies : une tentative de définition et de mise en perspective

1 Un sujet d'actualité en sciences de l'information

Depuis quelques années, les revues à destination des chercheurs en sciences de l'information documentation abordent de plus en plus les ontologies. Il en est de même pour la revue professionnelle *Documentaliste - Sciences de l'information* éditée par la première association de professionnels de l'information et de la documentation en Europe, l'ADBS. Ce constat s'appuie sur l'étude bibliométrique réalisée pour ce mémoire à partir de deux bases spécialisées dans les domaines de l'information documentation et des bibliothèques comme *Library, Information Science & Technology Abstracts* (LISTA) et la Banque de données bibliographiques de la littérature en information documentation (BDID)[9]. Depuis 2006, l'INTD intègre dans les enseignements de ses différents cursus des heures spécifiquement dédiées à sensibiliser les professionnels aux ontologies. Auparavant les intitulés de ce type d'enseignements étaient composés des termes « structure » ou « schéma ». Par ailleurs, depuis 2008, l'ADBS propose deux stages de formation par an dans lesquels il est question d'ontologie[10]. Par ailleurs l'Urfist de Rennes a proposé une formation dédiée à l'indexation du document numérique et aux ontologies dès 2005. L'Urfist de Paris organise des formations qui abordent les ontologies depuis 2008.

1.1 A l'origine de la résurgence du terme « ontologie »

L'Ontologie est d'abord une notion philosophique. Du grec onto- ὤν, ὄντος « étant », participe présent du verbe εἰμί « être », et de ogos, λὸγος lógos « parole, discours », c'est la théorie de l'être ou science de l'être. Elle est considérée au XVII[e] siècle comme synonyme de métaphysique ou comme en étant une partie[11].

Mais selon les communautés qui l'emploient, ce terme prend un sens différent. Ainsi pour les historiens de la médecine, il s'agit de la doctrine qui prétend étudier l'être de la maladie et à laquelle s'est opposée la doctrine physiologique au XIX[e] siècle.

---

[9] L'interrogation de ces deux bases est soumise à abonnement. La première est un produit EBSCO (plus de 600 revues indexées, mais aussi de la littérature grise et des ouvrages depuis 1960). < http://www.ebscohost.com/academic/library-information-science-technology-abstracts-lista >
La seconde contient 33 793 références et est éditée depuis 1985 par le Centre de ressources documentaires du Cnam-INTD (Conservatoire National des Arts et Métiers-Institut National des Techniques de la Documentation). < http://bdid-intd.cnam.fr >

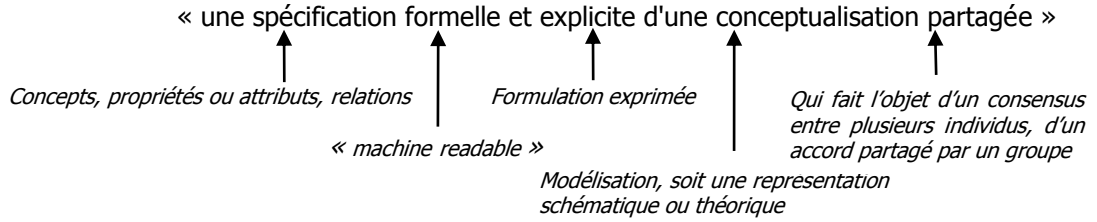
[10] Stage ADBS : « S'initier aux ontologies » et « Web sémantique : publication de données » renommées respectivement dans le catalogue de formation 2015 : « Traiter les data grâce aux ontologies » et « Comprendre le web sémantique »
Il est à noter par ailleurs que le vocabulaire de la documentation interrogeable en ligne depuis son site ne comporte pas d'entrée pour le terme « ontologie ». Il s'agit d'un vocabulaire élaboré en 1999.

[11] Encyclopaedia universalis : version 6 CD-ROM PC Windows. Paris : Encyclopaedia universalis, 2000. ISBN : 2-85229-296-3.
LITTRÉ É., BAUDENEAU J., MORHANGE-BÉGUÉ C. Dictionnaire de la langue française. Édition nouvelle. Versailles : Encyclopaedia Britannica France, 1999.

Cependant, la principale distinction est celle entre le sens philosophique et celui donné en informatique. L'intelligence artificielle (IA), la branche de l'informatique en liaison avec les sciences cognitives à l'origine des systèmes experts apparus vers le milieu des années 70, s'est progressivement élargie en s'intéressant aux outils et méthodes pour la conception de systèmes à base de connaissances (SBC). Ainsi s'est constituée la communauté internationale d'acquisition des connaissances, maintenant appelée ingénierie des connaissances (IC). Elle a emprunté à la philosophie le terme d'ontologie au début des années 90. Il ne s'agit donc plus ici de science de l'être, à laquelle seule on attribuera une majuscule et le singulier, l'Ontologie, mais d'objets informatiques [41, GUARINO ; 27, BACHIMONT]. Les ontologies informatiques, avec un « o » minuscule et la forme plurielle ont deux rôles. D'une part, elles sont conçues pour définir et fournir à des humains une sémantique interprétative d'un domaine du monde réel fondé sur un consensus, soit un langage commun. D'autre part, elles doivent définir et fournir une sémantique formelle pour l'information de ce domaine afin que celle-ci soit exploitable par un ordinateur. Ce sont donc des objets qui peuvent s'intégrer dans diverses applications informatiques.

1.2 Mais que faut-il entendre par ontologie ?

La définition la plus consensuelle dans la littérature de l'IC est celle de [31, STUDER et al.] qui fusionne la définition originelle de [32, GRÜBER] de 1993 et celle donnée par [29, BORST] en 1997. Une ontologie informatique est alors définie comme *« une spécification formelle et explicite d'une conceptualisation partagée ».* Cela signifie que c'est un objet qui est le produit d'un processus, soit d'une succession d'actions. Au cours de ce processus, il s'agit de fixer et valider, non par un individu seul mais par un groupe, les définitions des concepts (entités, attributs, processus) ainsi que leurs interrelations. Et ceci de façon formelle afin d'être exploitable par la machine. Le langage formel garanti l'exécution d'opérations informatiques pour la communication et l'échange de données entre applications.



Surtout, les auteurs de l'IC insistent sur le fait qu'une ontologie n'est pas produite pour elle-même. Elle est toujours un moyen et non une fin. Sa formalisation est fonction des usages et des applications comme celles liées :

- à la gestion des connaissances ;
- à la recherche d'information (sémantique) ;
- à l'aide à la navigation ;
- à l'aide à la décision.

1.3 Le retour des langages contrôlés ?

C'est en partie pour ces cas d'usages cités ci-dessus, que les professionnels de l'information et de la documentation (I&D) s'intéressent aux ontologies. Effectivement, ces derniers sont au cœur de leurs préoccupations. L'autre raison tient à leur interprétation de ce que sont les ontologies informatiques. En effet, la grammaire des relations opérationnalisée dans un langage informatique est associée à ce que sont pour ces professionnels les vocabulaires contrôlés et les référentiels dans des domaines particuliers : vocabulaires conceptuels et schémas de métadonnées. Le « vocabulaire de la doc » interrogeable sur le site de l'ADBS renvoie vers la définition de l'expression « langage documentaire » :

> « langage contrôlé et normalisé utilisé dans un système documentaire pour l'indexation et la recherche. Un langage documentaire permet de représenter de manière univoque les notions identifiées dans les documents et dans les demandes des utilisateurs, en prescrivant une liste de termes ou d'indices, et leurs règles d'utilisation. On distingue les langages combinatoires ou postcoordonnés dont les éléments peuvent être combinés entre eux a posteriori lors de l'indexation ou de la recherche (thésaurus, classifications à facettes) ou les langages précoordonnés, contenant des combinaisons de notions établies a priori (classifications hiérarchiques, listes de vedettes-matières) ; les langages polyhiérarchiques (chaque terme du langage peut avoir plusieurs termes génériques de niveau immédiatement supérieur) ou monohiérarchiques. »
> http://www.adbs.fr/langage-documentaire-17593.htm?RH=OUTILS_VOC

D'après la norme [7, ISO 25964-1:2011], les vocabulaires contrôlés :

> « prescribed list of terms, headings or codes, each representing a concept […] Controlled vocabularies are designed for applications in which it is useful to identify each concept with one consistent label, for example when classifying documents, indexing them and/or searching them. Thesauri, subject heading schemes and name authority lists are examples of controlled vocabularies.»

Ce sont donc des outils de description et de communication entre personnes. Avec leur informatisation, cette communication se fait par l'intermédiaire de la machine. Mais la plupart des logiciels et des SIGB représentent très mal ces langages documentaires. La constitution et la gestion de ces outils sont une des spécialisations de ces gestionnaires d'information. Ainsi, les avancées de l'IC et l'intérêt porté au web sémantique pour améliorer l'interopérabilité entre les données présentes sur le Web, actualisent la réflexion sur les outils de l'I&D. Bruno Menon propose de rassembler les vocabulaires contrôlés et les ontologies sous le vocable générique de système d'organisation des connaissances (SOC)[12] [6, MENON]. Mais ce retour peut être aussi pervasif que transparent pour les utilisateurs et les décideurs. En effet, les ontologies et leurs composantes terminologiques ne sont qu'une des briques des systèmes d'information. Elles renforcent la dimension sémantique des

[12] D'après [6, MENON p. 10] citant Gail Hodge, ce terme est « proposé pour la première fois en 1998 lors du lancement du groupe de travail NKOS, (Networked Knowledge Organization Systems). » L'expression SOC ou KOS (Simple Knowledge Organisation System), pour laquelle l'expression « langage d'organisation des connaissances » pourrait tout aussi bien convenir selon M. Zacklad [25, ZACKLAD p. 1], regroupe « dans une dénomination unique aussi bien les langages documentaires, les schémas de classification que les langages de représentation des connaissances issus de l'Intelligence Artificielle »

applications mises en place dans les entreprises (CMS, ERP, GED, RSE, etc.). [57, GANDON *et al.* p. 11] parlent « *d'intrawebs sémantiques* »[13]. Sur Internet, elles sont présentes dans les sites des éditeurs de contenus commerciaux ou non. C'est le cas par exemple d'Amazon et des portails de bibliothèque, mais aussi des réseaux sociaux grand public, via leur intégration dans les moteurs de recommandations [51, KEMBELLEC *et al.*].

En ce qui concerne la recherche d'information, les SOCs ont été éclipsés un temps par la recherche sur le texte intégral. Effectivement, la disponibilité croissante de documents sous forme numérique et l'augmentation des capacités informatiques ont favorisé ce type de recherche. A cela, il faut ajouter l'unique cartouche de recherche du moteur de recherche Google, imité ensuite par les autres moteurs, ainsi que la puissance de son algorithme. Il est vrai aussi que la recherche sur les formes lexicales, autrement dit sur les chaînes de caractères, associée aux méthodes statistiques est plus économique. A l'encontre, celle basée sur les indexats de métadonnées recourt principalement jusqu'à présent à l'analyse intellectuelle des contenus en amont de cette recherche. Celle-ci est devenue inadaptée du fait de l'augmentation du nombre de documents, d'informations et de données à traiter. Ainsi la recherche sur le texte intégral permet d'économiser un temps de structuration de l'information et d'indexation exclusivement manuelle, fondée sur une analyse thématique intellectuelle des contenus. Cependant, cette recherche montre elle aussi ses limites concernant la qualité des résultats de recherche. Elle est confrontée à l'ambiguïté du langage, à l'hétérogénéité des sources d'informations croisées. A présent, la recherche s'oriente avec une visée sémantique vers une combinaison de ces différentes méthodes : statistiques, linguistiques et ontologiques. Pour autant, l'ensemble de ces technologies sont lourdes à mettre en place et nécessitent des avancées scientifiques pour améliorer l'efficacité de leur conception. Celles-ci risquent peut-être d'être ralenties par l'importance des investissements consacrés à la recherche aux mégadonnées bien que l'une et l'autre soient complémentaires.

[13] « un web formé par les serveurs HTTP d'un intranet » [57, GANDON *et al.* p. 11]

2 Comment concevoir des ontologies ?

2.1 Les différents types d'ontologies

Si les ontologies sont, comme les autres SOCs, des outils participant aux opérations permettant de classer, d'indexer, de représenter, de formaliser et de modéliser le réel, il se dessine aussi une classification des ontologies elles-mêmes. Les critères peuvent être ceux de leur niveau de complexité et de leur niveau de généricité. Et c'est d'ailleurs l'objet de débats au sein de la communauté de l'IC. Cependant, la typologie suivante est admise :

2.1.1 Les ontologies de haut niveau

Elles sont aussi appelées ontologies formelles ou *Upper Level Ontology* (ULO) ou encore *Top-ontologies*. Les concepts doivent être les plus abstraits et génériques possibles afin d'être universels, et être indépendants d'un problème à résoudre ou d'un domaine [41, GUARINO p. 9 et 11]. C'est à ces concepts d'ordre très général, « *space, time, matter, object, event, action* » que doivent pouvoir s'accrocher des ontologies plus spécifiques présentées ci-dessous. Les concepts de l'ontologie de haut niveau font penser aux facettes de la *Colon Classification* (CC) élaborée en 1933 par le bibliothécaire et mathématicien indien Shiyali Ramamrita Ranganathan. La spécificité de cette classification à facettes est de permettre de décrire et de chercher des ressources selon plusieurs angles de vue. Multidimensionnelle et combinatoire par opposition aux classifications monohiérarchiques, elle s'affranchit de l'espace où un objet ne peut être rangé qu'à un seul endroit en même temps [75, PIERRE].

Colon Classification **(CC)**	**ontologie de haut niveau**
Espace (localisation géographique)	L'espace
Temps (localisation chronologique ou temporelle)	Le temps
Matière (substance ou une propriété)	La matière
Personnalité (sujet ou concept principal)	Les objets
Energie (opération ou action subie par le sujet)	Les évènements
	Les actions

2.1.2 Les ontologies noyaux et les ontologies de domaine

Les ontologies noyaux sont aussi appelées *Core-ontologies* ou *Top-domain ontologies*. Il est parfois difficile de les distinguer des ontologies de domaine appelées *Domain ontologies* en anglais.

Dans une ontologie noyau, les concepts, mais aussi les relations entre ceux-ci sont propres à une discipline ou à la pratique d'une activité comme par exemple la médecine, le droit, la finance, le cinéma etc. Ces concepts prennent appui non pas sur les sujets du domaine, mais sur les faits, les processus et évènements propres à cette pratique. Pour l'ontologie noyau de la médecine, les « *concepts primitifs* » dits encore de premier niveau, c'est-à-dire ceux qui sont en haut de la hiérarchie sont : diagnostic, signe, structure anatomique. Les relations sont celles liées à la localisation d'une pathologie sur une structure anatomique. Dans une ontologie noyau, les concepts et les relations qui structurent le domaine lui sont propres. En principe, pour un domaine il n'y a qu'une seule ontologie noyau.

Quant aux ontologies de domaine, elles s'intéressent aux termes servant à désigner les concepts dans les documents produits par les professionnels du domaine lors de leur activité. Chez ces dernières, il y a une prédominance de la dimension linguistique et terminologique. Par exemple, l'ontologie de domaine peut être plus ou moins riche de termes synonymes pour dénoter les concepts. Par ailleurs, la spécification peut être plus ou moins raffinée selon l'application à laquelle l'ontologie est destinée. Ainsi, pour un domaine donné, il peut y avoir plusieurs ontologies de domaine.

L'ontologie ToxNuc observée pendant le stage a été conçue comme une ontologie de domaine. Pour une part, elle s'apparente à un certain type de vocabulaires que sont les terminologies telles que la norme [8, ISO 25964-2:2013] en rappelle les principales caractéristiques : « *ensemble des désignations appartenant à une langue de spécialité* ». En effet, il s'agit d'une liste où les termes sont égaux aux concepts et sont utilisés par les professionnels de la toxicologie. Mais les « concepts primitifs » dits encore de premier niveau, c'est-à-dire ceux qui sont en haut de la hiérarchie, s'apparentent aux faits et évènements propres à ce domaine. Malgré cela, elle ne peut être considérée comme une ontologie noyau car il n'y a pas de relations spécifiques qui déterminent les concepts entre eux. C'est donc bien une ontologie de domaine.

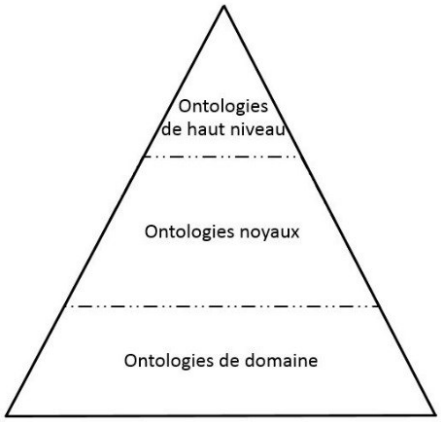


Figure 1 : les différents types d'ontologies, des frontières difficiles à distinguer.

2.2 Les caractéristiques des ontologies

Voici les éléments qui constituent une ontologie :

- Les concepts peuvent être appelés catégories ou bien encore classes ou instances lorsque l'ontologie est formalisée avec le langage OWL. Ils représentent des entités, autrement dit des objets matériels ou des notions ou des idées.
    - Les classes (*Class*) du premier niveau de l'ontologie sont appelées des concepts primitifs. Ils sont dénotés par des termes qui sont soit des mots mis au singulier ou des groupes de mots. Les termes qui les dénotent sont appelés labels. Lorsqu'il y a plusieurs labels pour un même concept (synonymes, équivalents dans d'autres langues), cela étend la portée de ce concept et offre une richesse linguistique. D'autres annotations que les labels peuvent qualifier les concepts. Il peut par exemple s'agir de définitions, de consignes d'usages. Quand une classe est définie par combinaisons d'autres classes et relations dans une classe d'équivalence (*equivalentClass*), c'est une *classe définie*.
    - Les instances sont les enfants d'un concept auquel il n'est plus possible d'ajouter d'attributs. Elles sont les extensions des concepts dont elles possèdent exactement les propriétés. Ce sont elles qui portent les connaissances de l'application à laquelle participe l'ontologie. Le nom d'individu (*NamedIndividual*) leur est aussi donné. Il est dit alors que les instances « peuplent » l'ontologie. Il n'est pas nécessaire que l'ontologie soit instanciée, ou peuplée. « En sciences du vivant, on crée souvent des ontologies que de classes. » [79, CHARLET diapo. 126]
- Les relations sont les liens qui structurent les concepts entre eux. Elles participent à caractériser les concepts les uns par rapport aux autres. Les associations entre les concepts sont explicites car chaque lien est typé :
    - relations de subsomption ou hiérarchiques dites aussi de genre-espèce (est_un) ou (is_a). Il y a alors une notion d'héritage. En linguistique, on parle d'hyponymes et d'hyperonymes. Lorsque l'ontologie est formalisée en RDF ou OWL, ces relations se

nomment *subClassOf*. Si les ontologues s'accordent sur le principe d'éviter les polyhiérarchies, les instances peuvent parfois hériter de plusieurs pères.

- relations de type associatif autres que celles de subsomption (*objectProperties*). Il existe par exemple un type d'association spécifique comme les relations partitives. Elles sont appelées aussi tout-partie (partie_de) ou (part_of). Les enfants y sont des composants. En linguistique, on parle de méronymes.

- Les attributs (*dataProperties*) sont des caractéristiques propres aux concepts. Ils relient les concepts de l'ontologie à des données typées (types choisis dans les types XML comme par exemple *Positive Integer*). Ils peuvent faire penser aux champs d'une table dans une base de données. Les attributs peuvent avoir différentes valeurs qui peuvent aussi être vues comme des facettes. Par exemple l'attribut « sexe » peut avoir la valeur mâle, femelle ou hermaphrodite et l'attribut « age », la valeur 33.

Ces éléments accroissent la puissance sémantique des ontologies. Ils font d'elles des outils qui modélisent de façon très précise des faits et des connaissances d'un domaine. En y associant un moteur d'inférences, des raisonnements peuvent avoir lieu et des déductions peuvent être faites.

L'organisation des concepts des ontologies sont formellement des treillis (ou graphes orientés dirigés). Visuellement, elles peuvent être représentées sous la forme d'arbre où un concept peut apparaître n fois s'il a n pères. Ces représentations peuvent faire penser aux cartes conceptuelles car les liens sont explicitement typés, contrairement aux cartes heuristiques.

2.3 Les langages de représentation des connaissances

2.3.1 Leurs caractéristiques et leur historique lié à celui du Web

Il est nécessaire d'encoder les modélisations afin d'en faire des ontologies formelles autrement appelées informatiques ou encore ontologies computationnelles par anglicisme, afin de les rendre exploitables par les machines. Les langages alors utilisés rendent exploitables les axiomes, c'est-à-dire les règles appliquées aux concepts ou aux relations, pour réaliser des inférences. La syntaxe utilisée est « *human readable* » pour qui connaît les langages à balises que sont HTML et XML. En effet, les langages pour les ontologies sont des extensions de ces syntaxes qui servent à présenter et échanger des données. Le premier d'entre eux, SHOE, pour Simple HTML Ontology Extension, a été créé en 2000 au sein de *l'University of Maryland* puis adapté à XML. Depuis, d'autres langages reposant sur la syntaxe XML sont apparus comme XOL évincé ensuite par RDF, et RDFS, OIL, DAM+OIL et OWL avec des niveaux d'expressivité supplémentaires. Aujourd'hui, c'est donc RDF, RDFS et OWL qui correspondent aux ontologies dans les piles des standards du Web sémantique. [26, RICHY *et al.*]

2.3.2 OWL dans l'architecture du Web sémantique

OWL dans l'architecture du Web sémantique : « En 2001, le W3C a formé un groupe appelé *Web-Ontology (WebOnt) Working Group* dont le but était de concevoir un nouveau *Ontology Markup Language* pour le Web sémantique. Le résultat de leurs travaux est le langage OWL. » [26, RICHY *et al.*]. Ainsi la modélisation ontologique est particulièrement compatible avec le web sémantique.

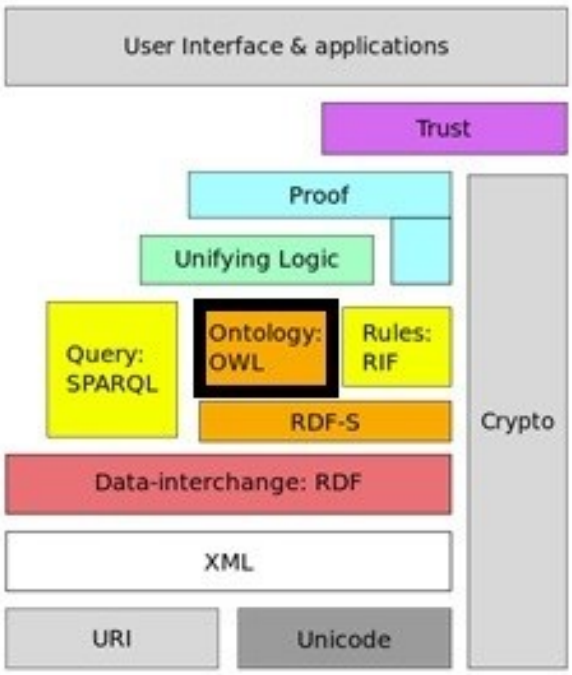


Figure 2 : OWL dans le Semantic Web Layer Cake ou la pile des standards du web sémantique d'après Le layer cake de Tim Berners-Lee de 2006 par Mhermans.[14]

OWL dont la deuxième version est devenue une recommandation du W3C en décembre 2012[15] est en réalité une famille de langages : OWL Lite (*lightweight ontologies*), OWL DL (DL pour logique de description), OWL Full (*heavyweight ontologies*). Chacun des langages est une extension de l'autre. Il s'agit de trois fragments d'expressivité croissante [57, GANDON *et al.* p. 21]. Ces langages permettent d'opérationnaliser les ontologies de façon formelle, tel qu'on les formalise sur papier avec des liens typés et des règles plus ou moins contraignantes. La machine peut alors réellement faire des raisonnements complexes.

14 Travail personnel. 24 juin 2007. Sous licence Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons : <http://commons.wikimedia.org/wiki/File:SW_layercake_2006.svg#mediaviewer/File:SW_layercake_2006.svg >

15 < http://www.w3.org/TR/owl2-new-features >

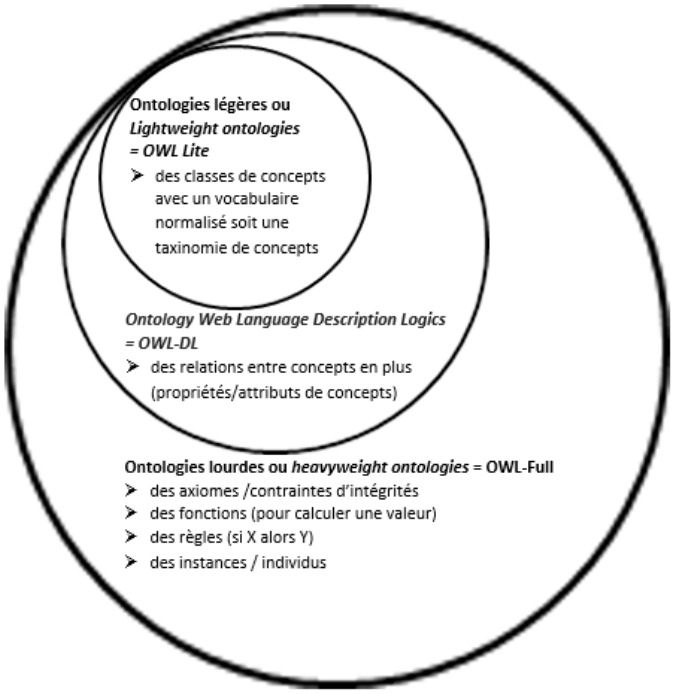


Figure 3 : les trois langages d'OWL

Figure inspirée par RICHY H., DESPRÉS S. Métadonnées, ontologies et documents numériques. Techniques de l'ingénieur. 2007. p. h7155v2 [26, RICHY *et al.*].

2.4 Les méthodes et les phases de construction

2.4.1 Construire un modèle et le faire expertiser

Pour toute ontologie, cette étape est nécessaire. Il n'existe pas un modèle unique. En effet, chaque ontologie est conçue pour une application particulière, pour réaliser des raisonnements et des tâches spécifiques. De plus un domaine peut faire l'objet de plusieurs consensus différents. Enfin les connaissances de ce domaine ne seront pas forcément mobilisées de la même façon dans diverses applications destinées pourtant au même domaine.

Dans l'ouvrage somme sur l'IA, [33, AUSSENAC-GILLES *et al.*] font la synthèse des différentes méthodes de construction de modèles dans le chapitre consacré à l'histoire de l'IC [*Ibid.* p. 621]. Il existe les méthodes ascendantes et les méthodes descendantes. Ils ajoutent que ces méthodes se combinent au cours de la construction.

- Les méthodes ascendantes (*bottom-up*) :

Il s'agit de partir des données pour dessiner le modèle selon les besoins. Par exemple, dans le cas de la création d'une ontologie de domaine, ce sont les concepts présents dans la documentation propre à ce domaine qui serviront de point de départ.

- Les méthodes descendantes (*top-down*) :

Dans le cas des méthodes descendantes, il n'est pas question de s'occuper des données ou des sujets abordés dans la documentation de ce domaine. A la manière dont est construit un thésaurus, il est porté attention « aux faits, évènements, objets et des processus ». Toujours selon ces auteurs,

ces méthodes privilégient l'adaptation et l'assemblage de composants d'autres modèles ainsi réutilisés[16]. Même si « récupérer une théorie faite par quelqu'un d'autre, la comprendre et la manipuler, ce n'est pas chose aisée non plus. » [27, BACHIMONT].

Comme pour la construction de systèmes experts et de systèmes à base de connaissances (SBC), le recours à l'expertise humaine pour valider de façon consensuelle le modèle est nécessaire. Peut-être pourrait-il aussi être question de solliciter ceux qui, au quotidien, seront amenés à manipuler ces modèles dans des applications métiers ? Peut-être que ceux qui seront chargés de les mettre à jour avec des logiciels plus ou moins bien appropriés pourraient eux aussi être sollicités ?

La sollicitation des experts est difficile car ils sont peu disponibles. Par conséquent, l'acquisition des connaissances s'est orientée par le développement d'outils et de méthodes. Elle a donc collaboré de plus en plus avec d'autres disciplines comme par exemple celles du génie logiciel et de la linguistique informatique. En 1997, elle a été renommée ingénierie des connaissances (IC) pour prendre acte d'un champ d'expertise dépassant la seule acquisition des connaissances pour un SBC.

2.4.2 Ingénierie des connaissances (IC) et traitement automatique des langues (TAL)

Les experts sont souvent peu disponibles. Pour contourner cette difficulté, des corpus textuels porteurs de la connaissance d'un domaine sont constitués afin d'y relever des candidats termes et des marqueurs de relations sémantiques. Cette extraction est aujourd'hui automatisée. La linguistique informatique a progressé avec le TAL ou traitement automatique du langage naturel (TALN et NLP pour *Natural Language Processing* en anglais). Avec l'élaboration de logiciels spécialisés d'analyse, cette pratique d'« utilisation de textes comme sources de connaissances a pris un essor autour de 2000 » [34, AUSSENAC-GILLES *et al.* p. 136]. Toutefois, il est nécessaire de consulter les experts pour faire valider ces corpus, les candidats termes, puis les concepts et les ontologies qui en découlent. « *Parce que sélectionner un terme, c'est une chose, savoir ce qu'il signifie en est une autre.* » [27, BACHIMONT]. Pour construire des ontologies de domaine, la terminologie est « *l'outil qui fixe le langage spécialisé d'une communauté de pratique*. » [39, DJAMBIAN p. 80]. Elle ne représente pas à elle seule l'ontologie. Il convient d'ajouter que ces techniques d'extraction automatiques de termes sont aussi utilisées pour élaborer d'autres types de ressources tels des vocabulaires contrôlés comme des thésaurus par exemple. La littérature en IC mentionne alors les expressions « ressources terminologiques ou ontologiques » ou encore « ressources termino-ontologiques », abrégées par l'acronyme RTO [34, AUSSENAC-GILLES *et al.*].

[16] Ontology Design Patterns (ODPs), initié dans le cadre du projet européen NeOn (Network Ontologies), est un Wiki où une communauté de spécialistes des ontologies discute et valide des modèles. Un catalogue de modèles est librement mis à disposition. Leur réutilisation permet minimiser les efforts nécessaires à la création d'ontologies riches et rigoureuses en bénéficiant du contrôle de la qualité effectué par la communauté d'ODPs.
< http://ontologydesignpatterns.org >. (Cf. Annexe 1- Livrable 3- État de l'art : répertoire des ressources termino-ontologiques disponibles dans le domaine. p. 22/34)

La littérature professionnelle en information-documentation les désigne plutôt comme des terminologies et des référentiels [55, DALBIN]. Une RTO peut aussi être une ontologie dont la dimension linguistique est bien développée.

Il est donc fait appel à un sous-ensemble du *Data mining*, le *Text mining,* pour extraire la terminologie de ces corpus [86, SAINT-LÉGER]. C'est « *une étape dans le processus d'extraction des connaissances, qui consiste à appliquer des algorithmes d'analyse des données* » [85, SAPORTA diapo. 8]. Le TAL, ou TALN, combine plusieurs processus complexes d'analyse œuvrant à désambiguïser le langage naturel. En effet, le langage est équivoque pour plusieurs raisons : polysémie, synonymie, implicite.

« *plusieurs niveaux d'analyse interagissent dynamiquement* »

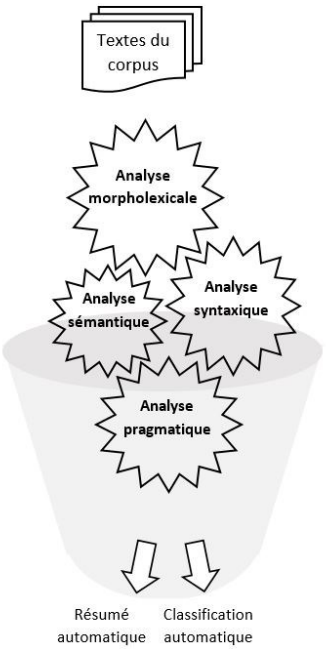


Figure 4 : le traitement automatique des langues
Figure inspirée par BALMISSE G. Guide des outils du knowledge management: panorama, choix et mise en œuvre. [36, BALMISSE pp. 259-268].

Le processus de classification est particulièrement intéressant pour la création et pour l'exploitation d'ontologies. Pour appréhender cette double fonction, il convient de distinguer les deux types de classification [81, 82, NIANG-KEITA] :

- la catégorisation. Elle est dite aussi classification supervisée, car elle est réalisée en fonction d'un schéma de classe représentatif de ce que contient le corpus.
- la segmentation *(clustering)*, pour laquelle le regroupement des termes (ou des textes dans une autre perspective) est fait par similarité et non en fonction d'un plan de classement préétabli.

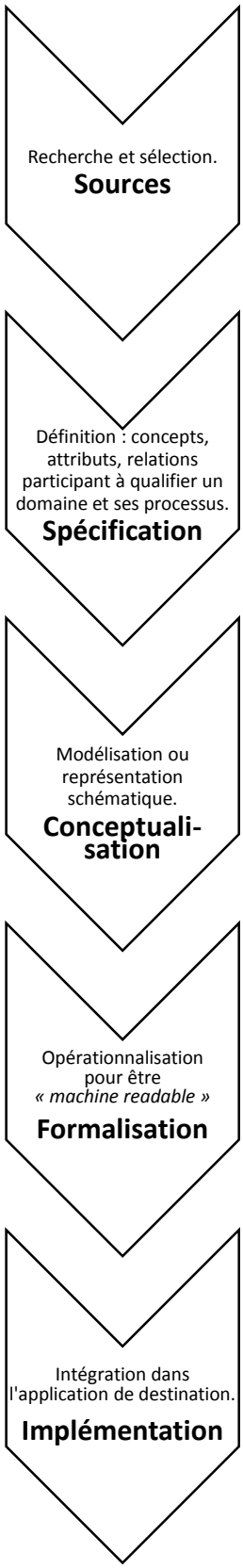
Le TALN est donc à la fois producteur et utilisateur de RTO [34, AUSSENAC-GILLES *et al.* p. 136]. Il est producteur de RTO grâce à l'usage de la lexicométrie qui utilise des méthodes statistiques, la segmentation pour dégager un vocabulaire. La lexicométrie se base sur la fréquence de certains termes dans le corpus mais aussi en fonction des cooccurrences de termes [86, SAINT-LÉGER]. Par ailleurs, le TALN est consommateur de RTO car elles participent à superviser l'indexation automatique de corpus au sein d'une application. La construction et les utilisations envisagées pour l'ontologie de la toxicologie nucléaire, dont l'étude est abordée par la suite, illustrent d'ailleurs ce double usage (7 Un référentiel ontologique pour le programme Toxicologie Nucléaire).

Aujourd'hui, la primauté donnée aux corpus est une pratique courante pour la création d'ontologies médicales dont les auteurs de cet article présentent plusieurs exemples [44, CHARLET *et al.*]. Il y ait souvent pris appui sur des classifications médicales et sur le thésaurus MeSH. Dans d'autres domaines scientifiques, il en est de même. Par exemple, le pôle de compétence Gestion des connaissances (GeCo) au sein du service Information scientifique et technique (IST) de l'INRA ne crée pas ex nihilo des RTO. Dans la majorité des cas, le pôle GeCo met ses compétences en *Text mining* ainsi que ses outils tels que Luxid®[17] au service des équipes de recherche et fusionne, adapte ou aligne des RTO.

2.4.3 Appliquer une méthodologie éprouvée et suivre une démarche projet

Dans le chapitre du 1er volume de la trilogie consacrée à l'histoire de l'IA, les auteurs du chapitre traitant de l'IC recensent les différentes méthodes et plateformes de construction d'ontologie [33, AUSSENAC-GILLES et al. p. 626]. C'est celle de B. Bachimont a inspiré la construction de l'ontologie de la toxicologie nucléaire, ToxNuc, étudiée durant le stage et présentée dans la 2e et 3e partie de ce mémoire. Les phases de la méthodologie de J.-M. Pinon se rapprochent quant à elles de celles de la construction d'un thésaurus. La méthodologie représentée par la figure suivante est une synthèse de ces deux méthodes.

[17] Logiciel de fouille de textes, de TALN pour guider la construction de vocabulaire Société Temis < http://www.temis.com/fr/luxid-6 >



- Ressources ontologiques : ontologies, ontology design paterns (ODPs)
- RTO* : thésaurus, taxinomies et classifications
- Ressources non ontologiques : experts, glossaires, lexiques, corpus (opérations de TALN)

➲ Candidats termes dont la signification peut se révéler ambiguë.

* cette expression est employée dans la communauté de l'IC lorsqu'il y a utilisation du TALN

- Normalisation linguistique

ou *sémantique différentielle* de Bruno Bachimont
- une seule signification des termes = une seule interprétation possible par un humain
- construction d'un système de différence entre les termes retenus

➲ *Engagement sémantique* car l'ontologie obtenue est valable pour un contexte particulier, elle n'est pas universelle (*arbre ontologique, concepts sémantiques, ontologie régionale*).

- Normalisation sémantique

- Définir les concepts et les relations entre les concepts de l'ontologie différentielle en fonction des besoins et de l'application, et en recherchant de la modularité (*définitions formelles et ajout de concepts formels définis dont la sémantique est référentielle pour former un treillis formel*)

➲ *Engagement ontologique* car l'ontologie est le résultats de choix spécifiques (*ontologie référentielle*)

- Utilisation des langages de représentation des connaissances : RDFS, OWL
- Ingénierie : utilisation d'éléments existants et assemblage (URI, alignement, formalisation de la documentation au sein du fichier de l'ontologie avec DC, VIAF, FOAF, VOAF, ccREL...)
- Opérations, règles (*sémantique computationnelle, ontologie computationnelle*).

➲ L'ontologie obtenue permet de générer des inférences

- Exécution de tâches

➲ Dans le cas de la recherche scientifique publique, l'ontologie peut faire l'objet d'une valorisation des données de la recherche avec une démarche de publication et de publicisation.

Figure 5 : les étapes méthodologiques de construction d'ontologie, une tentative de synthèse d'après [35, BACHIMONT ; 37 CHAUMIER].

Le cycle de vie idéal présenté par Fabien Gandon ressemble fortement aux phases de la gestion de projet avec sept notions à prendre en considération :

- détection des besoins mais aussi évaluation de ceux-ci
- conception
- gestion et planification
- évolution
- diffusion
- utilisation
- évaluation

Cependant, Fabien Gandon explique que ces étapes posent chacune des problèmes relevant encore de la recherche. C'est bien pour cela qu'il nomme ce processus « *La vie rêvée des ontologies* » [40, GANDON]. La figure suivante combine les activités de développement et celles propres à la gestion de projet comme la planification, le contrôle et l'évaluation de la qualité.

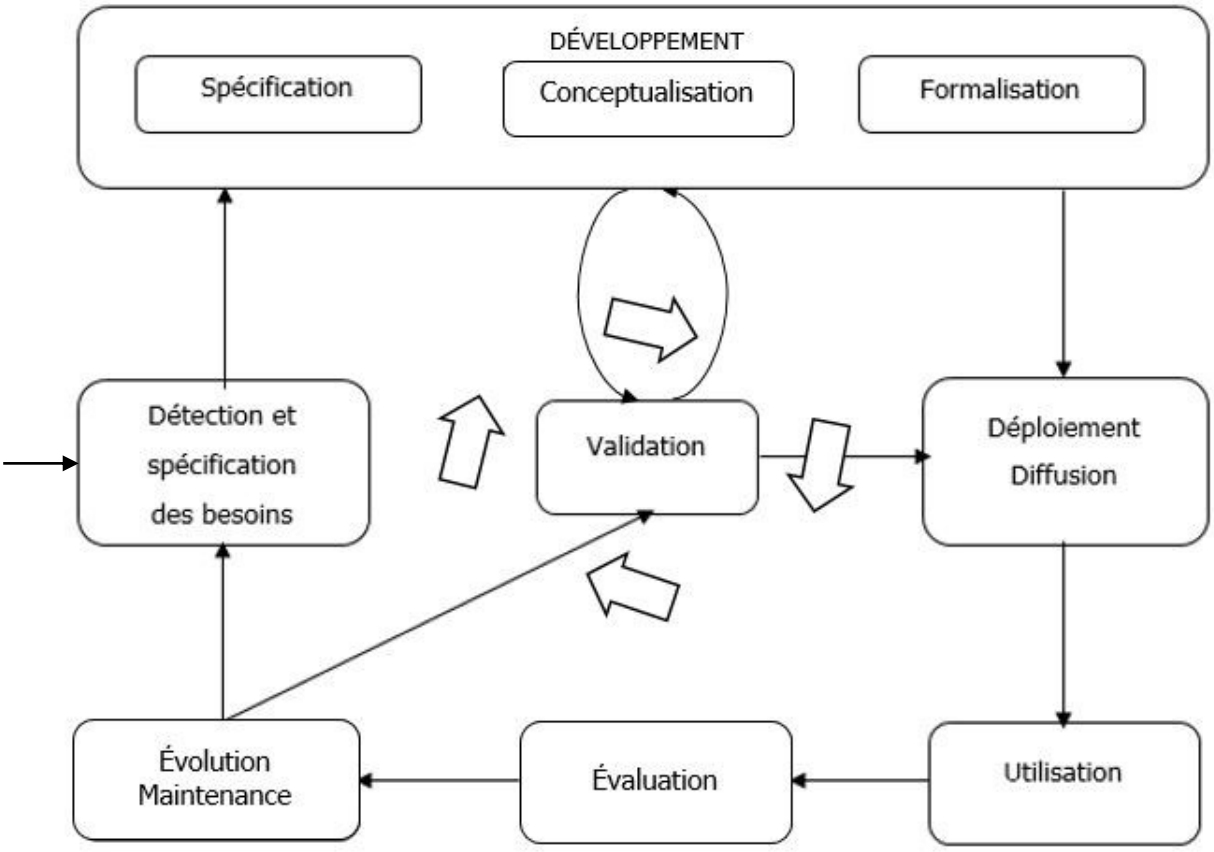


Figure 6 : le cycle de vie d'une ontologie[18]

[18] Reproduction de la figure 1 p. 150 de BANEYX A., CHARLET J. « Évaluation, évolution et maintenance d'une ontologie en médecine: état des lieux et expérimentation ». Revue I3 – Information, Interaction, Intelligence, numéro spécial « Corpus et ontologies ». 2007. p. 147-173.

3 Le point sur quelques confusions

3.1 Ontologies et taxinomies, ontologies et thésaurus… éclaircir le flou terminologique

Comme observé précédemment, la notion d'ontologie recouvre plusieurs significations au sein même de la communauté de l'IC. Cette diversité de définitions se propage aux autres communautés. Cela a été constaté au cours du stage par l'étude de la littérature et à l'occasion de diverses rencontres : sociologue des techniques d'information et de communication et de l'innovation, responsable SI de l'Inra, linguiste informaticien au sein du service IST de l'Inra, consultant en architecture de l'information spécialiste des thésaurus, consultants en technologies sémantiques. A l'échelle des recherches menées dans le cadre de la mission de stage, la communauté des techniciens du Web et celle des professionnels de l'I&D ne semblent pas faire exception. La situation n'aurait pas changé depuis le constat fait par Bruno Menon en 2005, date de publication dans *Documentaliste - Sciences de l'information* de son article sur les langages documentaires [6, MENON]. Désorientante et déstabilisante au premier abord, cette polysémie apparaît ensuite comme étonnante et amusante. En effet, si l'une des finalités de ces objets est de proposer un langage commun et univoque, cette polysémie n'est-elle pas paradoxale ?

Au-delà des termes… quelles sont les caractéristiques de ces trois objets ? Quelles sont les similarités ? Quelles sont les différences ?

3.2 Ontologies et taxinomies/taxonomies

> « Ontology is not a synonym of Taxinomy » [28] GANDON diapo. 14]
> « Taxonomical knowledge is a kind of ontological knowledge among others » [*Ibid*, diapo. 15]

Comme pour la notion d'ontologie, les professionnels de l'I&D ont également des difficultés à se saisir de la notion de taxonomie. La preuve en est certains des titres de l'article consacré à cette entrée dans la rubrique « Le Vocabulaire de la doc » sur le site de l'ADBS[19]. Pour preuve aussi, la longueur de cet article comparée à celle de l'article qui définit le terme « Thésaurus » dans ce même vocabulaire. Cette rubrique, des éléments de l'article de Bruno Menon préalablement cité sont repris.

Le terme *taxinomie*, créé en 1813 par le botaniste Augustin Pyrame de Candolle, est aussi couramment orthographié *taxonomie* en raison de son origine étymologique controversée[20]. La taxinomie vise à classer et à décrire de façon systématique et hiérarchisée les êtres vivants.

---

[19] « *Des définitions déroutantes* », « *Des conceptions hétérogènes* »
< http://www.adbs.fr/taxonomie-58346.htm?RH=OUTILS_VOC >
[20] TARDIEU C. « La bonne orthographe du mot taxinomie. Un concept important dont l'orthographe est malmenée ». *PALEO. Revue d'archéologie préhistorique*. 2011. n°22, p. 331–334.

Tout comme la taxinomie du vivant est bouleversée elle-même, la définition de ce que recouvrent les taxinomies s'est élargie. Depuis la fin des années 1990, ce terme, majoritairement orthographié taxonomies est employé dans un nouveau contexte. Il s'agit de celui des sites Web et des portails intranet des entreprises avec l'apparition d'arborescences hypertextuelles guidant la navigation. L'éditeur de logiciel de gestion de vocabulaires Mondeca et les professionnels de l'I&D et des SIB les nomment « taxonomies de navigation » [23, MONDECA ; 24 REMILLIEUX ; 13 AMAR]. D'ailleurs, la nouvelle norme ISO 25964-1:2011 sur les thésaurus confirme cette acception. La norme consacre les sections 17 à 24 aux définitions des autres types de vocabulaires n'ayant généralement pas fait l'objet d'une normalisation spécifique. La section 19 consacrée aux taxonomies regroupe ces diverses définitions [7, ISO 25964-1:2011 p. 59-63] :

> « The typical taxonomy is presented as a hierarchical vocabulary, used for classifying or categorizing, organizing, browsing, navigating, searching and/or filtering any type of content in networked environments. A common use case is to support navigation, especially by hierarchical organization and browsing through a broad set of electronic resources, e.g. websites, intranets, portals, wikis. Taxonomies are often used to provide website menus. To complement the navigational features with a search capability, taxonomies can include synonyms operating behind the scenes as entry terms, and "See also" references between related categories in the hierarchy." […] Taxonomies are commonly used for resource navigation in portals, intranets and websites, and they support retrieval primarily by enabling browsing. »

Ces fonctions d'aide aux utilisateurs pour explorer et repérer de l'information spécifique sur les sites Web « s'apparentent à celles des langages documentaires » [6, MENON p. 22]. Si elles sont arborescentes comme les ontologies, les taxonomies sont visibles directement des usagers des applications numériques. Elles ont une valeur heuristique qui est pertinente seulement pour l'humain. [25, ZACKLAD] les désigne comme relevant du paradigme de la navigation. Contrairement à cela, les ontologies sont intégrées dans d'autres outils et servent d'une autre façon la recherche d'information via les moteurs de recherche de ces sites. Ainsi, même si une ontologie peut se limiter à une organisation taxonomique, sa formalisation en fait un artefact plus sophistiqué. Car c'est la formalisation qui permet de contribuer à la recherche d'information, même si cette organisation est limitée au seul type de la relation hiérarchique. En revanche, comme pour les ontologies, les concepteurs des taxonomies ne sont pas prioritairement des professionnels de l'I&D. En résumé, les similarités et les différences entre taxinomies et ontologies s'observent du point de vue structurel, fonctionnel, front office/back office et en considérant aussi leurs concepteurs. Il est également possible de résumer cette assimilation en reprenant ici les mots de [57, GANDON *et al.* p. 84] :

> « Ce squelette taxinomique est […] la principale source de confusion entre les ontologies et d'autres structures ayant une hiérarchie pour squelette, tels les thésaurus. »

3.3 Ontologies et 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.» [7, ISO 25964-1:2011]

Pour les professionnels de l'I&D, le terme de thésaurus apparaît dans les années cinquante avec l'augmentation importante de la documentation technique [11, DALBIN p. 76 et 78]. Il s'agit d'un langage documentaire spécialisé.[21]

Il est utilisé pour l'indexation et la recherche, non de documents dans des collections, mais de l'information contenue dans les documents. Un thésaurus a pour vocation de décrire un seul domaine de la connaissance ou de quelques domaines liés entre eux. Exploité pour la recherche informatisée dans un fonds documentaire spécifique, son organisation interne caractérise également le champ d'activité ou la discipline qu'il permet d'indexer.

C'est un lexique spécialisé dans un domaine dont chacun des termes a été sélectionné pour désigner un concept de façon non ambigüe. Ces termes sont appelés des *descripteurs*, leurs synonymes dont l'emploi est proscrit car ils pourraient suggérer des ambiguïtés sont quant à eux nommés *non-descripteurs*. Ce lexique est complété par des définitions et notes d'application pour faciliter son emploi. Les thésaurus sont, comme les *listes de vedettes matières, des langages combinatoires* mais dits *postcoordonnés*. En effet, la majorité des concepts sont unitermes et peuvent être combinés au moment de la recherche pour représenter un sujet.

Les relations sémantiques entre les termes du lexique sont donc des relations d'équivalence. Mais les thésaurus sont aussi caractérisés par des relations hiérarchiques et associatives, voire interlinguistiques. Ces relations sont représentées par une notation symbolique :

- TG : terme générique - en anglais, BT : *broader term*
- TS : terme spécifique - en anglais, NT : *narrower term*
- EP : employé pour – en anglais, UF : *use for*

Ces signes conventionnels interprétables par les humains le sont aussi par les machines depuis que les thésaurus ont fait l'objet d'une formalisation informatique. Ainsi, leur représentation informatique permet des rebonds via des liens hypertextes. Leur formalisation informatique et les interfaces de recherche plus ergonomiques permettent surtout de guider les requêtes des utilisateurs

[21] Quelques exemples : le thésaurus de la Banque de données en santé publique (BDSP) est édité par l'École des hautes études en santé publique (EHESP). Le MeSH (Medical Subject Headings) est le thésaurus biomédical de référence. Il est utilisé pour indexer et rechercher des références dans le catalogue de la National Library of Medicine (NLM) et de la base MEDLINE/PubMed. Depuis 1986, il est traduit en français par l'Institut national de la santé et de la recherche médicale (Inserm).

finaux lors de leurs interrogations des bases de données documentaires. Ainsi des listes de termes à cocher issues des branches du thésaurus après une première recherche peuvent être proposées pour reformuler les requêtes. Sylvie Dalbin parle alors de « Thésaurus de recherche » [11, DALBIN]. L'extension à d'autres termes que ceux de la requête peut également être automatique. Ainsi, la recherche sur des termes désignant des concepts plus spécifiques du domaine, appelée autopostage spécifique. Elle s'oppose à l'autopostage générique qui permet d'étendre la recherche à des ressources indexées par des concepts du niveau supérieur. Ces opérations relèvent déjà d'une forme d'expansion sémantique.

Ainsi, avec une structure sémantique formalisée et opérationnalisée, mais aussi avec des applications dans le domaine de la recherche d'information, il est facile d'assimiler ontologies informatiques et thésaurus. Là encore, c'est une citation de [57, GANDON *et al*. p. 86] qui éclairent la distinction entre ces deux notions :

> « Le thésaurus est un recueil de termes (de la langue naturelle) et il organise directement ces termes et non des catégories, des concepts ou des classes comme dans le cas des ontologies. »

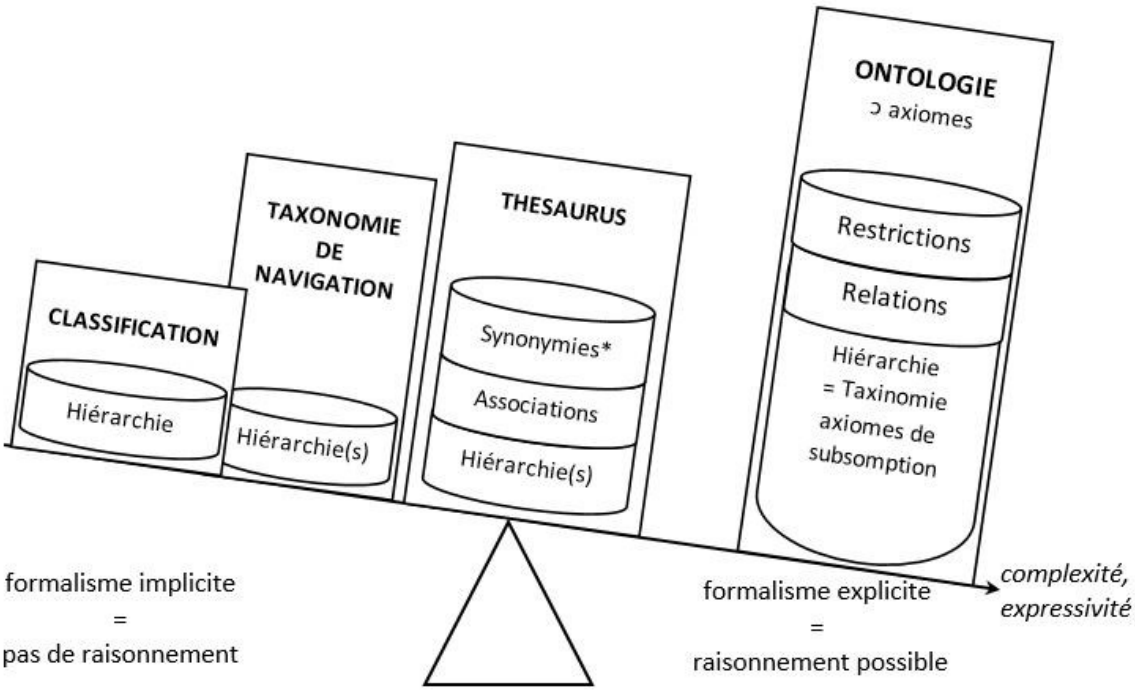
Les ontologies correspondent à « l'organisation des thésaurus poussée au bout de la formalisation ». La dimension linguistique des thésaurus en font des ressources intéressantes pour les ingénieurs de l'IC lorsqu'ils ont à concevoir des ontologies dont la dimension linguistique est bien développée [79, CHARLET diapo 119].

Cependant, la confusion peut encore subsister. En effet, ces termes du langage naturel, à la fois signifié et signifiant, représentent eux aussi des concepts d'un champ d'activité ou d'une discipline. Mais comme le montre la définition donnée par la nouvelle norme ISO 25964-1:2011 précitée, « *l'orientation "concept" a été renforcé et formalisé* » [10, DALBIN]. Pour tenter à nouveau de distinguer thésaurus et ontologie, il est possible d'insister sur le fait qu'une ontologie doit permettre de répondre à des requêtes et d'exécuter des tâches. La formalisation, ou la grammaire formelle, est sensiblement la même pour tous les thésaurus (au minimum TG/BT et TS/NT). Et cela, quel que soit le domaine, alors qu'un domaine peut faire l'objet de plusieurs ontologies pour des applications autres que purement documentaires (d'applications, de situation, d'activités, de tâches). En effet, il est possible d'inventer autant de grammaires pour spécifier les relations et les contraintes que d'applications pour un même domaine. Ce sont alors les classes et les types de RDFS ou OWL qui sont utilisés selon la complexité et la sophistication des raisonnements nécessaires. Ainsi, un thésaurus ou un autre type de SOC comme une classification peut tout à fait être « ontologisé ».[22] Les syntaxes SKOS ou OWL permettent de franchir la barrière des symboles de la notation conventionnelle pour faire entrer les SOCs dans le Web sémantique. En effet, elles formalisent les

[22] C'est d'ailleurs le cas du thésaurus Agrovoc créé et maintenu par l'organisation des Nations unies pour l'alimentation et l'agriculture (FAO) depuis 1980 : < http://aims.fao.org/fr/agrovoc/concept-scheme#.VHvCijGG_Ss >

relations et les concepts selon les standards du Web sémantique car elles respectent la structure basée sur les triplets RDF et elles utilisent des URI.

Comme cela a été vu précédemment à propos des méthodes de conception des ontologies, des RTO et plus spécifiquement parfois des thésaurus sont utilisés pour créer des ontologies. Cette pratique est peut-être une source de confusion supplémentaire entre les ontologies et les thésaurus. L'effort de normalisation du langage naturel fourni par les professionnels de l'I&D et des SIB bénéficie aux ingénieurs de l'IC. Pour autant, [43, CHARLET *et al.*] mettent en garde sur le fait que les thésaurus ne sont « pas des embryons de » mais que « des ressources pour amorcer une ontologie ».



*Synonymies comprenant ici aussi le multilinguisme

** Hiérarchie(s) : des relations de type hiérarchique et/ou partitive.

Figure 7 : l'expressivité des SOCs. Figure inspirée en partie par l' « ontology spectrum » de [50, MCGUINNESS pp. 171-194].

Dans cette figure, l'orthographe « taxonomie » a été choisie en cohérence avec les auteurs de l'expression « Taxonomie de navigation ». Quand il s'agit de la structure hiérarchique de l'ontologie, c'est l'orthographe « taxinomie » qui a été choisie. Contrairement à la taxonomie de navigation, la taxinomie de l'ontologie est construite uniquement via des instances de la relation de subsomption parfaitement précisées. Les informations contenues dans l'ontologie constituent un ensemble d'axiomes formels permettant l'exploitation par un ordinateur du modèle ainsi explicité.

3.4 Ontologies versus SKOS

SKOS dont l'acronyme est *Simple Knowledge Organization System* est une syntaxe dédiée à la représentation des SOCs en RDF (*Ressource Description Framework*), le formalisme standard du *Web of Data* ou *Linked Open Data* (LOD) soit le Web de données. SKOS fait l'objet d'une recommandation du W3C depuis 2009[23]. Par ailleurs, SKOS fait aussi l'objet d'un schéma de concepts (*Concept Scheme*), ce qui peut ajouter à la confusion. SKOS est inspiré par les thésaurus car ses concepteurs ont suivi les recommandations de bonnes pratiques des normes ISO 2788:1986[24] révisée par les normes [7, ISO 25964-1:2011 ; 8, ISO 25964-2:2013]. Mais SKOS est suffisamment générique pour représenter les autres SOCs que les thésaurus comme les taxinomies et les systèmes de classification. Ainsi, grâce à SKOS, l'héritage documentaire issu de la communauté de l'I&D et des SIB peut migrer vers le Web sémantique. SKOS s'appuie sur RDF, RDFS et dans une moindre mesure sur OWL dont la logique formelle est trop stricte pour prendre en compte les usages du langage naturel. Et par ailleurs, « *OWL peut être exploité pour étendre les possibilités de SKOS et inversement !* » [79 CHARLET diapo. 112-113].

Des données indexées avec des vocabulaires contrôlés formalisés avec le langage de représentation SKOS peuvent être portées dans le Web afin d'en faciliter l'échange. Ainsi le langage Rameau a fait l'objet d'une première expérimentation de conversion en SKOS [47, ISAAC *et al.*]. En outre, en plus de typer les liens qui font la sémantique propre des thésaurus, des relations spécifiques ont été déclarées. Il s'agit de relations dédiées à la spécification de différents liens d'équivalence conceptuelle entre plusieurs vocabulaires contrôlés afin de les interconnecter entre eux. Ainsi, le *SKOS Mapping* évite de fusionner des vocabulaires et permet à ces derniers de rester autonomes les uns vis-à-vis des autres. Une fois le LCSH de la Library of Congres converti en SKOS, une correspondance entre Rameau et le LCSH pourrait être exploitée pour différents usages. Le rôle que ces langages pourraient jouer une fois passés à l'échelle du Web de données est encore à inventer. Mais il est sûr que l'opérationnalisation de leur sémantique formelle les rend exploitables par les moteurs de recherche.

Pour autant, les vocabulaires contrôlés sont-ils solubles dans SKOS ? Sylvie Dalbin indique que SKOS est un format d'échange et d'exploitation des vocabulaires dans l'écosystème du Web de données. Elle rappelle qu'il « *ne contient pas tous les éléments qui pourraient être utiles en exploitation ou en gestion ou en maintenance* » [DALBIN, 10]. Cela est très gênant pour l'échange de thésaurus entre professionnels. Les échanger en SKOS fait perdre les données de gestion et l'historique des modifications par exemple. Si SKOS formalisait l'explicitation de la conceptualisation de ces données, il ne serait plus *Simple*. Il faut donc espérer que les éditeurs de logiciels de gestion

[23] < http://www.w3.org/TR/2009/REC-skos-reference-20090818 >
< http://www.w3.org/2004/02/skos >
[24] ISO 2788:1986 - Documentation -- Principes directeurs pour l'établissement et le développement de thésaurus monolingues. N.B. : N'étant plus à jour, elle ne figure plus au catalogue des normes.

de vocabulaires implémentent la norme ISO 25964 [7 - 8]. Par ailleurs, comme le schéma de concepts SKOS est suffisamment générique pour être commun à différents SOCs, il manque d'éléments pour représenter encore d'autres notions que celles de gestion. C'est par exemple le cas pour les notions liées à la coordination des vedettes-matières : têtes de vedettes, subdivisions, construites[25]. Il est alors nécessaire de combiner le schéma de concepts SKOS avec d'autres ontologies adaptées pour représenter les SOCs. Certains de leurs éléments peuvent compléter le schéma de concepts de SKOS : MADS/RDF, FRAD, FRASD afin de ne pas perdre en richesse lors de la seule « SKOSification ». Il est également possible de créer des extensions pour SKOS. Le W3C a d'ailleurs proposé une extension « *pour représenter les libellés en tant que ressources* » : Simple Knowledge Organization System eXtension for Labels (SKOS-XL)[26, 27]. Emmanuelle Bermès souligne aussi que la modélisation SKOS « *n'a pas vocation à supplanter les référentiels documentaires dans leur contexte d'utilisation originel* » [54, BERMÈS p. 53].

En insistant sur SKOS, il est possible que la confusion perdure avec les ontologies. Pour l'éviter, il est nécessaire de rappeler que les ontologies ne sont pas des thésaurus adaptés au Web de données mais le langage qui permet de les échanger dans le Web de données. Pour illustrer, il est possible de rappeler qu'il existe aussi des ontologies pour d'autres modélisations que celles des SOCs. Ainsi la modélisation *Dublin Core* (DC)[28] est aussi une ontologie. Ce modèle définit quinze éléments de base pour la description de ressources documentaires quelles qu'elles soient. Elle dispose d'URI et son schéma est exprimé en RDFS. L'ontologie *Dublin Core* n'est pas un thésaurus mais un schéma de métadonnées permettant l'exposition et l'échange de données bibliographiques dans le Web. Il n'est donc pas possible de réduire les ontologies informatiques à des thésaurus adaptés au Web. Cet exemple définit en creux ce qu'est une ontologie informatique. C'est un cadre logique et sémantique dont les langages expriment les relations entre instances, classes et attributs. Dans ce cadre il est possible d'élaborer soit un **vocabulaire de valeurs** *(Value Vocabularies)* soit un **schéma de données** [29] :

- un **vocabulaire de valeurs** *(Value Vocabularies)* ou autrement dit un SOC manipulable par les machines : une structure basée sur un réseau de concepts dont les termes ou labels lexicaux permettent d'indexer des informations spécialisées contenues dans un domaine. L'observation du vocabulaire de valeurs permet à lui seul d'appréhender le domaine de connaissance qu'il recouvre.

---

[25] « *Attention, SKOS est simple: on peut perdre de l'information !* » [48, ISAAC]
[26] http://www.w3.org/TR/skos-reference/skos-xl.html
[27] Le W3C propose un service de validation SKOS consultable à l'adresse < http://www.w3.org/2004/02/skos/validation > ; l'application Skosify permet de valider des fichiers SKOS, en ligne ou offline < http://code.google.com/p/skosify >. ThManager est une application open-source permettant d'éditer des thesaurus et de les exporter en SKOS < http://thmanager.sourceforge.netS >
[28] < http://dublincore.org/ >
[29] < http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-2011102 >

- un **schéma de données** ou autrement dit un **vocabulaire de métadonnées** *(Metadata Element Sets)* : une structure de description uniquement. En reprenant l'exemple du DC, les quinze éléments du schéma de métadonnées (*Title*, *Creator*…) permettent de décrire tous les types de ressources[30]. Si toutes les ressources ont un titre, un auteur, une date d'édition… la seule vue du schéma ne permet pas d'appréhender le périmètre du domaine de connaissance auquel appartient une collection de ressources spécifiques qu'on décide de décrire au moyen de ce schéma. Et donc, le modèle SKOS n'est pas un thésaurus, car son observation révèle seulement qu'il y a des concepts, que ceux-ci sont représentables par des termes, et que les concepts sont liables à d'autres concepts. SKOS fournit un modèle extensible pour représenter les SOCs comme les thésaurus dans le Web de données.

Par ailleurs et pour tenter d'être exhaustif, il est à rappeler qu'un autre standard que SKOS s'applique pour la recherche et le management de l'information ainsi que pour la représentation et l'échange des connaissances. Il s'agit de la norme *Topic Maps* (ISO 13250) pour les cartes de thèmes (ou cartes topiques). Il est accompagné de la norme XML Topic Maps (XTM), dérivée du langage XML, pour le rendre opérationnel sur le Web. Bien que publiée par l'ISO, cette norme semble moins plébiscitée que SKOS par les acteurs du Web. Cela est probablement dû au fait qu'elle est issue d'un autre consortium que le W3C : le consortium indépendant TopicMaps.Org[31]. Cela explique probablement aussi que la norme XTM ne soit pas encore interopérable avec d'autres formats sémantiques. Cependant, une réflexion est en cours pour qu'elle concorde avec « *la recommandation RDF du W3C, et […] OWL* »[32]. La modélisation propre aux cartes de thèmes ne peut donc pas passer à l'échelle du Web sémantique, à la différence des autres SOCs via SKOS. En conséquence, les cartes de thèmes dont l'objectif est de définir plusieurs vues d'un ensemble d'informations, sont « *Web opérables* » grâce à XMT, mais pas encore interopérables. De toutes les façons, ce modèle n'a pas été conçu pour faire des inférences sans intervention humaine (paradigme du Web cognitivement sémantique). En contrepartie, et comme pour SKOS, l'appropriation de ces cartes par un internaute est plus aisée par rapport à celles des ontologies informatiques [19, CAUSSANEL *et al. ;* 20, PEPPER].

[30] < http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/#Dublin_Core >
[31] < http://www.topicmaps.org >
[32] < http://fr.wikipedia.org/wiki/Cartes_topiques > et < http://www.w3.org/TR/rdftm-survey >

L'observation de ces « homologies » conduit à dire ceci, comme Bruno Menon lorsque celui-ci traduit Gail Hodge [6, MENON p. 27] : les ontologies peuvent servir de cadre général pour structurer les taxinomies, les terminologies comme les thésaurus et même les catalogues de ressources documentaires. Les ontologies partagent certaines caractéristiques avec les outils familiers des professionnels de l'I&D auxquels elles sont parfois assimilées à tort. Toujours selon Bruno Menon, langages documentaires et ontologies informatiques jouent un rôle similaire de contrôle. Pour ce qui concerne les langages documentaires, ils servent à contrôler l'usage du langage naturel pour l'indexation des ressources dans les systèmes d'information bibliographiques. Quant aux ontologies informatiques, elles servent aux machines pour le contrôle des métadonnées. Elles font office alors de cadre de référence pour l'expression des métadonnées. Elles contrôlent l'usage des concepts pour annoter et raisonner, et c'est pour cela qu'elles tendent à être bien plus formelles que les autres SOCs.

3.5 Ontologies et Web sémantique

L'objectif poursuivi par Tim Berners-Lee[33], était que les humains accèdent non seulement aux pages mais aussi surtout aux informations contenues dans ces pages. Le mythe des origines était aussi que les machines exploitent de façon « intelligente » ces informations et les restituent à l'usager de la même façon[34]. Ceci pouvant donner alors l'impression que les machines puissent être douées de sens. Voilà donc pourquoi son concepteur a pu appeler de ses vœux cet avenir du Web, le Web sémantique. En réalité, en guise d'information, les machines sont seulement capables de manipuler des données. Mais il est nécessaire que les informations soient structurées. Cela veut dire, qu'en amont, des humains ont intentionnellement attribué une signification à ces données et l'ont encodée dans un format informatique pour les rendre exploitables par les machines. C'est la raison pour laquelle dès cette date, Tim Berners-Lee parle de données et pas seulement de documents. D'ailleurs aujourd'hui, est-il encore possible de parler de pages quand celles-ci sont générées par calcul et à la volée à partir d'informations, de données inscrites à différents endroits du Web ?

En conséquence, aujourd'hui, l'expression Web de données tend à supplanter l'expression Web sémantique. Il se caractérise comme un environnement où les données sont fortement reliées entre elles. Dans cet environnement de type réseau, une donnée permet d'accéder à d'autres données. Le Web de données, aussi appelé *Linked Data*, est donc la matérialisation des principes du Web sémantique et de l'utilisation grandissante des technologies développées dans cette optique. Quelles sont ces technologies ?

---

[33] Inventeur en 1989 du Web lors de travaux menés au Cern.

[34] BERNERS-LEE T., HANDLER J., LASSILA O. « The Semantic Web : a new form of Web content that is meaningful to computers will unleash a revolution of newpossibilities ». *Scientific American* [En ligne]. 2001. Vol. mai, n°17. Disponible sur :
< http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html >

Les techniques du Web sémantique font de celui-ci une extension du Web car elles reposent elles aussi sur :

- l'hypertexte au travers de l'utilisation d'URI (*Uniform Resource Identifier*),
- le protocole de transfert http,
- les métalangages basés sur XML qui permettent de marquer les données.

Ces technologies comme celles décrites ci-dessous, font l'objet de standards élaborés au sein d'une large communauté regroupée au sein du W3C.

Ainsi, depuis 1997, le W3C développe et maintient le standard *Ressource Description Framework* (RDF) qui est devenu une norme en 2004. Il s'agit à la fois d'un modèle conceptuel et d'une syntaxe pour décrire de façon structurée et relier entre elles tous les types de ressources soit des objets, notions, termes, appelés ici données. Cette syntaxe est toujours composée de trois éléments reliés entre eux par un mécanisme d'identification et de liaison des URI. Chacune des données peut être décrite sous la forme d'une phrase simple « sujet-verbe-complément » ou plus précisément « sujet-prédicat-objet/valeur ». Les données peuvent être le « sujet » de certaines déclarations et tout aussi bien être le « sujet » ou l'« objet » d'autres déclarations par ailleurs. A chaque donnée est attribuée une URI. Les liens entre les données sont typés, c'est-à-dire que la relation est explicitée et décrite selon un formalisme fourni par RDF. Ces liens sont eux aussi identifiés avec des URI. L'ensemble de ces triplets interconnectés compose le *Giant Global Graph* du Web.

Plus que de permettre la navigation dans ce graphe, ces liens ont vocation à permettre aux machines de réaliser des raisonnements. Le *mapping* vise l'interopérabilité des données issues de diverses sources et stockées à différents endroits. Ainsi, l'interopérabilité est assurée par des équivalences conjointement véhiculées par des URI et la syntaxe RDF. L'interopérabilité est grandissante car le standard Ressource Description Framework (RDF)[35] commence à être partagé par de plus en plus de communautés. Ces communautés, en produisant leurs données et en les mettant à disposition en RDF, permettent à celles-ci de ne plus être enfermées dans des silos de données qui leur sont propres mais d'être présentes dans le Web par le biais des standards. Il est alors possible à d'autres communautés d'y avoir accès et de s'y relier bien qu'elles puissent être nativement créées dans des formats hétérogènes.

Pour permettre l'exploitation sémantique de ces données reliées entre elles, il est fait appel à d'autres modèles : les ontologies. En effet, il ne suffit pas seulement de relier les données pour permettre aux machines d'en exploiter les possibles significations. Il faut les contextualiser davantage. Pour rappel, une ontologie informatique est une « *spécification formelle et explicite d'une conceptualisation partagée* ». En rattachant les données à des catégories définies de façon précise, de même qu'en définissant les types de relations entre elles et en ajoutant aussi des attributs, les

[35] < http://www.w3.org/RDF >

ontologies sont donc une opportunité pour exploiter la masse de données qui, progressivement, sont reliées entre elles grâce à RDF. Pour cela, le W3C a permis l'homogénéisation des langages formels utilisés pour exprimer la connaissance avec l'élaboration du langage OWL (cf. 2.3.2 OWL dans l'architecture du Web sémantique). Avec un formalisme standard et un champ de données de plus en plus vaste, le Web de données offre aux ontologies l'opportunité de se développer davantage. Et en contrepartie, le Web profite de l'étiquetage des données par des ontologies. En effet, grâce à elles, les algorithmes peuvent encore davantage trier, filtrer, agréger entre les données reliées et encodées selon ces standards. Ce typage des liens et des données, quelles qu'elles soient, renvoie donc à la notion de métadonnées. Autrement dit, il s'agit de données pour qualifier et annoter les données. Elles décrivent, expliquent la nature, le type d'information porté par la donnée (juridique, administrative, technique). Les métadonnées peuvent donc servir à trouver, à utiliser et à gérer l'information. C'est la raison pour laquelle, A. Isaac qualifie le Web de données de Web de (méta-)données [48, ISAAC diapo. 14].

Le Web sémantique, principalement grâce à l'interopérabilité syntaxique qu'il met en œuvre fait que certaines ontologies se répandent dans le Web. C'est le cas de plusieurs petites ontologies qualifiées de virales comme FOAF (*Friend of a friend*) car elles sont réutilisées par de nombreuses communautés. Lorsqu'il s'agit de décrire les données d'un domaine en particulier, les ontologies informatiques sont en concurrence ou se complètent car plusieurs vues sur un domaine peuvent s'entremêler. Cela est dû au fait que des communautés proches peuvent manipuler les mêmes données mais avec des finalités différentes. Il en est ainsi des éditeurs et des bibliothécaires qui manipulent réciproquement des données bibliographiques. Ainsi, dans ce domaine, différents schémas de métadonnées cohabitent et sont parfois combinés (DC, Bibo, FRBR)[36]. La documentation en ligne du projet Data.bnf.fr pour l'exposition des données de la BnF dans le Web permet de l'appréhender[37]. Le temps où éventuellement une ontologie universelle a pu être envisagée semble révolu au regard par exemple des ontologies recensées par le catalogue LOV. Mais ce répertoire montre que les schémas de métadonnées s'échangent et se combinent tout autant que les données elles-mêmes. Dans le contexte mondial du Web, l'existence d'un seul langage commun est impossible et il serait dommageable qu'il n'existe qu'une seule langue. Mais, à l'extrême il serait inopportun que chaque communauté élabore son schéma en faisant abstraction de ceux préexistants et qui représenterait en partie un intérêt pour elle. Les ontologies sont comme des groupes linguistiques ou sociaux qui fonctionnent tout à la fois par distinction, mimétisme et assimilation et apportent de la richesse aux données ainsi contextualisées. Ce n'est donc pas forcément sans heurts ou tout du moins sans négociation. [54, BERMÈS *et al.* ; 57, GANDON *et al.* ; 83-84, RAÏS]

---

[36] Dublin Core < http://dublincore.org/documents/dcmi-terms >
Bibliographic Ontology < http://bibliontology.com >
Functional Requirements for Bibliographic Records
< http://metadataregistry.org/schema/show/id/14.html >
[37] < http://data.bnf.fr/semanticweb >

Les initiatives qui s'appuient sur le projet universitaire DBpédia[38] en sont une illustration. L'ontologie sous-jacente de l'encyclopédie collaborative Wikipédia a été formalisée. Cela a permis de mettre à disposition en RDF l'extraction des données de Wikipédia et de favoriser l'émergence du Web de données. Ainsi, plusieurs institutions comme la BBC ou des bibliothèques et des musées nationaux ont pu enrichir le Web et s'enrichir mutuellement sans pour autant que l'organisation du web se réduise à la représentation sémantique de Wikipédia. Plusieurs ontologies ou schémas de métadonnées cohabitent et se combinent.

Que se passe-t-il pour les organisations et institutions engagées dans l'ouverture de leur patrimoine informationnel et qui utilisent le jeu de données issu de Wikipédia ? Leurs données gagnent en visibilité par la mise en relation avec ce jeu de données. L'avantage de ce jeu est d'avoir une couverture de nature encyclopédique et d'être suffisamment massif pour que beaucoup de communautés et d'institutions y trouvent de quoi se relier. Ainsi les données d'autorité de la BnF sont enrichies de données biographiques issues de Wikipédia car elles sont reliées à DBpédia. Par ailleurs, elles sont enrichies par celles des homologues et partenaires de la Bnf exposant aussi leurs données en RDF pour permettre leur mise en correspondance. L'ensemble de ces données sont distribuées via data.bnf.fr. Data.bnf.fr n'est donc pas un nouveau silo ou une nouvelle collection fermée où des données étrangères à la BnF ont été dupliquées. Ce service expose et distribue des données issues d'horizons différents qui, à nouveau, peuvent être agrégées et exploitées pour divers usages : navigation enrichie, informations contextuelles... [54, BERMÈS *et al.* ; 57, GANDON *et al.*]

Ces technologies sont en majorité invisibles pour les internautes, si ce n'est par de nouveaux usages auxquels il n'est pas évident de les associer. Il est possible de s'interroger sur la neutralité ou plutôt sur l'intentionnalité de la technologie [38, DECLERCK *et al.*]. Peut-être cela explique-t-il les craintes quant à une représentation du monde sur le Web qui soit sinon unique ou du moins hégémonique ? Mais plus qu'au niveau des ontologies, les craintes à avoir pourraient plutôt être liées aux jeux de données et au caractère de ces données. Pour reprendre l'exemple de DBpédia, seules les données des pages anglophones ont été transformées en RDF. Ceci excluant de fait des sujets non partagés par la communauté des rédacteurs anglophones des pages de l'encyclopédie. Pour disposer d'un jeu de données francophones et de données qui n'existent pas dans le Wikipédia anglophone, le projet SemanticPedia a été lancé en 2012.[39]

[38] < http://dbpedia.org > et < http://fr.dbpedia.org >
[39] < http://www.semanticpedia.org >

Pour l'heure, quels jeux sont exposés sur le Web et quels sont ceux qui en deviennent des points d'ancrages ou autrement dit des *hub*[40] ? Lesquels sont privatisés ou rendus inaccessibles aux usagers alors qu'ils en sont soit eux-mêmes les fournisseurs, ou sont en position légitime d'en demander l'accès ?[41] Il s'agit alors de savoir quelle confiance il est possible d'accorder envers les institutions émettrices de données et de métadonnées. Il s'agit aussi de dénoncer les risques « d'enclosures » et autres « jardins fermés » qui empêchent la réalisation du Web de données[42]. En contrecarrant les principes du Web sémantique et son architecture volontairement distribuée, ils menacent également la création de nouveaux services. Il s'agit donc de question politique et aussi d'éthique si l'on pense aussi aux données personnelles que ces ontologies sont permettent de manipuler.

Le Web de données est favorable aux ontologies grâce aux données accessibles en de plus en plus grande quantité et grâce à l'utilisation du langage OWL. Plus il y a d'expériences de réalisation d'ontologies, plus y a de perspectives d'avancées de l'ingénierie ontologique. Il est possible de voir alors apparaître de nouvelles méthodologies et techniques pour créer mais aussi maintenir et faire évoluer les ontologies. La question des méthodes d'alignement n'est pas sans reposer des questions d'ordre philosophiques si on s'oriente vers un alignement avec une ontologie de haut niveau ou une ontologie noyau ou bien par lien direct ou par une structure pivot dans le cas des vocabulaires de valeurs. Elles seront aussi à l'origine de l'amélioration des services proposés par les éditeurs de contenus sur Internet. Par exemple, en fonction du contexte d'interaction si celui-ci a été conceptualisé, c'est-à-dire décrit précisément et avec la syntaxe standard et les ontologies répandues et partagées dans le Web, les services qui pourront être proposés à l'usager seront davantage adaptés à ses besoins [40, GANDON].

Mais ces opportunités offertes par le Web sémantique et les défis qu'ils posent ne doivent pas faire non plus oublier certains des freins méthodologiques et logiciels auxquelles les ontologies ont déjà été confrontées par le passé. C'est l'occasion de rappeler une nouvelle fois que les ontologies étaient présentes dès les recherches sur les systèmes experts et qu'elles étaient aussi connues par les professionnels de l'I&D pour la gestion des dossiers d'activités par exemple. Aujourd'hui, la puissance algorithmique, l'interopérabilité syntaxique du Web de données et les données qui y sont en partie présentes leur donnent de nouvelles opportunités pour sortir des laboratoires et des systèmes d'information des grandes entreprises. Mais si elles se heurtent à des problèmes encore méthodologiques, le défi est aussi logiciel comme par le passé.

---

[40] Des points d'interconnexion. http://fr.wiktionary.org/wiki/hub

[41] Loi n° 78-753 du 17 juillet 1978 relative à la liberté d'accès aux documents administratifs et à la réutilisation des informations publiques
http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068643&dateTexte=vig
Décret n° 2005-1755 du 30 décembre 2005 relatif à la liberté d'accès aux documents administratifs et à la réutilisation des informations publiques, pris pour l'application de la loi n° 78-753 du 17 juillet 1978
http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000265304&dateTexte=vig

[42] ERTZSCHEID O. « Documentation haute fréquence ». Affordance.info, ISSN 2260-1856. [En ligne]. 25 mars 2012. Disponible sur :
< http://affordance.typepad.com/mon_weblog/2012/03/documentation-haute-frequence.html >

Les outils manquent encore pour les manipuler par d'autres professionnels que des informaticiens. Par exemple les éditeurs de portails documentaires ou de systèmes intégrés de gestion de bibliothèques (SIGB) ne permettent pas facilement d'expliciter la connaissance des langages documentaires et d'exploiter les relations entre les concepts dans les interfaces destinées aux usagers finaux. Cela devient encore plus crucial aussi à l'heure où les données gérées par les bibliothèques et les centres de documentation sont encore plus hétérogènes et issues de sources diverses. La question des outils est également importante à l'heure où les professionnels ont à faire le choix entre plusieurs *Electronic ressources management systems* (ERMS) du marché ou à co-construire les bases de connaissances sur lesquelles ils devront s'appuyer.

4 Le point sur les aspects économiques

4.1 Quel est le coût des ontologies ?

Dans le cadre académique, presque toute publication scientifique s'intéressant à la construction d'ontologies évoque l'importance du coût de conception sans l'expliciter davantage. Cette cherté semble évidente à tous les auteurs sans qu'il leur semble nécessaire d'en détailler plus avant les raisons. Pour autant, ce coût n'apparaît pas être un argument suffisant pour renoncer à l'usage des ontologies. D'ailleurs l'émergence du Web de de données renforce ce pan de l'intelligence artificielle (IA) qui se consacre à l'acquisition des connaissances. Appelée ingénierie des connaissances (IC), celle-ci cherche à développer des méthodes et des outils pour contourner ce coût.

4.1.1 Le coût de construction et le coût de la réutilisation d'ontologies existantes

Il semble que ce coût se compte surtout en temps nécessaire à la réalisation d'ontologies. Il est lié aux difficultés de solliciter des experts peu disponibles. Les rencontres trop espacées ne permettraient pas de maintenir l'implication des experts et menaceraient la réalisation effective d'une ontologie de domaine. Comme abordé précédemment, le TAL est principalement exploité pour parvenir à pallier cette difficulté. En effet, dans le cas de la création d'ontologie de domaine, les écrits d'une communauté permettent de constituer des corpus représentatifs des connaissances partagées par celle-ci. Ils favorisent d'accéder à la langue qui, malgré toutes les ambiguïtés qu'elle peut véhiculer, est le principal vecteur de la transmission des connaissances. Le TALN est d'ailleurs l'une des solutions retenues et combinée avec d'autres dans le cas de la construction de l'ontologie étudiée au cours du stage (cf. 2.4.2 Ingénierie des connaissances (IC) et traitement automatique des langues (TAL) et 7.2.1 Combiner les méthodes ascendantes et descendantes).

La réutilisation d'autres ontologies participe aussi de cette ingénierie des connaissances et de cette tentative de réduction des coûts. Il peut s'agir dans une moindre mesure de l'utilisation de vocabulaires contrôlés existants pour un domaine. Mais la réutilisation a elle-même un coût pour plusieurs raisons. Ensuite les critères de leur évaluation sont en soi aussi un domaine en développement actuellement. A cela il faut ajouter le presque inévitable traitement des ontologies réutilisées. Ceci pour les abstraire de leur conceptualisation d'origine afin qu'elles conviennent à la tâche qui leur est assignée dans un nouveau contexte. De ces difficultés, et donc aussi de ce coût de réutilisation, se dégagent des préceptes de construction ciblant la modularisation des ontologies.

4.1.2 Le coût de la conception et le coût de maintenance

GANDON *et al.* amènent à penser qu'il faut envisager la conception et la réalisation d'ontologies comme tout autre projet lorsqu'ils écrivent :

> « Oui, créer une ontologie peut être coûteux. Une ontologie est un composant logiciel et, à ce titre, on peut être amené à l'inclure dans un cycle de gestion de projet logiciel. Mais ce qu'il est important de reconnaître avant tout, c'est qu'une ontologie n'est pas nécessairement très volumineuse ni nécessairement formelle. La couverture, la spécificité, la granularité et la formalité d'une ontologie sont des caractéristiques variables d'une ontologie à l'autre en fonction des buts visés et du domaine d'application. » [57, GANDON *et al.* p. 193]

Ainsi, il est possible de lister des types de coûts comme les salaires des membres de l'équipe de conception et le montant des charges. Il faut aussi penser aux honoraires des experts et des consultants s'il y a lieu pour le recueil de besoins et l'assistance à l'ingénierie de projet. A cela il faut ajouter les coûts engendrés par le recours à des prestataires pour le développement, l'administration et le graphisme de l'interface. En effet, la réalisation d'une ontologie est rarement l'objet unique d'un projet. Comme l'a mentionné Sylvie Dalbin, consultante en organisation et ingénierie documentaire, au cours d'un entretien suivi d'échanges durant la période de stage, « *le travail de modélisation dépasse le référentiel sémantique* ». Sylvie Dalbin rappelle ainsi que l'intérêt doit être porté plutôt sur « *l'économie globale du projet [qui] dépend de chaque projet, de chaque contexte* ». Ainsi, bien que spécialiste des vocabulaires, elle souligne : « *je suis exclusivement dans des environnements où les calculs ne sont pas centrés sur le vocabulaire, celui-ci n'étant qu'un des composants d'un projet. L'évaluation porte sur le projet lui-même. Mais la formalisation d'un référentiel peut être réutilisée dans d'autres projets, ce qui revient à en diminuer son coût.* »

Pour poursuivre cette liste de coûts de conception, il est possible d'ajouter les frais de déplacement pour l'équipe de travail et pour les membres du comité d'experts. Il peut y avoir des coûts d'extraction, de traitement de bases de données et de TAL. L'équipement informatique et logiciel est aussi à prendre en considération. Le coût, est donc difficile à estimer car propre à chaque projet et à chaque méthode utilisée. Il peut revenir au plus à environ 100 000 euros dans le cadre de la recherche publique. Cette estimation est basée sur le principe d'une ontologie créée durant les trois années nécessaires à la réalisation d'une thèse de doctorat. En sachant qu'un doctorant a d'autres activités que sa seule recherche et que la construction d'une ontologie. La recherche sur les ontologies de haut niveau et sur les ontologies en général devrait abaisser ce coût dans le futur.

Les coûts liés à la gestion et l'évolution à plus ou moins long terme de produits terminologiques sont eux aussi à envisager. Dans ce cas comme dans le cas d'utilisation d'outil gratuit, le temps d'apprentissage par les équipes ne doit pas être négligé. « *S'il n'y a pas de personnel en interne, il y a embauche* » rappelle aussi encore Sylvie Dalbin. Prévoir un budget pour absorber les coûts de mise à jour de l'ontologie et de son outil de gestion lui-même est donc également nécessaire.

Enfin des frais de promotion, de publication et autres frais divers sont à envisager. Pour résumer cet aspect économique, il est possible à nouveau de reprendre les propos de Sylvie Dalbin en réponse à la question des coûts et de leurs critères d'évaluation : « *Comme n'importe quel projet : charges et délais, nature (très variée) et niveau de compétences, reprise des systèmes précédents (aujourd'hui le plus couteux).* ».

Par ailleurs, selon les besoins applicatifs, une ontologie est parfois peu volumineuse, son expressivité limitée, son formalisme peu contraignant. Le coût engendré sera alors peut-être plutôt dû à l'extraction des données antérieurement rassemblées selon un autre modèle. L'opération la plus coûteuse étant peut-être alors de les transformer pour venir peupler cette nouvelle ontologie.

Ce qui a pu être appréhendé de loin également au cours du stage et au travers de quelques lectures, c'est une forme de salariat. Dans le cas de projets innovants, le recrutement d'un doctorant avec un contrat de thèse sous le régime Cifre[43] permet aux entreprises d'accéder à ces technologies.

4.1.3 Le coût symbolique et cognitif

Quel est le prix à payer pour les usagers ? Quel rapport coûts/bénéfices pour les usagers lorsqu'il s'agit de navigation dans les interfaces informatiques, de recherche dans les bases de données ? Quels bénéfices pour les uns et les autres lorsqu'il s'agit aussi d'une contextualisation de l'offre ? Quels coûts pour ces suggestions issues de calculs basés sur des modélisations exploitant des données de profils tout en s'appuyant sur les métadonnées des objets proposés dans les interfaces des applications sémantiques[44] ? [51, KEMBELLEC *et al.*] Entre efficacité, pertinence et sérendipité permises par la « personnalisation » du fait des différents types de traces laissées par les usagers [49, MERZEAU] et qui viennent « peupler » les modèles ontologiques, quel est le coût cognitif pour ces derniers ? Quelle liberté cognitive pour l'usager du « *Web computationnellement sémantique* » [19, CAUSSANEL]. L'ontologie est un objet social car c'est le produit d'une intention et c'est à la fois un sujet d'influence. Partant de cet état de fait, « *la recommandation algorithmique va-t-elle conduire à une ouverture ou à un entonnoir des pratiques des usagers ?* » [51, KEMBELLEC *et al.* p. 20]

Les interlocuteurs rencontrés au cours du stage, académiques ou interlocuteurs appartenant au secteur industriel et des services, ne semblent pas aborder le sujet sous cet angle. Il semble en être de même pour les auteurs traitant des technologies et du Web sémantique, et là encore quelle que soit leur position. Mais il est sûr que les pratiques changent, actualisent les besoins et les demandes que formulent les usagers.

---

[43] « En France, une convention industrielle de formation par la recherche (CIFRE) est un dispositif de financement de thèse qui aide les entreprises pour le recrutement de jeune chercheur-doctorant. »
< http://fr.wikipedia.org/wiki/Convention_industrielle_de_formation_par_la_recherche >

[44] *Collaborative models* ou système de recommandation par filtrage (approche *People-to-people correlation), content-based* ou moteur de recommandation orienté données ou métadonnées et *Content Based Filtering* (CBF) ou filtrage par le contenu qui mixe les deux : métadonnées sur les contenus couplées aux profils des utilisateurs.

4.1.4 Le coût stratégique et politique

Il est vrai que le parti pris de rendre payant des schémas de métadonnées contredirait les objectifs de certaines organisations. Effectivement, certaines, comme la majeure partie des grandes institutions, souhaitent diffuser leur modèle ontologique afin que leurs données elles-mêmes soient plus visibles sur le Web. Il en est de même pour les communautés scientifiques dont le souhait est de favoriser l'échange et l'accès aux données. Pour une frange d'informaticiens héritiers des idées des pionniers de l'Internet et du Web, la culture de l'échange s'accorde avec celle du libre accès à l'information des bibliothécaires. En effet, ces derniers sont moteurs de l'exposition des données patrimoniales et culturelles sur le Web. Mais cela n'est pas aussi sans une certaine compétition pour la définition de modèles ontologiques. Le point crucial étant surtout l'adoption la plus large des formats qui en découlent pour garantir l'échange effectif de données et l'exposition de celles-ci. Tout en étant une question d'éthique, c'est donc aussi une question de pouvoir qui passe par la présence et la visibilité dans l'écosystème du Web. Ainsi à plusieurs reprises [57, GANDON *et al.* p. XII] alertent sur l'importance *« des métadonnées pour annoter les ressources du Web et exploiter la sémantique des schémas de ces annotations pour les traiter avec intelligence. Il est possible que, demain, ceux qui contrôlent les métadonnées contrôlent le web, et des initiatives comme schema.org confirment cette intuition »*.

Ces propos et l'argument soulevé montrent qu'en effet, c'est aux travers des usages actuels dans le cadre du LOD que ces technologies sémantiques semblent être validées. Ces usages ne sont plus seulement à la fois ceux des grandes institutions culturelles et ceux des sites marchands. Ainsi, ces investissements, certes coûteux, commencent à porter leurs fruits. La preuve en est par la nouvelle orientation prise par les moteurs de recherche généralistes les plus populaires. Après avoir utilisé les métadonnées puis décidé de les ignorer, ces géants du Web manifestent de nouveau leur intérêt pour celles-ci. Comme le rappellent les auteurs de l'extrait précité, et comme l'explique Emmanuelle Bermès [54, BERMÈS *et al.*], depuis 2011 les géants du Web sont réunis pour définir l'ontologie Schema.org[45]. Dans ce même mouvement, l'achat par *Google* de l'entreprise *Metaweb* et, par ce fait, de l'encyclopédie *Freebase* pour mettre en place *Google Knowledge Graph*[46] ainsi que l'annonce par *Facebook* du *Graph search*[47] démontrent bien l'intérêt grandissant pour la structuration, pour la sémantisation des données et pour leurs potentialités[48]. Même si bien sûr, le témoignage de Sylvie Dalbin n'a pas été sans rappeler que des technologies reposant sur la modélisation sont plébiscitées depuis un certain temps déjà dans les SI d'entreprises. Mais l'introduction des technologies du web sémantique arrivant à maturité, la réalité change aussi pour

[45] Google, Bing de Microsoft et Yahoo! : < http://schema.org >
[46] < http://fr.wikipedia.org/wiki/Knowledge_Graph > ;
[47] < http://fr.wikipedia.org/wiki/Facebook_Graph_Search >
[48] « Le Knowledge Graph de Google ferait baisser le trafic de Wikipedia ». In : *Actualité Abondance* [En ligne]. [s.l.] : [s.n.], [s.d.]. Disponible sur :
< http://www.abondance.com/actualites/20140114-13550-le-knowledge-graph-de-google-ferait-baisser-le-trafic-de-wikipedia.html >

ces SI qualifiés « *d'intrawebs* » par [57, GANDON *et al.*]. Il est plus facile de formaliser et donc de rendre calculables par les machines les modélisations de connaissances métiers ou de retours d'expérience. Par exemple encore, il est aussi maintenant possible de les affranchir de progiciels de gestion propriétaires comme *Lotus Notes*. Par ailleurs, l'accueil au laboratoire Limics Inserm durant le stage a été l'occasion de prendre encore davantage conscience de l'intégration des objets ontologiques dans les applications médicales.

4.2 Qui vit des ontologies ?

4.2.1 Un rappel des usages

Comme cela a été abordé précédemment, les ontologies sont au cœur d'applications en pleine expansion du fait de la standardisation et de l'adoption progressive des standards du Web sémantique. Ces applications sont nombreuses. Elles peuvent varier et se cumuler selon les objectifs définis :

- la gestion de connaissances d'un domaine particulier et le partage de ces connaissances ;
- l'interopérabilité entre les systèmes avec l'amélioration de la communication entre les machines en plus de celles entre les hommes et les machines ;
- le traitement et la recherche d'information.

4.2.2 Les acteurs et leur domaine d'activité : l'informatique et le conseil

> « Pour Ivan Herman, responsable des activités liées au web sémantique au W3C [...] les projets de recherche, initiés au début des années 2000, sont aujourd'hui sortis des labos. Il existe maintenant une communauté de petites entreprises innovantes, mais également de grands acteurs tels qu'Oracle ou IBM qui se sont lancés dans l'aventure. Des sociétés, des institutions commencent à utiliser ces technologies : la BBC, Chevron, Novartis… En parallèle, un mouvement est apparu pour créer et lier les données entre elles : *Linked Data* ou le web des données. […] Pour Fabien Gandon, chercheur à l'Inria, on observe depuis 2007 une montée en puissance de la présence industrielle. La France est, pour une fois, bien positionnée, avec des acteurs comme Logilab, Semsoft, Antidot, etc. » [49]

Certaines grandes entreprises semblent donc percevoir l'intérêt des technologies sémantiques pour leurs activités. D'autres entreprises ou jeunes pousses en font semble-t-il le cœur de leur activité d'innovation. Cependant, une recherche rapide sur le site *Societe.com* montre bien que les activités liées aux ontologies ne constituent pas un seul et unique secteur. En effet, en interrogeant par les noms de sociétés citées ici par Fabien Gandon et par le nom de celles rencontrées au cours du stage, constat est fait que les intitulés et codes NAF[50] diffèrent d'une société à l'autre :

- Conseil en systèmes et logiciels informatiques (6202A)

---

[49] TRAN P. « Semweb.pro : la France bien positionnée sur le web sémantique ». *01net* [En ligne]. 20 janvier 2011. Disponible sur : http://pro.01net.com/editorial/527170/semweb-pro-la-france-bien-positionnee-sur-le-web-semantique

[50] Nomenclature d'activités française : elle est élaborée par l'Insee (Institut national de la statistique et des études économiques).

- Autres activités informatiques (6209Z)
- Edition de logiciels applicatifs (5829C)
- Conseil en systèmes et logiciels informatiques (6202A)
- Programmation informatique (6201Z)
- Ingénierie, études techniques (7112B)
- Conseil pour les affaires et autres conseils de gestion (7022Z)

En cherchant plus avant, il est possible de constater que l'une appartient au groupe Thales. Une autre a fait l'objet d'un rachat par cette grande société spécialisée dans l'aérospatial, la défense et les technologies de l'information.

4.2.3 Les produits et les services proposés

Pour la plupart, ces sociétés développent des solutions liées à l'analyse sémantique de textes comme des logiciels de *Text mining,* c'est-à-dire de fouille de textes, dits aussi d'extraction de connaissances et d'annotation de textes. Soit ces solutions font aussi partie de leur gamme, soit elles collaborent fréquemment avec des sociétés qui les développent et les intègrent chez un client commun. Ces clients les sollicitent dans une optique de référencement. Dans ce cas, ces solutions sont souvent mises en liaison avec des techniques rassemblées sous le vocable *Search engine optimization* (SEO) pour techniques d'optimisation pour les moteurs de recherche [65, ECONOCOM-OSIATIS]. Comme l'indique l'observation en 4.1.2, le traitement de l'information par l'amélioration de son indexation et l'amélioration de sa recherche semble être l'application la plus manifeste. Quelle en est la raison ? La réponse est la suivante : cette technologie est un réel atout pour aider à la valorisation des contenus et est donc une source indirecte de rentabilité.

4.2.4 Des besoins et des clients différents

Les clients de ces sociétés qui ont émergé depuis 2007 peuvent être des éditeurs de contenus qui souhaitent mettre en place des annotations normalisées pour guider les contributeurs dans la valorisation de leurs contenus. Par ailleurs, les sociétés clientes peuvent souhaiter améliorer la structuration de leur portail pour proposer une navigation logique et des informations contextuelles selon le public visé. La demande peut être également de réduire le nombre d'échecs des requêtes sur leur portail. Dans ce cas, une modélisation des contenus de leur portail sous la forme d'une ontologie peut être réalisée. Elle est ensuite reliée au vocabulaire d'interrogation des usagers récupéré via les logs du moteur de recherche. Ainsi, les sociétés spécialistes des technologies sémantiques réalisent des modèles en créant un référentiel instancié avec des noms de rubriques, des mots-clés, et pour lequel sont déclarés des liens typés entre les éléments. Et dans ce cas, c'est la méthode ascendante qui est utilisée (cf. 2.4 Les méthodes et les phases de construction). Pour autant, il n'est pas toujours nécessaire de pousser l'expressivité plus loin qu'une taxinomie pour augmenter le rappel sur un portail [46, FRANCART]. Pour produire des raisonnements, il est en revanche très important de pousser la structuration du modèle et d'en faire des bases de connaissance.

4.3 Quelles sont les caractéristiques du modèle économique ?

4.3.1 Un rappel du modèle d'affaire des thésaurus

Au sujet des ontologies, la littérature mentionne souvent comme pratique celle de la réutilisation de ressources déjà existantes. Mais, s'agit-il de réutilisation contre paiement ? Peut-on vendre, peut-on acheter des ontologies comme il est parfois nécessaire d'acheter le droit d'accès aux thésaurus et donc le droit de leur usage pour certains d'entre eux[51] ? Y a-t-il un modèle d'affaire propre à ce type d'objet, de ressources ? Vend-on plutôt les ontologies ou plutôt l'accès aux bases de connaissances que les ontologies participent à rendre opérationnelles ? Ou bien encore vend-on les données qu'elles agrègent ?

Auparavant, les thésaurus étaient édités principalement sur support papier, parfois sur support cédérom. Certaines institutions diffusent encore leur thésaurus sous forme imprimée. Ces éditions sont généralement payantes. Les institutions qui éditent et diffusent leur thésaurus proposent de plus en plus des versions électroniques sous la forme de fichiers téléchargeables. Lorsqu'il s'agit de la version imprimable au format PDF, celle-ci peut tout aussi bien être payante que gratuite. Parfois, des versions électroniques pour exploitation sont aussi proposées. Et dans ce cas, c'est plus un modèle basé sur le service avec un modèle d'abonnement qui est constaté. Plus qu'au coût du fichier en lui-même, le coût correspond alors davantage au fait d'appartenir à un réseau, de participer à sa gouvernance, de bénéficier de mises à jour et de prestations techniques pour l'intégration du thésaurus dans des applications. Mais la diversité des modèles ici exposés ne doit pas faire oublier que le modèle d'affaire le plus répandu est celui qui va de pair avec les réseaux documentaires. Et c'est d'ailleurs principalement dans ce cadre de mutualisation au sein des réseaux qu'ils ont été et sont encore développés. C'est le cas du Thésaurus Santé Publique (TSP) du réseau Banque de données en santé publique (BDSP)[52] développé par l'Ecole des hautes études en santé publique (EHESP) qui diffuse toutes les versions du thésaurus, dont le fichier natif, gratuitement et au-delà même du réseau de ses contributeurs.[53]

[51] Lu sur Twitter : « Olivier Le Deuff @neuromancien 25 sept. 2014 les thésaurus sont comme les sous-vêtements, tout le monde en veut mais personne ne veut les partager #websem14 ».
[52] Page d'information à propos du TSP sur le site de la BDSP :
< http://asp.bdsp.ehesp.fr/Thesaurus >
[53] Voir la liste maintenue à jour des thésaurus accessibles via le Web par Sylvie Dalbin sur le répertoire Open Directory à l'adresse suivante :
< http://www.dmoz.org/World/Français/Références/Thésaurus >

4.3.2 Une technologie encore récente

Au niveau temporel, puisque les origines des ontologies remontent à l'intelligence artificielle, elles peuvent faire les frais des déceptions générées par l'IA. Mais avec le développement des technologies du Web sémantique, elles revivent une deuxième jeunesse. Elles font l'objet de bouillonnements à la croisée des disciplines linguistiques et informatiques. En conséquence, le consensus autour de la définition des ontologies et de leurs finalités est instable. Cela fait bouger les lignes aussi pour les professionnels de l'I&D. Cette technologie est par ces aspects encore une innovation non stabilisée car elle s'inscrit plus ou moins directement dans le Web, qui est contraint lui-même de combiner différentes logiques. Celles-ci sont d'une part la gratuité du fait de l'histoire d'Internet mais aussi les logiques des secteurs marchands et celles, plus hybrides, de l'économie de la connaissance d'autre part. Cette technologie est protéiforme puisque son essence est de spécifier les connaissances pour s'adapter aux besoins et aux finalités qui varient d'une communauté et d'une application à l'autre. De plus, les besoins sont parfois difficiles à définir. Et présumer, à la place des usagers, l'utilisation idéale et optimale de certaines technologies a souvent été une gageure comme le rappelaient les différents intervenants des journées tenues à la BnF en novembre 2014[54] [78, CARMES].

4.3.3 Des standards pour faire du sur-mesure

Les organisations utilisant des ontologies ont tout intérêt à utiliser des standards éprouvés et validés par le W3C. Ils sont plus largement maitrisés par les professionnels des ontologies. Mêmes les sociétés dont le système d'information est un environnement fermé y ont intérêt pour diverses raisons : flexibilité et efficacité des standards, évolution éventuelle vers une entrée dans le Web de données avec une volonté d'interconnecter leur système et leurs données à d'autres. Cela leur permet également de ne pas être enfermées dans un système propriétaire verrouillé et de pouvoir ainsi récupérer et retravailler leurs données. Utiliser les standards est donc aussi surtout un gage de pérennité. Enfin, utiliser les modèles mis à disposition par d'autres communautés avec des licences d'utilisation ouvertes permet de diminuer le coût de modélisation en bénéficiant de modèles déjà validés. Standards et économie basée sur la réutilisation sont donc principalement mis en avant par les promoteurs de cette technologie qui s'inscrit dans le Web sémantique.

[54] BIBLIOTHÈQUE NATIONALE DE FRANCE. « Quelles réutilisations des métadonnées mises à disposition par la BnF ? Jeudi 6 et vendredi 7 novembre 2014 ». [s.l.] : [s.n.], 2014. Disponible sur : http://www.bnf.fr/fr/professionnels/anx_journees_pro_2014/a.jp_2014_rameau_donnees_bnf.html

4.3.4 Des catalogues d'ontologies commerciales versus des entrepôts d'archives ouvertes

Dans l'univers du Web sémantique, l'important est que les modèles ontologiques, ou schémas de données, puissent s'interrelier entre eux sinon il y a de nouveau un risque de reproduction des silos, avec des graphes restreints ne permettant pas de dépasser l'entre soi. Pour favoriser le Web de données, comme ressource incontournable, il existe depuis 2011 le *Linked Open Vocabularies* (LOV). Il recense 475 ontologies au format RDFS ou OWL utilisées et interconnectées dans le Web de données[55]. Par ailleurs, dans un souci de faire avancer la recherche, il existe en guise de place de marché, un catalogue de modèles librement mis à disposition comme *OntologyDesignPatterns.org* déjà évoqué précédemment. Dans le domaine biomédical, les ontologies créées dans le cadre de la recherche publique sont souvent signalées et déposées sur *BioPortal*. Il s'agit d'une ressource mise à disposition par le *National Center for Biomedical Ontology* (NCBO) soutenu par l'*US National Institute of Health* (NIH). Elles sont pour la plupart signalées comme libres de droit. Les fichiers sont consultables et téléchargeables directement. Cette archive contient 402 ontologies contre 379 en août 2014 lors du stage. Les chercheurs hors États-Unis d'Amérique y déposent également les ontologies créées dans le cadre de leurs travaux. Cela traduit un véritable besoin d'outil de signalisation et de partage pour ce type de données. Cependant, bien qu'une liste de diffusion y soit rattachée, les chercheurs des autres pays peuvent-ils réellement participer à la gouvernance de cet outil de partage d'ontologies ? Par ailleurs, le taux de réutilisation de ces ontologies est difficile à mesurer de façon objective car les statistiques fournies sont basées sur des données déclaratives.[56]

Pourquoi cette absence de catalogues commerciaux d'ontologies ? Pour tenter d'expliquer cet état de fait, il est possible de poser la question suivante : quels intérêts les acteurs trouveraient-ils à vendre des ontologies dans l'écosystème actuel ? Tirer des rentes des ontologies conduirait probablement à l'immobilisme et pourrait être défavorable à l'émergence de nouvelles activités liées aux technologies sémantiques. Jusqu'ici, le terme « ontologies » ne semble pas avoir un potentiel mobilisateur suffisant pour en faire un argument commercial en lui-même. Ce terme apparaît peu sur les sites commerciaux des spécialistes en la matière. Même si d'après Fabien Gandon : « *la notion d'ontologie, qui précède largement l'utilisation du mot, ne semble pas prête à disparaître. Au contraire, le spectre d'applications et de domaines s'intéressant aux ontologies ne cesse de s'élargir* » [40, GANDON]. Avec cet élargissement des applications et des domaines, les ontologies devraient continuer à sortir des laboratoires de recherche. Si la gratuité actuellement prônée dans ce secteur s'inscrit dans une logique de confiance et d'ouverture, cette situation n'est peut-être pas définitive.

---

[55] 458 en septembre lors de la remise du Livrable 3- État de l'art : répertoire des ressources termino-ontologiques disponibles dans le domaine. p. 21/34. < http://lov.okfn.org >

[56] Voir également la description Annexe 1- Livrable 3- État de l'art : répertoire des ressources termino-ontologiques disponibles dans le domaine. p. 20/34. < http://bioportal.bioontology.org >

Peut-être sera-t-il constaté leur progressive émancipation d'une certaine culture informaticienne liée aux utopies des technologies de l'Internet ?

La monnaie, comme les ontologies, fait elle aussi office de langage commun et donc d'instrument de confiance. Et aujourd'hui c'est un langage doublement chiffré puisque numérique aussi du fait des écritures informatiques. En l'occurrence, la monnaie est donc un outil ayant pour vertu de simplifier l'échange, celle de biens et de services. Mais dans l'écosystème actuel, la marchandisation des ontologies ne semble par leur donner plus de valeur ou les rendre plus désirables. Au contraire, la dimension privative paraît a priori contreproductive et en contradiction avec la volonté d'étendre les principes du Web sémantique. L'objectif est de parvenir à constituer un Web de données étendu à de nombreux domaines dont les données sont interreliées. Leur tarification serait une barrière artificielle à la propension nativement sociale et virale des technologies et des standards du Web. Pour l'instant, la gratuité enclenche une dynamique vertueuse car le Web de données s'agrandit[57]. Mais pour autant, cette gratuité, avec l'exposition et la mise en relation des données qu'elle permet, favorise-t-elle effectivement l'innovation, la rentabilité qui, elles, vont générer ensuite des profits ? La prédominance de la gratuité signifie-t-elle que les ontologies font l'objet d'un retour sur investissement nul ?

4.3.5 Quel retour sur investissement ?

De très grands acteurs misent sur cette technologie : Air France, Boeing, Thales. Ainsi il s'agit d'entreprises possédant des intrawebs complexes et de taille importante [63, BARBAUX ; 64, DALBIN ; 65-66 ECONOCOM-OSIATIS]. Dans le domaine médical, les ontologies sont plébiscitées pour des applications d'aide à la décision, mais aussi pour catégoriser de façon automatique les dossiers « patient » et les actes médicaux selon des modèles plus riches que la classification internationale des maladies[58]. Par ailleurs, elles sont aussi sollicitées pour la catégorisation des contenus dans les portails marchands ou documentaires.

Pour favoriser le web de données, comme il a précédemment été présenté, mais aussi pour diminuer les coûts de conception, les discours actuels recommandent d'arrêter de multiplier les modèles ontologiques. Cependant, de nouveaux modèles continuent d'être créés pour répondre à des besoins spécifiques. Et s'ils sont créés à partir des modèles existants, le coût n'est donc pas à l'achat mais reporté sur l'effort de transformation. Quant aux bénéfices, ils sont ultérieurs. Ils

---

[57] La version du *LOD cloud diagram* datant d'août 2014 du diagramme réalisé par Chris Bizer et son équipe de l'Universität Mannheim en Allemagne est consultable à l'adresse suivante : < http://data.dws.informatik.uni-mannheim.de/lodcloud/2014 >. En complément, se référer au billet de Thomas Francart sur son blog Sparna : < http://blog.sparna.fr/linked-open-data-cloud-nouvelle-version >

[58] Classification internationale des maladies (CIM) dont l'appellation complète est Classification statistique internationale des maladies et des problèmes de santé connexes ou *International Statistical Classification of Diseases and Related Health Problems* (ICD), est publiée par l'Organisation mondiale de la santé (OMS). Le sigle est en principe suivi du n° de version. Ainsi la littérature dans ce domaine fait référence à la CIM-10. La CIM-11 correspond à la révision qui devrait être publiée en 2015.

peuvent être appréciés en fonction de l'efficacité de l'outil auquel ils sont intégrés par exemple. Lorsqu'il s'agit d'ontologies pour améliorer les performances d'un SI, ces dernières comptent au titre d'actif immatériel de l'entreprise. Les ontologies peuvent dans ce cas représenter un avantage concurrentiel pour ces entreprises qui n'ont alors pas d'intérêt à les revendre. Cependant, cet avantage ne peut persister que si les ontologies, comme les thésaurus, font l'objet d'une mise à jour pour rester cohérentes avec les données traitées par les applications. Par ailleurs, d'un point de vue technique, du fait de leurs caractéristiques spécifiques à un domaine et aux données traitées, les sociétés prestataires participant au développement de ces ontologies ne peuvent présenter un catalogue d'ontologies génériques prêtes à l'emploi. En effet, ces ontologies ne pourraient convenir à différentes organisations, surtout si ces derniers travaillent avec des sources de données et des contraintes différentes. A l'inverse la vente d'un modèle générique n'a pas de sens puisque des standards sont déjà disponibles et largement connus et répandus. Cependant, l'économie a parfois des lois contraires à celles de la technique. Parfois les lois de l'économie mettent en place de la rareté et de la cherté là où pourtant il n'y en a pas du point de vue technique. Et parfois les lois économiques n'obéissent pas au droit, le détourne ou le déforme. Entre modèles ontologiques et modèles économiques, pourrait-il y avoir une opposition des valeurs ?

4.3.6 Un modèle d'affaire basé sur celui du logiciel libre

La recherche de profitabilité économique n'est pas basée sur la vente d'ontologies en tant que produits commerciaux. Les sociétés qui travaillent avec des ontologies ou pour en concevoir ne vendent même pas toujours des solutions logicielles. Ce qu'elles vendent c'est du conseil et du service. Leur savoir-faire repose sur leurs compétences pour architecturer les données et les systèmes d'information. Ainsi, c'est leurs compétences à modéliser, à traduire les modèles de travail existants en modèle d'échange de données comme l'a précisé Nicolas Chauvat, fondateur et directeur général de la société Logilab lors d'une intervention lors d'une journée d'études à la BnF[59]. Dans le contexte du Web sémantique, l'efficience est garantie par la réutilisation des ontologies standards. Elles permettent de réaliser des modèles avec un fort degré de « connectabilité » pour reprendre aussi les mots de Julien Homo lors de la journée Semwep.pro 2014[60]. Il entend par là leur capacité à être reliés à d'autres modèles, d'autres vocabulaires afin de rendre les systèmes et les données interopérables. Le partage est un gain d'économie et d'efficience autant pour celui qui réutilise que pour celui qui donne à réutiliser. A la manière des auteurs des livrables de DataLift, les ontologies sont, comme les autres SOCs, « *des constructions sociales* » [61, VANDENBUSSCHE *et al.* p. 8][61].

---

[59] BIBLIOTHÈQUE NATIONALE DE FRANCE. « Quelles réutilisations des métadonnées mises à disposition par la BnF ? Jeudi 6 et vendredi 7 novembre 2014 ». 2014. Disponible sur : http://www.bnf.fr/fr/professionnels/anx_journees_pro_2014/a.jp_2014_rameau_donnees_bnf.html

[60] Définition consultable sur le site de l'éditeur de dictionnaires et d'encyclopédies Larousse : http://www.larousse.fr/dictionnaires/francais/connectabilité/18283 ; programme de semweb.pro consultable à l'adresse suivante : http://www.semweb.pro/semwebpro-2014-details-pres.html dont quelques-unes des présentations sont accessibles depuis ce lien : https://twitter.com/semwebpro

[61] Les auteurs des livrables de DataLift nomment sciemment les schémas de métadonnées, c'est-à-dire les ontologies, avec le terme « *vocabulaires* » : « *des* ontologies – *appelés ci-après des* vocabulaires » Ce choix terminologique étant expliqué p. 7 du présent document : « *le mot « ontologie » reste un peu intimidant* ».

4.3.7 La valeur n'est peut-être pas seulement dans le moteur ?

Comme cela a été vu précédemment ce qui fait leur intérêt c'est aussi leur degré de « connectabilité ». La dimension des projets dépassent les ontologies et in fine la dimension exclusivement technique. Leur déploiement est peu visible bien qu'elles soient pourtant présentes. Elles sont actives dans les systèmes utilisés tous les jours sous la forme de schéma de métadonnées ou sous la forme de taxinomies de vocabulaires contrôlés. Simplement, elles font partie du back office, elles sont du côté des serveurs. En 2008, l'entrepreneur Alex Iskold, dans le célèbre blog *ReadWriteWeb*, relayé par le site *Internetactu*, considérait que le Web sémantique approchait de la maturité[62]. Évoquant les ontologies sans en prononcer directement le nom, il prévenait bien que « *les utilisateurs ne seront pas sensibles à la technicité des langages RDF ou OWL […] mais bien à des applications concrètes qui sauront répondre à leurs besoins* ». Ce qui compte, parce que cela fait vendre, ce n'est pas la technologie « ontologie », ce sont les applications. Pour l'heure, il semble que l'on attende encore la « *killer app* » (« *killer application* ») qui popularisera le terme « ontologie », si tant est que cette popularité soit réellement attendue par les promoteurs des ontologies. Ceux-ci attendent certainement plus l'ouverture de jeux de données pour faire la démonstration des potentiels d'usage des ontologies et autres technologies sémantiques. Et ceci notamment pour les données de santé où l'accès s'avère délicat pour des questions non pas seulement techniques et économiques mais d'éthique. Car c'est dans les données que résident potentiellement la valeur. Ainsi Nicolas Chauvat, indique que c'est « *l'analyse des données qui produit la véritable valeur ajoutée* »[63]. L'ontologie est l'un des outils qui permet de la révéler. Un an avant Alex Iskold, sur son blog « *Les petites cases* », Gautier Poupeau désignait *DBpédia* comme le « *killer data store pour le Web sémantique* »[64]. Il ne suffit pas que l'ontologie soit un moteur de règles. Il faut pouvoir y insérer des données auxquelles associer de la puissance de calcul. Ainsi applications et données se partagent la valeur dans la nouvelle économie numérique dont les ontologies informatiques sont une des briques logicielles.

Dans cette tentative de synthèse et de prospective économique, il est difficile de dissocier l'analyse des modèles d'affaire de celle des usages. Concernant ces usages, il est possible de dire qu'ils courent encore derrière l'évolution rapide des technologies et sont encore à inventer.[65]

---

[62] GUILLAUT H. « Web sémantique : y aura-t-il une application qui tue ? « ». In : *InternetActu.net* [En ligne]. 21 janvier 2008. Disponible sur : http://www.internetactu.net/2008/01/21/web-semantique-y-aura-t-il-une-application-qui-tue

[63] BEKY A. « Nicolas Chauvat (Logilab) sur CubicWeb et Dataconnexions ». In : *Silicon* [En ligne]. 14 février 2013. Disponible sur : http://www.silicon.fr/nicolas-chauvat-logilab-cubicweb-83522.html

[64] POUPEAU G. « Petite pelote pour pull multicolore ». In : *Les petites cases* [En ligne]. 16 avril 2007. Disponible sur : http://www.lespetitescases.net/petite-pelote-pour-pull-multicolore

[65] Ces usages font l'objet du congrès national FAN2014 des professionnels pour la maîtrise et la valorisation des contenus numériques regroupés au sein de l'association APROGED : < http://www.congres-fan.com/fr/accueil >

4.4 Quels enjeux pour un dispositif socio-technique ?

4.4.1 Faire connaître et faire la preuve de son utilité

Dans l'expression « socio-technique », c'est surtout la notion de « socio » qui semble prédominer. Le fait qu'une ontologie résulte d'un consensus n'en est pas la seule raison.

La première est que « *le mot « ontologie » reste un peu intimidant* » [61, VANDENBUSSCHE *et al.* p. 7]. Les ontologies nécessitent d'être expliquées. Il n'est pas possible de déduire automatiquement la signification de ce terme. La compréhension des potentiels de l'ontologie par les acteurs qui pourraient y trouver des intérêts peut entrainer une plus grande spécification des besoins et attentes, ainsi qu'une plus grande implication pour faire aboutir son implantation. Et la réussite de l'implantation de l'ontologie dépend du processus d'implication de la part de ces acteurs. Certains ontologues parlent d'évangélisation[66]. Mais ce sont les usages qui en définitive valident et « actualisent »[67] une innovation et font se dessiner un modèle économique bien identifiable. Les modèles d'affaire liés ne peuvent se développer s'il n'existe pas déjà quelques applications reposant sur les ontologies. C'est pourquoi on parle de la nécessité de développer des démonstrateurs[68]. Car il faut pouvoir prouver que leur potentiel n'est pas seulement théorique. Ainsi, la preuve de son utilité et donc de l'utilité du temps investi à concevoir des ontologies informatiques et à transformer les architectures en place est décisive pour convaincre. Mais, la preuve peut-elle suffire quand il y a tout un historique ? Suffit-elle quand des outils reposent sur des architectures d'information propres et des habitudes de travail déjà bien ancrées ? Sont-elles assez flexibles pour répondre à différentes pratiques de travail ?

---

[66] Cf. Annexe 4 - Entretien avec Éric Dagiral à propos de l'ontologie pour Orphanet

[67] CARMES M. *Introduction générale aux enseignements du Titre 1, cours Titre 1 INTD*. 2013.
CARMES M. *Introduction à l'analyse des controverses, cours Titre 1 INTD*. 2013
CARMES M. *Introduction au management des connaissances de l'organisation : perspectives – problèmes, cours Titre 1 INTD*. 2014

[68] Le calculateur du domaine public français [En ligne]. *Cblog, Le labo du numérique*. Disponible sur :
< http://cblog.culture.fr/projet/2013/11/08/un-calculateur-du-domaine-public-francais >

4.4.2 Soigner les outils pour améliorer l'expérience

Quand bien même la technologie est efficace du point de vue technique, l'expérience n'est pas toujours perçue comme intelligente et plus simple. La valeur accordée aux dispositifs sous-tendus par des ontologies ne porte pas seulement sur son efficacité technique mais sur la qualité de l'expérience d'utilisation proposée. Ainsi, les problèmes d'efficience et autres problèmes de performance desservent souvent les nouvelles applications. La faiblesse de l'ergonomie fonctionnelle et l'esthétique assez pauvre des interfaces graphiques peuvent jouer en défaveur de technologies pourtant très intéressantes. La sous-évaluation des usages potentiels entraine ces lacunes qui découragent l'usage. Dans la mesure du possible, un point fort devrait aussi être apporté à ces aspects pour les prototypes ou du moins pour les démonstrateurs[69]. La dimension projet est importante pour recueillir à la fois les besoins et les retours lors de tests utilisateurs mais aussi pour procéder à des observations et des évaluations des usages. Sinon il y a un risque pour ces technologies comme pour tant d'autres que leur soit reproché de ne pas être adaptées, et que cela soit aux utilisateurs de s'y adapter. L'une des conséquences est aussi qu'en définitive, la masse critique d'utilisateurs ou de données disponibles pour en révéler le potentiel au sein de certaines communautés ne soit pas atteinte. Des voies d'harmonisation restent encore à trouver entre technologies avancées et usages.

4.4.3 Animer pour faire aboutir et pérenniser les ontologies

En conséquence, un travail d'accompagnement est à développer. Les ontologies gagneraient à ce que des « *infomédiaires* »[70] soient à l'interface des différentes professions en présence. Cela permettrait de recueillir non seulement les besoins mais aussi la réception des dispositifs pour améliorer l'ergonomie des fonctionnalités offertes. Et il y a surtout besoin d'animer les interactions entre les différents acteurs, dont certains seront amenés à maintenir à jour l'ontologie. Seule leur collaboration effective permet l'aboutissement des projets de conception et le maintien par la mise à jour des référentiels qui en découlent. Sans animation, sans création et activation des liens du réseau des experts, informaticiens et techniciens ainsi rassemblés, le réseau risque de mourir. Les ontologies qui doivent servir de langage commun risquent de ne pas aboutir ou de devenir obsolètes. Elles risquent alors d'être inutilisées car inutilisables, et cela sans avoir eu peut-être même le temps de faire leurs preuves.

---

[69] Propos entendus lors de la journée Semweb.pro2014

[70] « mot valise constitué de information et intermédiaire et qui désigne un ensemble de professions gravitant autour des fonctions de documentation et de veille, de knowledge et records management. » d'après CHAUDIRON S. « La notion de médiation en information et communication ». In : *Laboratoire Geriico - Groupe d'études et de recherche interdisciplinaire en information et communication* [En ligne]. [s.l.] : [s.n.], [s.d.]. Disponible sur : http://geriico.recherche.univ-lille3.fr/index.php?page=annee-2011-2012

4.4.4 Développer un argumentaire

Animation donc, mais peut-être aussi que des stratégies de communication et de traduction devraient être mises en place à l'exemple des différents récits et stratégies d'enrôlement qui se succèdent par rapport aux différentes technologies informatiques. Ainsi les acteurs économiques, qui perçoivent les potentiels de rentabilité et de contrôle des données développent des mythes technologiques comme autant de stratégies d'intéressement. Actuellement il s'agit de rendre les infrastructures et les villes plus intelligentes, plus économes en énergie grâce aux capteurs et objets connectés. A l'échelle des systèmes d'information dont c'est plus l'objet de cette étude en ingénierie documentaire, l'argument est plutôt celui du gain de temps et de l'efficacité que la grandeur « intelligence ». D'autres grandeurs s'opposent à ces récits comme celle du partage, la « tradition » de l'accès gratuit et de l'expérimentation. Cependant dans le secteur de la recherche académique, le transfert de technologie est un moyen de valorisation et de financement non négligeable.[71]

[71] CARMES M. *Introduction à l'analyse des controverses, cours Titre 1 INTD*. 2013
CARMES M. *Changement organisationnel et TIC : introduction à la sociologie des usages et synthèse sur les transformations technico-organisationnelles, cours Titre 1 INTD*. 2014

Deuxième partie

La généalogie d'une ontologie et les enjeux associés

5 L'historique et le contexte du projet

La réflexion par rapport aux ontologies et la question des motivations qui peuvent conduire à l'élaboration d'une ontologie de domaine a été menée dans un contexte particulier, celui d'un programme de recherche multidisciplinaire. Dans ce cadre une plateforme répondant à un premier besoin de stockage des documents et des publications scientifiques issues de ce programme a été mise en place.

5.1 Un programme de recherche en toxicologie nucléaire

Le programme scientifique pluridisciplinaire en toxicologie nucléaire a été lancé en 2001 par la Direction des sciences du vivant (DSV) du Commissariat à l'énergie atomique et aux énergies alternatives (CEA) pour une durée de 5 ans. Mariant les compétences des médecins, des biologistes, des chimistes, des physiciens des différents pôles du CEA, il s'ouvre en 2004 aux chercheurs de trois autres organismes de recherche français :

- le Centre national de la recherche scientifique est un organisme public de recherche (CNRS) placé sous la tutelle du Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche ;
- l'Institut national de recherche en agronomie (INRA) placé sous la double tutelle du ministère chargé de la Recherche et du ministère chargé de l'Agriculture ;
- l'Institut national de la santé et de la recherche médicale (INSERM) placé sous la double tutelle du ministère de la Santé et du ministère de la Recherche.

Le Programme Transversal Toxicologie Nucléaire vise à identifier les mécanismes moléculaires et cellulaires des réponses aux éléments chimiques ayant des applications dans le domaine du nucléaire. Il vise également à déterminer des stratégies de défense aux différents niveaux d'organisation du vivant, « de la bactérie à l'Homme ».

5.2 Une plateforme informatique pour le programme Transversal Toxicologie Nucléaire

Dès 2001, le Laboratoire de génie informatique et d'ingénierie de production (LGI2P) de l'École Nationale Supérieure des Mines d'Alès (EMA) et le CEA ont collaboré pour ce projet. Le LGI2P avait déjà mis en place des plateformes informatiques de stockage de documents pour d'autres programmes scientifiques du CEA. En effet, en 1999, avait été créée une unité mixte de recherche (UMR) entre ces deux organismes. De même le LGI2P a répondu à la nécessité de mettre en place une plateforme de stockage des documents du programme pour garder trace et mémoire du projet. Il y a associé des fonctions collaboratives pour la gestion des documents. Le LGI2P a ajouté également un espace de diffusion et de communication à cette plateforme appelée ToxNuc, puis ToxNuc-E[72] et aujourd'hui Toxcea.

Ce partenariat a permis au CEA de bénéficier de technologies innovantes pour l'époque et au LG2IP d'expérimenter des concepts, des méthodes et des outils en ingénierie de l'information. Ainsi, le LGI2P a pu prendre comme terrain applicatif ce dispositif au service d'une communauté de 600 chercheurs impliqués dans le programme Transversal Toxicologie Nucléaire. A l'analyse des usages d'une telle plateforme par les collectifs de chercheurs, et des non usages de certaines de ses fonctionnalités, le LGI2P a par ailleurs mené un projet de recherche « Intelligence collective et travail collaboratif CYCLONE ». Ce dernier avait autant pour but de visualiser des collaborations entre les différents spécialistes rassemblés autour de la toxicologie nucléaire, que de les favoriser via un dispositif numérique. Dans cette optique, différentes pistes ont été étudiées pour concevoir et introduire dans la plateforme un référentiel commun aux différents spécialistes associés à ce programme transversal. Pour y parvenir, plusieurs procédés ont été testés comme la construction de réseaux sémantiques étendus mais également la création d'une ontologie [73, RICCIO].

L'ontologie, pensée ici comme un référentiel commun, est alors un sous-projet au sein d'un projet lui-même affilié à un programme scientifique. Elle fait l'objet d'une réflexion entre divers spécialistes de la toxicologie nucléaire avec des spécialistes des technologies de la communication et des spécialistes de l'ingénierie des connaissances.

[72] Toxicologie nucléaire **e**nvironnementale

6 Les fonctionnalités de la plateforme ToxNuc : d'un entrepôt à une plateforme de KM

Les données ayant trait à la toxicologie sont générées par diverses disciplines comme la physique, la chimie, la biologie et la médecine. Si elles ne sont pas toujours librement accessibles, au moins les références des publications ou de la littérature grise le sont. Cependant, ces références sont indexées selon les disciplines et les environnements propres aux bases de données qui les répertorient. Ainsi la recherche de documentation et la recherche de nouveaux collaborateurs n'est pas aisée.

6.1 Une archive fermée

De la rareté de données constatée il y a une dizaine d'années[73], notamment en chimiotoxicité, les chercheurs sont confrontés aujourd'hui à la complexité d'interroger et d'exploiter les données issues de multiples nouvelles bases. Mais nombre d'entre elles sont encore souvent construites en silos. De même les publications issues du programme Transversal Toxicologie Nucléaire sont diversement référencées dans plusieurs d'entre elles.

C'est une des raisons pour lesquelles la coordinatrice du programme a émis le besoin d'une plateforme de stockage pour les publications scientifiques relatives au programme dès son démarrage en 2001. Sa volonté était d'y déposer également les documents relatifs à la coordination et au pilotage du programme. Ainsi, la nécessité de garder trace, et donc de garder mémoire des décisions prises et des orientations du programme a été également un élément moteur. En effet, par expérience, la coordination du programme avait été confrontée à la difficulté de rendre des comptes aux tutelles sur des projets antérieurs. Mobiliser a posteriori des informations restées informelles ou disparates sur d'anciens projets s'était révélé difficile. Et cela d'autant plus lorsque leurs membres, partis à la retraite ou mobilisés sur de nouveaux projets, ne pouvaient apporter de réponses. Anticiper l'accès pérenne à l'information au bénéfice des futurs coordinateurs de programme et capitaliser les expériences sont une des composantes du *knowledge management* (KM). A celle-ci s'ajoute la facilitation des échanges et de la collaboration. L'outil de stockage doté de fonctionnalités de communication et d'échange alors mis en place, un temps nommé plateforme K-Hub, rend bien compte de la dimension KM du projet [69, GACHET pp. 35-36].

[73] PubChem Project qui s'est concrétisé par la mise en relation de trois bases de données en 2004 : < https://pubchem.ncbi.nlm.nih.gov >

6.2 Une plateforme collaborative pour communiquer et diffuser

C'est le LGI2P, de l'Ecole des Mines d'Alès, qui fut sollicité, en tant que partenaire du CEA dans le cadre d'une UMR depuis 1999. Il proposa une solution qui répondait à la fois au besoin exprimé d'un site de stockage et favorisant l'échange. Il s'agissait de l'adaptation d'un dispositif préalablement produit pour un autre collectif du CEA sur l'enfouissement de matières dangereuses [72, RICCIO *et al.*]. Par ailleurs, la solution offrait des outils innovants pour le début des années 2000 au sein d'un organisme comme le CEA. Ainsi la plateforme combinait les fonctionnalités :

- d'une Gestion électronique de documents (GED) dédiée au stockage à la description, la sauvegarde et le partage de documents et des publications. Celles-ci sont, soit issues du programme, soit considérées par ses membres comme d'intérêt pour le programme ;
- d'un Content management system (CMS) pour la rédaction par certains membres de la communauté du programme des pages du site web sans soucis de mise en page. Sa fonction est de permettre de communiquer des informations relatives aux colloques et séminaires. Elles sont destinées à l'ensemble des membres, mais aussi à destination du grand public. En effet, une partie des contenus du site est accessible librement sans besoin d'identifiant et de mot de passe ;
- d'un annuaire des membres de la communauté dont il a pu être observé qu'il était d'ailleurs très largement prisé par celle-ci. En effet, au début des années 2000, une partie des institutions autres que le CEA ne disposait pas d'annuaire de ses membres ;
- d'envoi d'une lettre de diffusion à l'adresse électronique des membres ;
- d'un forum qui n'a pas été reconduit dans les versions ultérieures de la plateforme en raison de son non-usage. Celui-ci fut étudié et les parades mises en place pour encourager son utilisation restèrent sans succès.

Accessible par Internet et pour partie par attribution d'identifiants et mots de passe, ce dispositif a permis de contourner les réticences des services informatiques du CEA à élargir le périmètre de l'intranet aux chercheurs des autres EPST. La coordination du programme avec le soutien du LGI2P s'est dotée alors dès 2001 d'un outil adapté. Il a été précurseur sans pour autant cumuler toutes les caractéristiques des *groupwares* annonciateurs des RSE futurs. Il a facilité le partage, la communication, la valorisation des réalisations et le travail à distance. La présence sur Internet a aussi donné de la visibilité au programme Transversal Toxicologie Nucléaire. A travers ce site, il a aussi été donné à des chercheurs en sciences de l'information et de la communication d'étudier les pratiques de cette communauté pluridisciplinaire et multiorganisme ainsi équipée d'un outil collaboratif.

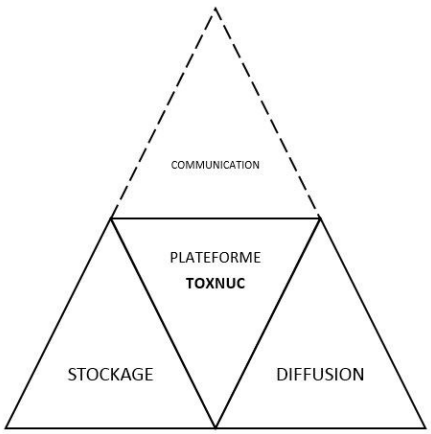


Figure 8 : la pyramide de fonctionnalités de la plateforme pour le programme Transversal Toxicologie Nucléaire. La dimension « communication » n'a pas atteint les objectifs escomptés. Le forum a été abandonné faute d'être utilisé par les chercheurs.

7 Un référentiel ontologique pour le programme Toxicologie Nucléaire

Après avoir retracé l'histoire de la plateforme ToxNuc élaborée pour servir le programme scientifique Toxicologie Nucléaire, il s'agit ici de s'intéresser plus précisément aux motivations et conditions de réalisation d'un premier objet de type ontologique. Très rapidement l'idée de constituer une ontologie de domaine a émergé pour favoriser encore davantage la collaboration entre chercheurs. En effet, introduire un vocabulaire dans la chaîne de traitement des documents archivés sur cette plateforme peut participer de l'amélioration de l'indexation et la recherche de ces derniers. Pour les spécialistes de l'IC, l'idée était aussi d'expérimenter des méthodes et des outils pour réaliser une ontologie à partir des usages collaboratifs et des contenus documentaires de la plateforme.

7.1 Une ontologie de domaine, pour quoi faire ?

7.1.1 Pour servir une communauté et la stratégie de sa direction

La Toxicologie Nucléaire faisant appel à plusieurs disciplines, la recherche bibliographique dans ce domaine y est plus difficile : les mots-clés utilisés pour interroger les bases de données spécialisées sont propres à chaque discipline et le vocabulaire utilisé dans les publications est lui aussi spécifique en fonction de l'origine des auteurs, alors même que la thématique s'avère souvent transversale. En effet, des chercheurs de spécialités différentes travaillent sur les mêmes sujets mais avec une méconnaissance des uns et des autres. Ils n'interrogent pas les mêmes bases de données et ne publient pas dans les mêmes revues. En outre, les chercheurs d'un même projet ne sont pas pour autant regroupés géographiquement. Cette problématique pousse la direction du programme à chercher de nouvelles solutions pour faire émerger les collaborations potentielles entre chercheurs regroupés au sein de quinze projets pluridisciplinaires. En effet, la plateforme décrite plus haut étant en place, « *la première découverte pour les acteurs du programme a été que l'outil ne crée pas le lien social* » [72, RICCIO *et al.* p. 5]. Les chercheurs n'y déposent pas d'informations. L'ensemble des informations mis à disposition sur la plateforme est le fait de la direction du programme. Et chaque tentative de déléguer à des représentants de chaque projet l'enrichissement de la plateforme reste vaine. L'utilisation de la plateforme se limite à la consultation sans utilisation des fonctionnalités d'échange. La nouvelle stratégie de la direction est alors de leur donner à voir des publications d'autres chercheurs et disciplines qui seraient toutefois d'intérêt pour eux pour stimuler le décloisonnement et l'échange. Mais pour les raisons précitées, trouver ces publications s'avère difficile de même que démontrer l'intérêt que pourrait avoir une publication pour un chercheur dans ce contexte de travail. Pour cela, la direction du programme fait appel de nouveau aux compétences des ingénieurs et enseignants en sciences de l'information et de la communication du LGI2P.

> « Dans ce cadre, la direction du programme ToxNuc-E nous sollicite […] : comment identifier dans un assez gros volume de données - ensemble des publications liées au domaine de la toxicologie nucléaire environnementale - les documents susceptibles d'intéresser un ou des chercheurs. » [73, RICCIO. p. 3]

7.1.2 Pour observer l'impact des TIC sur l'activité scientifique

A la suite de cette demande, le constat réalisé conduit à « *initier la construction [d'une] ontologie* » afin de l'instrumentaliser pour répondre au besoin de la direction du programme. Comme cela a été abordé dans les premières pages de ce mémoire, les traitements automatiques et mécanismes de raisonnements permis par un tel objet informatique peuvent être appliqués à la recherche d'information. Il y a plusieurs façons d'optimiser la recherche d'informations et de publications. Cependant, pour les chercheurs du LGI2P, « *élaborer un modèle de connaissance [leur] permettant de faciliter la recherche d'information dans un contexte local* » [ibid. p. 5], crée aussi le terrain favorable à l'observation de leur propre objet d'étude : l'impact des Technologies de l'Information et de la Communication sur l'activité scientifique collaborative.

7.1.3 Pour favoriser la naissance d'une communauté

Historiquement les grands « penseurs » du XVII^{e} siècle, comme Blaise Pascal, étaient souvent érudits dans plusieurs disciplines. Mais au fur et à mesure des découvertes, la couverture d'un domaine représentant une masse de connaissance en soi, les scientifiques se sont spécialisés. Or c'est souvent du croisement de plusieurs disciplines qu'émane l'innovation et à l'aube du XXI^{e} siècle, les progrès technologiques et en particulier les traitements informatiques aidant, on assiste à une volonté de renforcer ces croisements : les sciences du vivant en sont sans doute l'exemple le plus marquant. Mais comment observer une communauté reliée par une thématique commune mais qui s'ignore elle-même ? Qu'est-ce qui fait communauté ? Certainement la reconnaissance mutuelle, le fait de produire de manière collaborative. Mais avant cela, il est nécessaire de mieux se connaître, de dialoguer pour faire communauté. Or souvent, les chercheurs ignorent le fait même que leur thématique est également l'objet d'études, d'un point de vue différent, au sein d'une discipline voisine. Il en résulte une absence d'échange plus liée à une méconnaissance qu'à une volonté propre. Et ce constat était particulièrement présent au sein de la communauté ToxNuc qui regroupe des biologistes, des chimistes, des physiciens, des médecins etc. Après avoir étudié les raisons de cette difficulté de communication, et pour aider les chercheurs à « faire communauté », il est apparu opportun de les faire collaborer à l'établissement d'un langage commun. Celui-ci étant en servant de socle à la fois pour l'indexation, et la recherche d'information, pose les bases de l'identité du collectif. In fine, l'hypothèse à vérifier pourrait être d'observer si cette proposition engendre des collaborations concourant elles-mêmes à de nouvelles publications :

> « un des principaux verrous à lever pour favoriser l'efficacité collective et en particulier la diffusion des travaux était celui de la construction d'un langage commun de la toxicologie nucléaire environnementale. » [*Ibid*. p. 3]

7.1.4 L'ontologie de domaine : ciment invisible de la communauté

Si la construction de l'ontologie de domaine favorise les échanges et peut permettre de créer les fondements de la communauté scientifique, une fois établie et acceptée, elle peut également être le garant de sa pérennité. En effet, elle dote la communauté d'un vocabulaire commun et favorise donc la communication. Mais elle sert également de socle à diverses applications dédiées à la communauté et renforce donc les liens au sein de celle-ci. Par exemple en participant à améliorer la pertinence et l'efficacité des moteurs de recherche. Mais pour ça, il reste un verrou à lever : l'indexation des ressources disponibles dans la communauté à laquelle peut œuvrer le TAL. Comme il a été vu précédemment, le TAL est à la fois producteur mais aussi consommateur de RTO (2.4.2 Ingénierie des connaissances (IC) et traitement automatique des langues (TAL)). Ainsi, dans le cas de la communauté ToxNuc, et d'une ontologie de la Toxicologie Nucléaire, plusieurs applications pouvaient être envisagées : annoter des publications scientifiques avec les différents labels des concepts afin de les indexer, et améliorer la recherche de publications sur la plateforme TocNuc.

7.2 Une ontologie de domaine, par où commencer ?

7.2.1 Combiner les méthodes ascendantes et descendantes

Les ingénieurs du LGI2P ont commencé à travailler avec deux approches différentes : l'une basée sur des interviews d'experts par des étudiantes des sciences humaines et sociales, l'autre par traitement à partir d'un corpus. Ainsi, les méthodes se combinent au cours de la construction, comme l'indiquent [33, AUSSENAC-GILLES *et al.* section 2.4.1]. Ce corpus était constitué de la somme des textes décrivant le programme de travail de chacune des quinze équipes de projet. Sa particularité est de résulter d'une négociation entre les chercheurs des différentes disciplines concernées. On retrouve ici l'idée de consensus et de négociation autour de ce qu'est la toxicologie nucléaire.

Ce travail manuel, inspiré par :

> « la méthode de construction d'une ontologie mise au point sur le projet MENELAS, théorisée par Bruno BACHIMONT, outillée et améliorée par la suite et présentée de façon synthétique par Jean CHARLET […] a permis de dégager assez rapidement 7 catégories principales : disciplines, organismes, modèles biologiques, toxiques d'intérêt, molécules, outils et type d'études et un ensemble d'environ 1.200 « termes candidats ». [73, RICCIO p. 4.]

Finalement, les ingénieurs du LGI2P ne sont pas allés au bout du processus de construction ontologique théorisé par B. Bachimont.

> « L'objectif du travail n'étant pas de construire une ontologie finalisée de la toxicologie nucléaire environnementale, mais d'élaborer un modèle de connaissance nous permettant de faciliter la recherche d'information dans un contexte local, nous avons décidé d'arrêter le processus de construction d'une ontologie à ce stade et de nous contenter de l'arbre des concepts linguistiques ou ontologie régionale » [*Ibid*. p. 5]

Ce n'est qu'en 2011 que la formalisation a été réalisée afin de la transformer en ontologie informatique. L'encodage en OWL (*Ontology Web Language*) par transformation de feuilles de calcul au format Excel® a été réalisé par deux chercheurs affiliés au Limics (Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-Santé) et spécialistes des ontologies informatiques. L'ontologie formelle a ensuite été enrichie par le travail d'une stagiaire supervisée par une chercheuse du LGI2P spécialiste des ontologies informatiques. Cependant, des suites n'ont pu être données avant 2014 aux lacunes constatées ainsi qu'aux réserves émises au moment de la formalisation en 2011. L'ontologie ToxNuc n'a donc toujours pas été testée pour étudier dans quelle mesure elle faciliterait la recherche d'information dans le contexte local de la plateforme ToxNuc. En effet, il était convenu de déposer les publications à venir générées par les avancées et expérimentations du programme. Cette ontologie a cependant déjà encouragé la collaboration recherchée par la direction du programme grâce au processus de construction mis en place. Elle a donc pleinement rempli son 1er rôle théorique. Celui-ci étant de définir et fournir à des humains une sémantique interprétative d'un domaine du monde réel fondé sur un consensus, soit un langage commun.

7.2.2 Explorer des méthodes alternatives

Parallèlement, les ingénieurs du LGI2P ont imaginé et expérimenté un autre dispositif technologique avec des cartes sémantiques pour répondre directement au besoin de la direction du programme : découverte de documents d'intérêts et donc aussi de collaborateurs potentiels. Les ingénieurs du LGI2P ont eu recours aux outils du *Text mining*, notamment le TALN et la classification automatique. Mais le parti pris était celui d'une « *instrumentation technique* » qui donnerait « *toute sa place à l'expert (traitement manuel) tout en utilisant au mieux la capacité des calculateurs (traitement automatique)* » [*Ibid*. p. 5]. Ce dispositif prenait appui sur les recherches dans le domaine des réseaux sémantiques et des réseaux de proximité. Ainsi, ils ont réalisé « un prototype d'éditeur de réseaux sémantiques » pour « *la construction rapide de modèles conceptuels par des experts en situation, puis l'enrichissement de ces modèles conceptuels en prenant appui sur une cartographie de termes issus du traitement automatique d'un ensemble de documents soigneusement sélectionnés*. » Celui-ci a donné lieu à des démonstrations très positives grâce à l'implication de certaines équipes de projet du programme, notamment celle du projet Arabidopsis[74], et a satisfait la communauté ToxNuc. Au final, c'est une alternative pour créer des référentiels ontologiques plus rapidement.

[74] Projet ayant pris le nom de la plante modèle Arabidopsis thaliana. Cette plante, dont le génome est séquencé, est utilisée pour étudier l'impact des métaux lourds et des radionucléides chez les végétaux. (cf. http://www.toxnuc-e.org/document.php?pagendx=118).

7.3 Une ontologie de domaine, vers où continuer ?

7.3.1 Les folksonomies, une nouvelle perspective pour cette ontologie ?

Le recours à la visualisation de cette deuxième expérience ainsi que la manipulation de cartes sémantiques sont inspirantes à plus d'un titre. L'ajout possible de termes comme autant d'étiquettes venant éclairer le domaine de multiples points de vue a inspiré une nouvelle expérimentation au LGI2P. Cette expérimentation s'inspire également de la folksonomie, néologisme issu du croisement entre *folks* (les gens) et *taxonomy* (taxonomie), autrement dit de l'indexation sociale[75]. Encore non abordées dans ce mémoire, les *folksonomies* se distinguent des autres SOCs abordés précédemment. La raison en est que les étiquettes ou tags, soit les mots-clés utilisés pour qualifier des ressources dans l'univers du Web, ne sont pas contraints par une liste fermée de termes. Ils n'appartiennent pas à un vocabulaire contrôlé comme dans le cas des autres SOCs évoqués jusqu'ici. Muriel Amar, faisant la synthèse de plusieurs études et questionnements sur l'indexation sociale, la caractérise comme une catégorisation a posteriori, décentralisée et spontanée. C'est l'utilisateur qui choisit son propre langage d'indexation. Indexation dont les *tags* (étiquettes) personnels servent tantôt à la recherche, tantôt à la navigation pour lui comme pour les autres usagers des sites dits de *social bookmarking* [12-13, AMAR]. Pour citer les applications les plus connues, il y a pour les ouvrages *LibraryThing* et *Babelio*, *BibSonomy* pour les publications, *Deli.ici.us* et *Diigo* pour les ressources en ligne et *Flickr* pour les photos. Aujourd'hui beaucoup d'autres applications du Web social ou dit encore Web 2.0 intègrent de telles fonctionnalités. Ainsi les logiciels de gestion de références bibliographiques (LGRB) comme *CiteUlike* et *Zotero* proposent aux utilisateurs de taguer les références enregistrées. Pour paraphraser les auteurs évoqués par Muriel Amar, il s'agit de tirer parti de l'intelligence collective et de combiner logique individuelle et logique de collaboration. Et c'est d'ailleurs cette pratique, dont l'essor remonte à 2004, qui est à l'origine du nom attribué à l'outil développé par le LGI2P : *Folksonomies*. Ce projet, tout comme d'autres et comme le considèrent certains, illustre bien le fait qu'ontologie et folksonomie peuvent se compléter [14, ERTZSCHEID ; 15-16, POUPEAU ; 17, VAN DAMME *et al. ;* 28, GANDON].

7.3.2 Les folksonomies pour combiner les nouvelles formes d'expression des experts et la puissance algorithmique

Avec l'outil *Folksonomies*, l'idée est de proposer plusieurs graphes représentant les publications du programme aux utilisateurs de ToxNuc. Ainsi, avec un nombre de publications suffisamment important, le premier prototype réalisé permet d'appréhender des regroupements de publications. Au sein des graphes, chaque nœud correspond à une publication et est étiqueté à la demande, soit par la thématique dominante de celle-ci, soit par ses références bibliographiques. Le calcul pour représenter ces graphes se base sur la proximité entre les publications.

[75] < http://www.culture.fr/layout/set/print/franceterme/terme/INFO756 >

A l'avenir, la proximité combinerait les distances sémantiques issues de l'analyse automatique de texte appliquée aux publications, pondérées par les distances sémantiques entre les termes de l'ontologie ayant permis une annotation des textes. Cette association enrichirait la segmentation avec de la sémantique issue non plus seulement de calculs mais aussi d'un consensus d'experts. Une fois connecté à son compte, le chercheur étiquetterait avec ses propres tags les publications. Cela permettrait, pourquoi pas, d'en faire des candidats termes pour la mise à jour de l'ontologie ? De plus, à chaque chercheur pourrait être présentée une vue qui lui serait propre car appelée par ses propres mots-clés et/ou aussi étiquetée avec eux.

7.4 De nouvelles opportunités pour cette ontologie de domaine

7.4.1 Construire une ontologie de domaine pour améliorer l'indexation ?

Qualifier l'information au plus juste et en fonction des besoins de chacun, grâce à une indexation manuelle semble dépassé dans l'écosystème actuel : explosion des publications scientifiques et « infobésité ». De plus, elle nécessite que l'opérateur humain maîtrise une double compétence : connaissance du domaine et aussi connaissance du vocabulaire pour un respect des règles de l'indexation. Mais par ailleurs, la recherche d'information basée uniquement sur des index automatiques ne permet pas de profiter pleinement de la richesse propre des textes. Une ontologie, quant à elle, permet d'acquérir de la connaissance sur un domaine et d'enrichir les processus d'indexation automatique en apportant une plus grande finesse sémantique. Cependant, l'indexation ne peut jamais vraiment être totalement automatique. Elle nécessite toujours un contrôle humain [18, CHARTRON *et al.*]. Celui-ci peut être réalisé par des documentalistes qui, en amont, valident ou non, et réajustent l'indexation proposée par le système. Mais comme il a été vu avec l'outil *Folksonomies*, ce contrôle peut être apporté de façon plus originale grâce à une indexation à la fois collective et a posteriori. En effet, dans ce cas présent, le contrôle humain est déporté en aval. Il est reporté sur les chercheurs-tagueurs qui peuvent ne pas le percevoir comme un contrôle ou comme une forme de *curation*. Il peut être plutôt perçu comme une appropriation des connaissances. Il se double d'une dimension sociale car profitable à tous par l'ajustement du référentiel ainsi permis. C'est une sorte de réintermédiation où négocient entre elles les logiques de personnalisation et de normalisation. Il s'agirait là également d'une nouvelle exploitation des traces laissées par les pratiques d'annotation intellectuelle. Ces traces d'annotation manuelle mais outillée par l'informatique, sont des dérivées numériques des manicules et autres pointeurs apparus dans les manuscrits médiévaux.

7.4.2 Construire une ontologie de domaine : moteur de connaissance mais aussi d'innovation

L'application *Folksonomies* est une façon innovante et ludique de solliciter des experts. Cependant, la masse critique d'experts-tagueurs pourrait ne pas être atteinte car la proportion des individus à s'approprier de nouveaux outils est toujours minoritaire [78, CARMES]. De même, pourraient concourir à ce risque de mauvaises performances ou un graphisme peu attractif. Par ailleurs, la faiblesse déjà constatée des dépôts fait courir le risque de ne pas atteindre la masse critique de publications. Enfin, il semble que la communauté ToxNuc se soit distendue.

Pourtant, l'interaction et la personnalisation sont pleinement dans l'air du temps de même que l'aspect visualisation[76]. Les vues offertes par *Folksonomies* sont en cohérence avec les offres du secteur privé, comme par exemple celles proposées par la start-up *Dataveyes-Human Data Interactions*[77]. En effet, la *Dataviz* est plastique, au sens où elle explose actuellement [76, ARRUABARRENA]. Cette dernière l'est aussi dans la mesure où les vues changent en fonctions des flux de données qui les alimentent. Que pourra en déduire notre cerveau, qui est lui aussi plastique aux dires des découvertes des neurosciences ? Et que pourra-t-on lui faire interpréter de ces formes nouvelles de représentation[78] ? Il faudra devant et derrière l'écran des compétences et des individus métamorphosés pour travailler à ces formes[79].

Toujours dans une perspective d'innovation, pourquoi aussi ne pas envisager d'étudier les concepts des *serious games* pour la co-construction de ce type de référentiels avec les experts ? Pourquoi ne pas adapter de façon numérique les outils dérivés des *innovation games* ? Mais les toxicologues participeraient-ils plus facilement à construire des référentiels ontologiques ? Les sociologues et les épistémologues étudient les façons et les raisons dont les technologies numériques transforment le rapport des chercheurs à leur domaine. Comment ces technologies numériques deviennent-elles des outils supplémentaires ? Comment actualisent-elles la façon de faire de la science ? Qui et quelles compétences peuvent aider les chercheurs à utiliser leurs potentialités ? Questionner ce qui fait Science au travers de l'observation de comment elle se fait aujourd'hui pourrait rejoindre les préoccupations des métaphysiciens auxquels le terme d'ontologie a été emprunté. C'est en tous les cas une nouvelle transdiscipline aux perspectives multiples que certains se proposent de définir[80]. Ils se regroupent derrière l'étiquette *Digital Humanities* où le numérique y est à la fois sujet et objet.

---

[76] MERZEAU L. « Traces captées traces éditorialisées ». In : *Mémoire numérique. Publics, ressources et bibliothèques en mutation, Journée d'étude organisée par Médiadix et l'URFIST de Paris*. Médiadix, Saint-Cloud : [s.n.], 2014. Disponible sur : http://merzeau.net/traces-captees-editorialisees

[77] http://dataveyes.com

[78] MALABOU C. *Que faire de notre cerveau ?* 2nde édition. Montrouge : Bayard Jeunesse, 2011. 189 p. ISBN : 978-2-2274-8313-2

[79] LE DEUFF O. Humanités numériques : un concept en définition [En ligne]. *Le guide des égarés.* 2 mars 2012. Disponible sur : http://www.guidedesegares.info/2012/03/02/humanites-numeriques-un-concept-en-definition

[80] DACOS M. *Manifeste des Digital humanities* [En ligne]. *ThatCamp Paris 2010*. 26 mars 2011. Disponible sur : < http://tcp.hypotheses.org/318 >
LE DEUFF O. *Un contexte scientifique interdisciplinaire : Transdiscipline et translittératie* [En ligne]. *HUMANLIT*. 29 octobre 2012. Disponible sur : < http://humanlit.hypotheses.org/13 >

8 L'ontologie ToxNuc, d'hier à aujourd'hui : une ontologie à orienter, pourquoi ?

Formalisée en 2011 avec le langage OWL et l'utilisation de quelques labels SKOS avec l'éditeur d'ontologies *Protégé* mais présentant quelques erreurs et lacunes, l'ontologie informatique ToxNuc n'était cependant pas encore corrigée début 2014. L'annotation et l'indexation des publications du programme au moyen de cette ontologie n'ont donc pas été testées. D'après les informations recueillies pour établir l'historique de l'ensemble plateforme-ontologie-publications, plusieurs suppositions peuvent alors être émises pour expliquer l'état actuel de l'ontologie. Mais qu'est-il d'abord possible d'observer du point de vue technique à propos de l'ontologie ToxNuc ?

8.1 Le bilan technique de l'ontologie ToxNuc à l'été 2014

8.1.1 Du point de vue formel

L'ontologie a donc fourni une sémantique interprétative du consensus entre les différents acteurs de la toxicologie nucléaire réunis au sein de ce projet entre 2005 et 2006. Mais en revanche, la sémantique formelle permet-elle son exploitation par un ordinateur et d'effectuer des inférences ?

Pour mémoire, ToxNuc a été formalisée en 2011 pour devenir une ontologie informatique. Ainsi, l'observation qui suit a pu être réalisée après ouverture du fichier OWL (*Ontology Web Language*) transmis en début de stage avec l'éditeur d'ontologie *Protégé*[81]. ToxNuc est caractérisée par « *7 catégories principales : disciplines, organismes, modèles biologiques, toxiques d'intérêt, molécules, outils et type d'études.* » [73, RICCIO p. 4]. Ces sept catégories ou classes sont composées de sous-classes, elles-mêmes subdivisées encore en sous-classes. Le tout correspond à 646 classes. « *Une classe est quelque chose d'abstrait, plutôt qu'un élément particulier de l'ensemble d'objets qu'elle décrit* » [79, CHARLET diapo. 74]. La hiérarchie de classes s'apparente alors à un arbre de concepts différentiels où les sous-classes héritent des attributs et des valeurs des classes qui les incluent. Pour l'instant, il n'y a pas d'instances déclarées. A l'avenir, les valeurs porteuses de la connaissance pourront soit peupler l'ontologie en étant associées aux catégories, soit être renvoyées au système opérationnel auquel s'intégrera l'ontologie en fonction de la redéfinition potentielle de sa destination.

Cependant, est-il possible de parler de conceptualisation ? Oui car il y a des classes reliées par des relations taxinomiques. Mais la présence de seulement ce type de relations fait qu'il n'y a pas d'engagement ontologique. En effet, aucun attribut, ni aucun arbre de relations ne caractérisent ces classes. Il n'y a ainsi pas la possibilité de construire des classes définies par conditions nécessaires et suffisantes. Des raisonnements et une classification automatique des classes les unes par rapport aux autres ne peuvent être inférés. Ainsi ToxNuc est une taxinomie de concepts avec un faible niveau de connectivité.

[81] Pour une présentation de Protégé, se référer à l'Annexe 1- Livrable 3- État de l'art : répertoire des ressources termino-ontologiques disponibles dans le domaine p. 22/34.

Du point de vue formel, cette ontologie s'apparente plus à une taxinomie, soit un squelette ontologique. Elle pourrait aussi être qualifiée d'ontologie légère. L'ontologie ayant été validée, la relation de subsomption est en principe de type genre-espèce (est_un) ou (is_a) comme il convient pour ce genre d'artefact. Puisqu'elle est opérationnalisée avec le langage de représentation des connaissances OWL (*Ontology Web Language*), elle constitue une base pour envisager une expressivité plus importante désormais.

8.1.2 Du point de vue syntaxique et documentaire

« l'OWL ne fait pas l'ontologie » [*Ibid*, diapo. 121]

Avant d'envisager à nouveau, presque dix ans après sa constitution, à quelle exploitation par un système informatique la destiner, il est déjà nécessaire de corriger le fichier. En effet, celui-ci comporte des erreurs qu'il convient de rectifier afin de pouvoir le soumettre à des éventuels tests :

- L'encodage pose problème car il y a des caractères accentués pour le nommage des classes.
- La syntaxe des URI (*Uniform Ressource Identifier*) n'est pas correcte.
- L'écriture des labels, quand ceux-ci sont présents, n'est pas homogène d'une classe à l'autre.

Les labels évoqués ici sont le seul type d'annotation présent dans ce fichier OWL. Les annotations donnent des informations sur les concepts. Elles peuvent être des définitions, des identifiants... De ce fait, elles participent à documenter les ontologies. Des indications de labels préférentiels ou alternatifs offrent une richesse linguistique supplémentaire pouvant aussi être de l'ordre du multilinguisme si une précision quant à la langue est également donnée. Ici, un soin a été apporté pour garantir à chaque concept un label préférentiel en langue française et un label préférentiel en langue anglaise. Cependant il y a quelques lacunes. Par ailleurs, certains de ces labels sont encapsulés dans de mauvaises balises. Ainsi, certains ont pour annotations « *Comment* » ou encore « *Definition* » au lieu de « *Label* ».

Quelques annotations figurent aussi pour documenter le fichier dans son ensemble. Il s'agit du nom de l'ontologie, d'une phrase indiquant la date de création du fichier ainsi que la manière dont il a été constitué à partir d'un fichier Excel contenant la hiérarchie de classes, et enfin des contributeurs à ce fichier, en utilisant respectivement les annotations de type « *Label* », « *Comment* » et « *Contributor* ».

Ainsi, le fichier contient peu d'informations sur lui-même et les classes. En conséquence, très peu d'informations peuvent en l'état être manipulées par un système informatique, un langage de requêtes, ni même être interprétées par un humain qui consulte ce fichier via une interface d'édition. Ce fichier n'est pas autoporteur de sa documentation. Cette documentation non-embarquée dans le fichier fait défaut par ailleurs (8.3 Les dimensions organisationnelle et technique).

L'ontologie semble donc au moins inachevée pour ce qui est de répondre à la demande « *d'identifier dans un assez gros volume de données - ensemble des publications liées au domaine de la toxicologie nucléaire environnementale - les documents susceptibles d'intéresser un ou des chercheurs.* » [73, RICCIO p. 3]. Pour le comprendre ou tout du moins émettre des hypothèses à ce sujet, les lignes qui suivent s'efforcent d'apporter des éléments de contexte propre à la constitution de l'ontologie ToxNuc.

8.2 Les dimensions socio-technique et techno-politique

Plusieurs causes exogènes peuvent expliquer la situation actuelle du projet d'ontologie pour le programme Transversal Toxicologie Nucléaire.

Tout d'abord, pour ce qui est de sa dimension socio-technique, il est possible que les avantages d'une ontologie n'aient pas été perçus par la communauté ToxNuc car la bio-informatique en était à ses balbutiements en 2005. En effet, le potentiel des ontologies comme par exemple, l'annotation des données et l'interopérabilité des bases de données génomiques grâce à un vocabulaire contrôlé comme la Gene Ontology (GO), n'était pas encore aussi connu qu'aujourd'hui[82]. Cependant, le coordonnateur du projet Arabidopsis[83], assisté d'un professionnel de l'information documentation, se sont particulièrement impliqués, tout comme pour l'expérimentation sur les réseaux sémantiques étendus abordée en 5.2 (Une plateforme informatique pour le programme Transversal Toxicologie Nucléaire). C'est d'ailleurs la hiérarchie de termes et les catégories définies pour ce projet qui ont inspiré l'ontologie ToxNuc commune à l'ensemble des projets du programme scientifique.

De plus, la possible absence de vision est probablement responsable du fait que des objectifs n'ont pas clairement été définis pour cette ontologie. Il ne semble pas non plus qu'il y ait eu de mise en œuvre d'une organisation méthodologique caractéristique de la gestion de projet. Par ailleurs, il est intéressant de noter que la conception du portail ToxNuc a eu lieu elle aussi sans formalisation d'un cahier des charges fonctionnel et technique. Ceci peut être expliqué par le fait que le portail était une duplication d'une plateforme générique, déjà mise en service au CEA par le passé. Le LGI2P était donc en mesure de proposer rapidement une solution technique de stockage et de diffusion. Ceci pouvait justifier de ne pas alourdir le processus de conception par un recueil de besoins.

Du point de vue socio politique, la volonté était de faire collaborer les membres d'un collectif de chercheurs venus d'horizons divers. En prenant en compte cette considération, il est autrement possible de s'interroger sur le projet de conception de l'ontologie du domaine de la Toxicologie Nucléaire. Comme cela a été abordé précédemment (7.1 Une ontologie de domaine, pour quoi faire ?), la réflexion entamée pour la co-construire a été une opportunité pour inciter au dialogue entre les membres du collectif. Si une ontologie n'est jamais une fin mais un moyen, dans ce cas précis elle aura été un artefact supplémentaire aux séminaires communs et aux cours de spécialités croisés [72, RICCIO et al.]. En effet, ces derniers ont été imaginés pour amorcer l'intercompréhension nécessaire au travail commun et à distance par l'intermédiation d'une plateforme.

---

[82] ASHBURNER M., BALL C. A., BLAKE J. A., BOTSTEIN D., BUTLER H., CHERRY J. M., DAVIS A. P., DOLINSKI K., DWIGHT S. S., EPPIG J. T., HARRIS M. A., HILL D. P., ISSEL-TARVER L., KASARSKIS A., LEWIS S., MATESE J. C., RICHARDSON J. E., RINGWALD M., RUBIN G. M., SHERLOCK G. « Gene Ontology: tool for the unification of biology ». *Nat Genet* [En ligne]. mai 2000. Vol. 25, n°1, p. 25-29. Disponible sur : < http://dx.doi.org/10.1038/75556 > PMID PMCID: PMC3037419 < http://geneontology.org/ >

[83] Projet ayant pris le nom de la plante modèle Arabidopsis thaliana. Cette plante, dont le génome est séquencé, est utilisée pour étudier l'impact des métaux lourds et des radionucléides chez les végétaux (cf. http://www.toxnuc-e.org/document.php?pagendx=118).

8.3 Les dimensions organisationnelle et technique

Hormis la dimension sociale préalablement abordée, l'ontologie de la Toxicologie nucléaire ne peut être utilisée que si elle est nécessaire à une application dans laquelle elle doit s'insérer. Le cas d'usage prévu était d'utiliser l'ontologie pour superviser la catégorisation des publications. Autrement dit, il s'agissait de classifier les publications par rapport à des catégories prédéfinies et désignées par les concepts de l'ontologie. Or, « *très peu de documents (comparativement à la quantité produite par les équipes de recherche) ont été mis sur les plateformes successives de ToxNuc.* »[84] En conséquence, leur nombre, mais aussi leur qualité et celle de leurs métadonnées, ne permettent pas de tester l'ontologie. Référencer et déposer dans l'archive les publications issues du programme scientifique étaient naturellement loin d'être les préoccupations premières des chercheurs et des coordinateurs de projets. Ainsi, même le peu de publications déposées est inexploitable pour plusieurs raisons :

- certaines sont sous des formats PDF non exploitables par les outils de TALN ;
- la description bibliographique n'est pas homogène. Les métadonnées descriptives nécessiteraient d'être revues et corrigées. Une des solutions faisant appel à des traitements semi-automatiques est envisagée, mais pas à court terme ;
- les chercheurs n'ont pas été informés des conséquences de la non-qualité de cette première indexation. Ils n'ont pas non plus été formés au dépôt dans l'application. Aucun personnel non chercheur n'a été désigné pour le faire ou pour contrôler la qualité des signalements. Il aurait pu s'agir d'un professionnel de l'I&D car cette profession est sensibilisée à la normalisation des données bibliographiques. Y a-t-on songé ? La réponse penche plutôt du côté du manque de ressources et de coordination. Cela aurait été pourtant nécessaire pour la réalisation de ces travaux du fait de l'aspect multi-compétences d'un tel projet. Mais il faut rappeler que ce projet était annexe au programme Transversal Toxicologie Nucléaire. Il semble alors normal que certains moyens n'aient pas mis à disposition.

L'expérimentation sur les réseaux d'acteurs et la création d'une ontologie sont donc des besoins secondaires par rapport à la mise en place de la plateforme collaborative et à la recherche sur la toxicité nucléaire. Et malgré l'enchevêtrement des projets, des sous-projets et des projets parallèles, ces recherches furent cependant à l'origine de plusieurs expérimentations concluantes et que l'on peut qualifier de « gagnant-gagnant ».

[84] Propos recueillis auprès de Sylvie Ranwez, maître de conférences en Informatique habilitée à diriger des recherches (Université de Montpellier II, LGI2P Ecole des Mines d'Alès), qui a été amenée à travailler ponctuellement pour l'ontologie de la Toxicologie Nucléaire.

Le financement de la recherche dans un contexte de rigueur budgétaire, explique en partie le ralentissement de ce projet d'ontologie. En effet, l'Agence Nationale de la recherche (ANR) accepte et reconduit des programmes de recherche pour des durées de deux à trois ans. Ceci peut être préjudiciable à l'aboutissement de certains projets. Ainsi le programme Transversal Toxicologie Nucléaire a connu trois années de césure entre les périodes 2004-2007 et 2010-2014. L'ontologie, tout comme les autres expériences autour des réseaux sémantiques étendus et des cartographies dynamiques, a pu pâtir des durées limitées des financements et de la non extensivité des ressources allouées.

Ce projet, dense par la coordination que requiert un programme pluridisciplinaire, l'a aussi été au travers des expérimentations innovantes menées en parallèle et en complémentarité avec des professionnels de l'information-communication et de l'ingénierie des connaissances. Le portail et l'ontologie sont des outils innovants répondant à cette densité et sont eux-mêmes sources d'innovation. Ils avaient comme mission originelle que « *les successeurs [chercheurs et futurs coordinateurs] sachent ce qui a été fait et n'aient pas à le refaire* ». C'est là le paradoxe de ce projet orienté collaboration scientifique et gestion des connaissances. En effet, les ressources affectées à ce projet l'ont été à temps partiel et de façon intermittente. Il s'agissait de personnes encore en formation doctorale, en Sciences de l'information et de la communication ou encore en Sciences et technologies de l'information et de la communication (STIC), et en master professionnel Communication et technologie numérique, et donc non-expertes en ingénierie des connaissances, ni en gestion de projet documentaire [73, RICCIO]. Par conséquent, quand il s'agit de retracer l'historique du projet, le constat est celui d'une capitalisation parcellaire, dispersée et incomplète à son propos. En effet, la prise de connaissance de l'historique de l'ontologie et de la plateforme n'a pu se faire par la consultation d'une documentation intentionnellement constituée. L'historique relaté ici est une reconstitution, probablement lacunaire si ce n'est inexacte. Il a été réalisé à partir d'entretiens auprès de quelques acteurs des différentes institutions partenaires, disponibles et facilement joignables. Cet historique a pu être complété grâce à une recherche documentaire des publications disponibles en ligne. Il s'agit des publications des anciens doctorants et stagiaires dont les noms ont émergé lors des entretiens, ainsi que des publications de leurs encadrants. La figure suivante est également une tentative de synthèse. Elle vise à dresser le bilan de l'environnement stratégique et organisationnel de la plateforme ToxNuc ainsi que du projet d'ontologie de domaine associée.

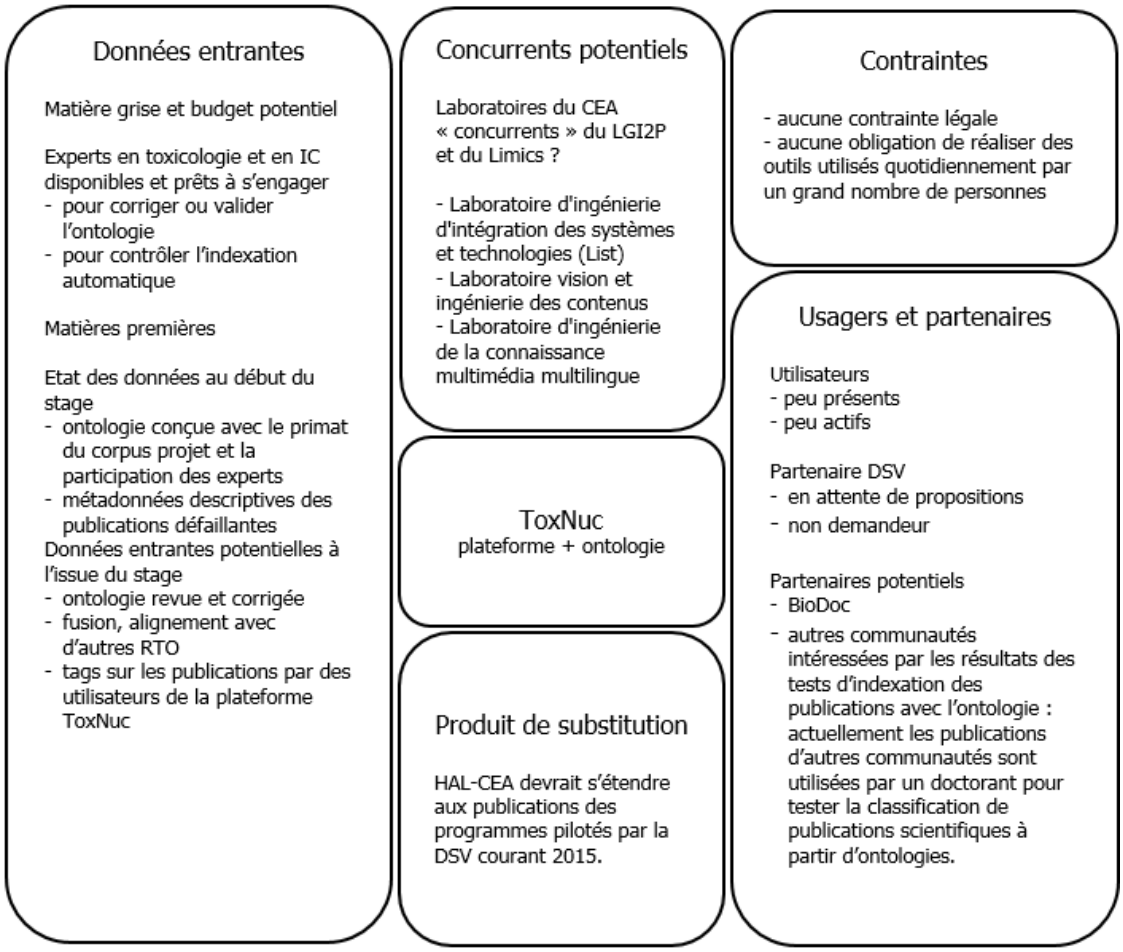


Figure 9 : l'analyse stratégique de l'environnement de la plateforme ToxNuc et de l'ontologie associée.

Cette figure est inspirée du diagramme Michael Porter représentant les cinq forces de l'intensité concurrentielle d'une organisation. Elle a été adaptée pour analyser le projet d'ontologie pour son intégration dans la plateforme ToxNuc en fonction des informations recueillies pendant le stage[85].

[85] Une présentation du modèle de Michael Porter est consultable à l'adresse suivante : < http://chohmann.free.fr/strategie/modele_porter.htm >

Troisième partie

Le cycle de vie d'une ontologie, de la réutilisation à la valorisation : éléments de préconisation

9 L'opportunité de reprendre le projet d'ontologie de la toxicologie nucléaire

Depuis 2001, date de mise en œuvre du programme Transversal Toxicologie Nucléaire, c'est 532 publications avec de forts taux de citations qui ont été recensées [69, DSV/DIR-BIODOC-CEA]. En plus de ces publications, une science ou une discipline se reconnaît aussi aux outils qu'elle produit pour élaborer sa recherche. Se pose alors aujourd'hui la question de poursuivre l'élaboration de cet outil ontologique initié en 2005 puis repris en 2011.

9.1 Réaliser une étude d'opportunité à partir d'un état de l'art

9.1.1 Un projet aligné sur la stratégie du CEA

Comme le montre le bilan bibliométrique du programme produit par le service de documentation BioDoc, de la Direction des Sciences du Vivant du CEA, et présenté lors de la journée scientifique Toxicologie, la recherche dans ce domaine reste une spécialité où le CEA conserve le *leadership*. En elle-même, l'ontologie participerait logiquement de l'apport du CEA à cette discipline où il fait figure de pionnier. Ce stage peut être positionné comme l'un des éléments participant à l'une ou plusieurs des étapes du cycle de vie d'une ontologie. Si comme pour tout projet de conception les phases sont celles illustrées par la Figure 6 : le cycle de vie d'une ontologie p. 33, le stage peut à la fois être lié aux étapes suivantes :

- « Développement » et « Validation » dans la mesure où il s'agit de parfaire l'ontologie telle qu'elle a été commencée ;
- « Évaluation » puisqu'il s'agit aussi de l'étudier pour la comparer à d'autres ;
- « Détections et spécification des besoins » car à l'issue des réflexions auxquelles amènent les étapes précédentes, et étant donné aussi les années écoulées depuis l'émergence de l'idée de réaliser une ontologie de la toxicologie nucléaire, ceux-ci tout autant que le contexte peuvent avoir évolué et nécessiter d'être réexaminés.

9.1.2 Une phase d'avant-projet pour clarifier le contexte stratégique et technique

Cependant et comme le rappellent de nombreux articles scientifiques consacrés aux ontologies, construire un tel artefact est long et mobilise les compétences de différents types d'expertises. Un projet d'ontologie requiert d'être traité d'une façon particulière, c'est-à-dire en accordant de l'importance à l'anticipation. Au préalable, il est nécessaire de clarifier le contexte stratégique et technique d'un tel projet. En effet, les années passant, des projets similaires ou des ressources complémentaires ont pu émerger au sein des communautés voisines. Dans cette hypothèse, l'ontologie ToxNuc est-elle obsolète parce que surpassée par une ontologie déjà opérante ? Ou a-t-elle à gagner d'autres ressources produites ailleurs depuis son initiation en 2005 ? Ainsi, la première démarche est alors de réaliser une étude d'opportunité avant même d'établir une étude de faisabilité.

9.1.3 Un état de l'art pour répertorier des ressources termino-ontologiques (RTO)

L'état de l'art est l'une des commandes du stage. L'autre consistait à l'affinage de l'ontologie à partir duLe bilan technique de l'ontologie ToxNuc à l'été 2014 et présenté en 8.1. L'ontologie comportait encore des erreurs ainsi que des lacunes au niveau des labels. Renforcer l'état de l'art devrait permettre d'évaluer l'impact potentiel d'une telle ressource pour décider ou non de la reprise. En cas de reprise, il devrait permettre de guider l'orientation des travaux de construction de l'ontologie sur la toxicologie nucléaire. L'état de l'art consiste à la fois en une recherche bibliographique et une recherche de ressources terminologiques et ontologiques (RTO).

9.1.4 La recherche documentaire informatisée à l'aide des sources traditionnelles

La première phase de la recherche bibliographique est d'abord la délimitation du sujet de recherche. Dans le cas présent il s'agit d'identifier des articles relatant des mises en œuvre ou l'exploitation de RTO dans le domaine de la toxicité. Il est ensuite nécessaire d'identifier les sources à interroger. En se référant à l'étude bibliométrique réalisée par le service BioDoc pour considérer les champs disciplinaires recouverts par les publications du programme, voici les ressources traditionnelles privilégiées pour la recherche et la veille documentaire[86] :

- les bases de données spécialisées dans le domaine biomédical
    - PubMed
    - ToxNet
- la base de référence dans le domaine nucléaire produite par l'*Atomic Energy Agency* (AIEA)
    - INIS (*International Nuclear Information System*)
- les bases spécialisées en informatique et en technologies de l'information et de la communication
    - ACM *Digital Library*
    - LISTA (*Library, Information Science & Technology Abstracts*)[87]
- les moteurs de recherche
    - CiteSeerX
    - Google scholar

[86] Voir aussi la liste constituée p. 26/34 de l'Annexe 1- Livrable 3- État de l'art : répertoire des ressources termino-ontologiques disponibles dans le domaine.
[87] Base consultable uniquement sur abonnement.

9.2 Les résultats de la recherche et de la veille bibliographique

9.2.1 L'ontologie ToxNuc, une initiative demeurée unique

Il semble que l'initiative ToxNuc reste unique. Cependant, la communauté OpenTox a été identifiée. Dans le cadre d'un projet européen, elle s'efforce de développer un environnement applicatif favorisant la toxicologie prédictive grâce à l'interopérabilité des données. Parmi les initiatives de cette communauté, il y a la création d'une ontologie et d'un vocabulaire contrôlé de la toxicologie. Sa stratégie de communication basée entre autres sur les réseaux sociaux numériques semble notamment un modèle à étudier. Par ailleurs, une autre communauté interdisciplinaire autour du développement et de l'application des nanotechnologies pour la détection, le diagnostic et le traitement du cancer a publié une ontologie et des éléments concernant la méthodologie associée.

9.2.2 De l'utilité de connaître les différents modes d'alimentation et d'interrogation des bases de données

L'interrogation de ces ressources a permis de repérer et de cibler des revues. La surveillance[88] de leurs sommaires respectifs peut s'avérer plus utile que d'attendre le référencement puis l'indexation de leurs articles dans ces bases de données. Elle est plus utile aussi que d'attendre les alertes paramétrables via les fonctionnalités de veille offertes par les bases de données. Au passage, il est intéressant de noter que le terme « *Biological ontologies* » a été introduit dans le thésaurus MeSH de PubMed seulement en 2014. Ceci incite donc à construire plusieurs équations de recherche selon la période surveillée. Dans le thésaurus de LISTA, il est également intéressant de noter que la notion d'ontologie se rapporte à celle de la recherche d'information. En effet, la seule occurrence présente dans le thésaurus est « *ONTOLOGIES (Information retrieval)* ». C'est en effet pour leurs utilisations pour la recherche de l'information que les professionnels de l'I&D s'y intéressent même si elles ne peuvent à elles seules la faciliter [46, FRANCART].

9.2.3 Des résultats intéressants à plus d'un titre

Par ailleurs, les résultats ont confirmé qu'une ontologie est développée pour un usage si spécifique que sa réutilisation et son assemblage avec d'autres apparaissent difficile. Mais ces résultats présentent un autre intérêt. En effet, les articles trouvés relatent des expériences de « débogage »[89] d'ontologies ou de méthodes alternatives de construction d'ontologies. Or, comme il a déjà été mentionné, ces derniers sujets font l'objet de recherches au sein de l'IC et du LGI2P.

[88] Paramétrable grâce à la récupération ou la création de fil de syndication RSS.
[89] < http://www.culture.fr/layout/set/print/franceterme/terme/INFO109 >

9.3 Le repérage, la sélection et la récupération de RTO

9.3.1 Des moteurs et des répertoires spécialisés en passant par les options de recherche des moteurs généralistes

La recherche bibliographique a donc permis d'identifier une communauté et par rebond sur le Web de récupérer des fichiers éditables pour certains avec le logiciel *Protégé*[90]. Cependant plusieurs autres stratégies de recherches de fichiers en OWL ont été mises en place. Bien que des recherches puissent être exécutées avec les moteurs généralistes, il est plus efficace de recourir aux moteurs et aux répertoires spécifiques à la recherche et l'identification de ce type de sources[91]. La littérature recense ici et là ces moteurs et répertoires. Cette recherche particulière fut l'occasion d'établir un répertoire de répertoires dans la pure tradition bibliothéconomique qui est de fournir des listes ainsi que des outils d'identification et de localisation.

9.3.2 Un répertoire d'ontologies spécialisées et les ressources terminologiques des grandes institutions

C'est *BioPortal*, l'un d'entre ces répertoires, qui a été particulièrement exploité. En effet, malgré l'instabilité constatée, il fait référence en raison de l'institution qui le maintient, le *National Center for Biomedical Ontology* (NCBO) soutenu par l'*US National Institute of Health* (NIH). Pour cette même raison il apparaît également comme celui référençant le plus d'ontologies dans le domaine des sciences biomédicales. Par ailleurs, il offre de nombreuses et intéressantes fonctionnalités de recherche et de sélection accessibles à un public non informaticien. De même pour les ressources terminologiques uniquement, c'est-à-dire les ressources non-ontologiques, des répertoires ont été identifiés. Ceux-ci venant en complément des thésaurus des bases de données documentaires préalablement identifiées pour la recherche bibliographique. Il faut y ajouter également les lexiques mis à disposition par les sociétés savantes et par les institutions concernées. D'ailleurs, l'*US Environmental Protection Agency* (EPA) dispose d'un service dédié à la terminologie.

9.3.3 Les ressources internes à la communauté du programme Toxicologie Nucléaire

Cependant, les ressources existent aussi en interne au sein de la communauté ToxNuc. Ainsi le lexique de la communauté ToxNuc présent sur la plateforme a été exploité. Il en a été de même de l'index de l'ouvrage collectif issu des recherches du programme [70, MÉNAGER *et al.*]. Par ailleurs, même s'il n'a pu être exploité durant le stage il existe un fichier résultant d'un traitement linguistique effectué à partir des textes de cet ouvrage. Il serait intéressant de le confronter avec l'ontologie ToxNuc.

---

[90] Pour une présentation de Protégé, se référer à l'Annexe 1- Livrable 3- État de l'art : répertoire des ressources termino-ontologiques disponibles dans le domaine p. 22/34

[91] Pour le détail des méthodes employées pour la recherche de RTO pendant le stage et la description des différents outils, se référer à l'Annexe 1- État de l'art : répertoire des ressources termino-ontologiques disponibles dans le domaine

9.3.4 La méthode de recherche et les critères de sélection

Quelle que soit la source utilisée, la même démarche de recherche a toujours été suivie. Celle-ci peut être qualifiée d'empirique. Elle a été inspirée par les conseils et avis des différents praticiens rencontrés, ainsi que des retours d'expériences fournis par la littérature. En effet, il semble qu'il n'existe pas de méthode gravée dans le marbre. Elle a consisté à partir des termes présents dans l'ontologie ToxNuc, comme par exemple les éléments chimiques, les instruments ou les molécules. Ceci afin de sélectionner les ressources ayant le plus de recoupements avec ToxNuc, pour ensuite les comparer du point de vue de leur structure, tout en confrontant aussi les objectifs pour lesquelles elles ont été conçues [61, VANDENBUSSCHE *et al.*]. Sur ce dernier point, il est parfois difficile d'identifier les auteurs et les intentions qui ont conduit à réaliser ces ressources. Or c'est un élément important pour choisir ou non d'utiliser une ressource. Comme pour toute ressource, l'auteur et sa réputation sont des gages de sérieux et de pérennité d'une ressource. C'est la raison pour laquelle les portails et répertoires ont été privilégiés par rapport aux moteurs. En effet, dans ce cadre, des éléments d'identification sont fournis en plus des ontologies elles-mêmes. C'est ainsi que dix ressources librement accessibles via le Web ont été sélectionnées et cataloguées. Leur catalogage fait l'objet, en partie, d'un des livrables[92] remis à l'issue de ce stage aux chercheurs.

9.4 Les éléments d'identification et de description des RTO

9.4.1 Des ressources directement identifiables ?

Avec l'accès direct aux ressources numériques, penser que la description n'est plus nécessaire est tentant. Aujourd'hui, les ressources numériques sont en capacité de porter elles-mêmes leurs propres métadonnées et d'alimenter automatiquement les catalogues de ressources grâce à l'interopérabilité des données. Autrement dit, en étant embarquées et elles aussi codées comme le reste des données, elles peuvent être interrogées, réagencées dans un autre système comme un portail d'information par exemple. Pourtant, cette expérience de recherche de ressources ontologiques témoigne bien du fait que celles-ci, comme les autres ressources électroniques d'ailleurs, ne sont pas autoporteuses de leurs propres métadonnées. Il manque très souvent des éléments de contextualisation pour ces ressources. Et ceux-ci ne sont ni encapsulés dans la ressource ontologique ni fournis par une documentation annexe sur des pages Web. Plutôt que de se lancer dans une ecdotique ontologique des ressources trouvées, l'idée ici a été d'éliminer toutes les ressources difficilement identifiables. Pour celles qui l'étaient plus facilement, une grille de description a été constituée. Les ressources ainsi décrites devraient permettre aux experts de la toxicologie nucléaire et de l'IC de déterminer de leur intérêt ou non pour compléter ToxNuc. Inversement, cette description devrait aussi permettre d'envisager si ces ressources pourraient être complétées par ToxNuc. Enfin leur recensement et leur description devraient permettre d'évaluer l'impact que pourrait avoir une ressource comme ToxNuc à leur côté.

[92] Cf. Annexe 1- Livrable 3 : État de l'art : répertoire des ressources termino-ontologiques disponibles dans le domaine

9.4.2 Une description qui ne fait pas encore l'objet d'un standard ou d'une norme

Pour décrire des ontologies, il est intéressant de s'inspirer des présentations qui en sont faites sur les différents répertoires. Il est également intéressant de s'inspirer des schémas de métadonnées eux-mêmes. Mais l'idée n'est pas d'aller jusqu'à produire un nouveau schéma les combinant et utilisant les métalangages RDFs ou OWL pour opérationnaliser cette description. Seule une grille classique de catalogage a été élaborée au cours du stage. Elle s'inspire des quinze éléments de base du schéma Dublin Core (DC). S'y retrouvent des métadonnées d'identification, descriptives, de gestion, structurelles et des métadonnées juridiques. Il n'y a pas de métadonnées d'exposition et d'accès aux données de ces ontologies. Pour les informations relatives à ces sujets, un renvoi est fait vers le répertoire ou le site y donnant accès selon différentes voies. Pour une meilleure comparaison entre ces ontologies sélectionnées et ToxNuc, cette dernière a également été décrite selon cette même grille dont voici les éléments :

- attributs des métadonnées d'identification : titre, acronyme, éditeur, date du 1er dépôt, date du dernier dépôt ;
- attributs des métadonnées descriptives : domaine(s) couvert(s), description *BioPortal*, documentation ;
- attributs des métadonnées de gestion : format, recommandation SKOS (c'est-à-dire s'il y a une présence ou non de valeurs issues de SKOS pour rdf:type) ;
- attributs des métadonnées structurelles : nombre de classes, (nombre de) niveaux de profondeur, nombre de relations, nombre d'instances, visualisation *BioPortal ;*
- attribut des métadonnées juridiques : licence d'utilisation.

En creux, ces différents éléments de la grille fournissent des éléments sur la pérennité des ontologies en présence. Cette information n'est en effet pas explicite. Elle résulte de l'interprétation qu'en feront les toxicologues et les chercheurs en IC par rapport à leur connaissance et leur appréciation des autorités de publication. En effet, évaluer leur pérennité au travers de leur notoriété et leur capacité à maintenir les ontologies diffusées, revêt une importance lorsqu'on souhaite lier son modèle à un autre. Par ailleurs, sans pouvoir donner réellement à mesurer la qualité intrinsèque des ontologies en question, certaines des métadonnées structurelles et descriptives leur permettront d'évaluer le niveau de spécificité, de profondeur et de couverture de ces ontologies. En outre, elle permet en partie d'informer de la conformité ou non aux standards afin d'évaluer la capacité d'interconnexion entre les données portées par ces ontologies.

10 L'affinage et les réflexions portées sur l'ontologie

Les corrections portant sur la qualité de la syntaxe et la dimension qualitative apportée par l'ajout de labels ont permis notamment d'amorcer une réflexion et quelques premiers changements au niveau du formalisme de ToxNuc.

10.1 Des corrections portant essentiellement sur la syntaxe

10.1.1 Les outils de références pour effectuer les corrections

Chacun des 646 noms de classe a été vérifié et corrigé. En corrigeant le problème d'encodage de caractère pour les noms de classes, les chaînes de caractères accolées à l'URI pour pointer chaque classe, appelées identificateurs de fragment (fragURI), ont été contrôlées et corrigées systématiquement elles aussi. Cependant, même sans problème au niveau des fragURI, chaque URI a été vérifiée. Sa syntaxe fut rétablie lorsque manquaient ou étaient tronqués le scheme ou l'autorité, en l'occurrence ici respectivement « http : » et « www.cea.fr/ontotoxnuc » (cf. 8.1.2 Du point de vue syntaxique et documentaire). Lorsqu'il manquait des labels, c'est-à-dire les libellés en langage naturel composés de termes permettant d'expliciter le sens des entités désignées par les classes (cf. 2.2 Les caractéristiques des ontologies), ils ont été recherchés et soumis à la directrice du programme Toxicologie Nucléaire. Plusieurs sources ont été consultées à cette occasion :

- le glossaire en ligne sur la plateforme ToxNuc-E ;
- le glossaire, l'index et les articles de l'ouvrage collectif publié dans le prolongement du programme scientifique Toxicologie Nucléaire [70, MÉNAGER et al.] ;
- l'*Energy Technology Data Exchange Thesaurus (ETDE Thesaurus)* de l'*International Nuclear Information System* (Inis) de l'*International Atomic Energy Agency* (AIEA) ;
- le portail terminologique multilingue et multidisciplinaire Termsciences proposé par l'Inist (Institut national de l'information scientifique et technique) ;
- le portail terminologique de santé multilingue HeTOP (*Health Terminology / Ontology Portal*) proposé par l'équipe CISMeF (Catalogage et indexation des sites médicaux de langue française) du CHU de Rouen[93] ;
- l'encyclopédie collaborative en ligne Wikipédia ;
- les services de traduction en ligne *Reverso*, *WordReferenc*e et *Google traduction ;*

L'avantage de consulter les thésaurus et les portails terminologiques, par rapport aux glossaires, est qu'ils ont une structuration arborescente qui donne des éléments de connaissance. Cependant, il faut garder à l'esprit que leur sémantique est plus ou moins légère. Ainsi, il peut y avoir avec un mélange de liens hiérarchiques et de composition (dits aussi relations partitives ou encore tout-partie) non recommandé pour établir la structure taxinomique d'une ontologie.

[93] Pour une présentation d'*ETDE Thesaurus*, Termsciences et HeTOP, se reporter aux pages 13 et 25/34 de l'Annexe 1- État de l'art : répertoire des ressources termino-ontologiques disponibles dans le domaine

10.1.2 L'enrichissement terminologique, une nécessité

De cette manière, les lacunes ont été comblées et les corrections effectuées. Cependant, si la vocation de cette ontologie est d'injecter dans le moteur de la plateforme ToxNuc du vocabulaire présent dans les articles sur la toxicologie nucléaire, elle mériterait d'être enrichie. Plusieurs labels alternatifs pourraient être ajoutés à chacun des noms de classes pour compléter cette ressource sémantico-linguistique de la toxicologie nucléaire. Par exemple, les labels pour les classes d'éléments chimiques une fois complétés de leurs différentes notations scientifiques permettraient de mieux annoter les publications scientifiques.

> « Incertitudes et lacunes terminologiques sont fréquentes dans le champ des systèmes d'organisation des connaissances, alors que le contrôle terminologique devrait être dans ce domaine une préoccupation centrale. » [6, MENON p. 11].

Ainsi, comme le constate Bruno Menon dans l'extrait qui précède et de même que Sylvie Dalbin [55, DALBIN], l'ontologie ToxNuc ne semble pas exempte de ces lacunes. Elle aurait probablement à gagner à s'inscrire dans le prolongement des référentiels terminologiques en usage dans ce domaine comme l'*ETDE Thesaurus* par exemple. Cet intérêt pourrait probablement être stimulé si l'AIEA s'orientait vers l'exposition de son référentiel en RDF. Cependant, les seules informations recueillies à propos des ambitions de l'AIEA pour l'*ETDE Thesaurus* est sa volonté de développer des interfaces utilisateurs multilingues et d'intégrer ce thésaurus à des systèmes d'indexation assistée par ordinateur.[94]

10.2 Quelques réflexions et suggestions à propos de la formalisation

10.2.1 La gestion des polyhiérarchies

La vérification manuelle, et sans requête d'aucune sorte, a permis de détecter également quelques erreurs dans la formalisation. A cette occasion, de nouveaux agencements dans la hiérarchie des concepts ont été suggérés. Par ailleurs, quelques polyhiérarchies ont été détectées. Celles-ci apparaissent parfois dans les thésaurus. Cependant, elles sont à éviter à ce stade de la construction d'une ontologie pour laquelle la méthode suivie tente de respecter celle élaborée par B. Bachimont qui repose sur le paradigme différentiel.

> « Une même unité ne peut avoir qu'une et une seule unité parente. » [35, BACHIMONT p. 311]

[94] D'après les rencontres PIST 2007 à Nancy
< http://rpist.inist.fr/sites/rpist/IMG/pdf/INIS_CEA.pdf >

> « On retrouve ainsi les conditions d'une sémantique pour laquelle les unités signifient toujours la même chose quels que soient les énoncés qui les contiennent. Ce qui n'est donc pas vrai d'un point de vue linguistique doit être imposé par la normalisation sémantique. » [*Ibid*. p. 310]

Signaler ces polyhiérarchies et tenter avec les experts de les annuler a été l'occasion de spécifier de façon encore plus explicite la définition des concepts les uns par rapports aux autres. En général, cette pratique permet de mieux fixer la sémantique des concepts pour un contexte donné. Surtout, cela évite à terme des oublis et des incohérences dans les inférences demandées à l'ontologie. Il est donc préférable que ces polyhiérarchies soient introduites suite à l'activation d'un logiciel appelé raisonneur. Celui-ci « *tire les conséquences logiques des assertions données en entrée* » [79, CHARLET]. En tirant les conséquences logiques des propriétés attribuées aux concepts, il reclasse les concepts. Ce n'est cependant pas un principe qu'il est toujours facile de respecter. A terme, cette démarche pour éviter les polyhiérarchies par simples assertions amènera à spécifier plus précisément les concepts ou les classes avec des propriétés comme des conditions nécessaires et suffisantes. Les polyhiérarchies pourront peut-être alors apparaître après l'activation du raisonneur si certaines notions ne sont pas disjointes. Dans ce cas, cela pose moins de problème pour les évolutions futures de l'ontologie.

10.2.2 Définir des relations supplémentaires entre les concepts

Dans le cadre du stage, il n'a pas été possible d'aller plus loin et d'expérimenter plus avant la méthode indiquée par Sylvie Dalbin et évoquée en 2.4.1 (Construire un modèle et le faire expertiser).[95] Ainsi, des relations inspirées du processus de la toxicité pourraient être définies. Il faudrait ensuite tester les inférences induites par ces relations. Par exemple, des relations pourraient être établies entre les classes et sous-classes désignant les concepts des différentes formes de l'iode avec le concept « Thyroïde ». Cela pourra être fait si la reprise du projet est validée. Et par ailleurs, comme il a déjà été précisé, il est nécessaire pour cela de bien définir la destination de l'ontologie en question. En conséquence, les besoins des toxicologues restent à redéfinir à nouveau quelques années après cette première tentative de représentation formelle des connaissances de leur domaine. Cependant, l'ontologie est considérée comme « *suffisamment propre et bien pensée pour être réellement utilisée* »[96] dans le cadre du projet lié au prototype *Folksonomies*.

[95] Se référer également à l'Annexe 3 - Livrable 5 - Correction de l'ontologie formalisée en OWL
[96] Propos recueillis auprès de Sylvie Ranwez, maître de conférences en Informatique habilitée à diriger des recherches (Université de Montpellier II, LGI2P Ecole des Mines d'Alès), qui a été amenée à travailler ponctuellement pour l'ontologie de la Toxicologie Nucléaire.

10.3 La visualisation comme outil d'aide à la conception, de maintenance et de partage avec les utilisateurs[97]

10.3.1 Visualiser l'ontologie pour la concevoir, la faire évoluer et en assurer la qualité

Après ces modifications et aussi après quelques aménagements de la taxinomie de concepts, cette ontologie n'est plus composée de 646 classes mais désormais de 649 classes. Elle a pu être donnée à voir aux différents interlocuteurs sous différentes formes grâce à l'utilisation de l'outil *SKOS Play !*. C'est un service de visualisation de thesaurus, taxo/inomies ou vocabulaires formalisés selon la recommandation SKOS[98]. Comme cet outil permet également de transformer des fichiers OWL en SKOS, il a été utilisé pour visualiser l'ontologie ToxNuc.

En réalité, cet outil dédié à la visualisation a permis de plus facilement détecter les erreurs du fichier OWL sans avoir de connaissances en programmation informatique. En effet, des requêtes SPARQL (*Simple Protocole and RDF Query Language*) auraient pu être exécutées afin de détecter certains types d'erreurs, comme celles liées aux annotations évoquées ci-dessus. De même l'utilisation d'un plugin pour la visualisation sous la forme de graphe dans *Protégé* permet probablement de repérer des erreurs ou du moins les polyhiérarchies. Mais le manque de connaissances informatiques a conduit à explorer les possibilités offertes par *SKOS Play !* afin de poursuivre ce repérage et le travail de correction. Celui-ci a consisté en une succession d'allers et retours dans *Protégé* pour effectuer les modifications du fichier OWL. *SKOS Play !* a donc été très utile pour améliorer la qualité du fichier OWL sans qu'il soit nécessaire de maîtriser aucun de ces deux langages SKOS et OWL.

Ainsi, cet outil a permis d'amorcer le dialogue pour susciter des retours de la part des toxicologues sur les labels et les relations hiérarchiques de l'ontologie actuelle. Enfin il a permis de générer très facilement des listes et des index, qui sans cela, n'auraient probablement pu être présentés à l'issue du stage.

---

[97] Ce titre revient à Sylvie Dalbin suite à un échange autour de *SKOS Play !*

[98] Il s'agit d'un service en ligne gratuit. Il est mis à disposition par le consultant et formateur Thomas Francart depuis son blog (http://blog.sparna.fr) < http://labs.sparna.fr/skos-play >.

10.3.2 Partager des vues pour partager un langage commun, partager des vues pour faire connaître ce langage

Le dialogue avec les utilisateurs a pu être amorcé grâce aux différentes vues générées avec cet outil. Il s'est révélé intéressant pour dialoguer avec les interlocuteurs du projet situés à distance les uns des autres et loin du laboratoire d'accueil pour ce stage. En effet, cet outil ne nécessite pas d'être installé comme le logiciel *Protégé* car il est utilisable directement en ligne sur le Web. Il a aussi pour avantage d'avoir une interface conviviale. La navigation en ligne peut aussi être considérée comme plus facilement manipulable que des listes et index sous forme de fichiers PDF que *SKOS Play !* est par ailleurs aussi en capacité de fournir. En revanche ces listes sont intéressantes à l'issue du stage pour accompagner le fichier OWL corrigé et en présenter une version stabilisée.

En outre, rappelons qu'à l'origine cette terminologie se présentait sous la forme d'un tableur car informatisée avec le format Excel®. Ce test est donc aussi l'occasion de constater que le choix d'opérationnaliser la terminologie en un fichier OWL la rend facilement manipulable depuis le Web. Surtout ce test démontre qu'elle est conforme aux standards et qu'elle possède déjà certains éléments lui permettant d'être transférable dans l'univers du Web de données.

Ce type d'outil pourrait être envisagé comme un outil de management visuel et comme un outil de promotion supplémentaire :

- aide à la conception en cas de reprise du projet pour représenter les hiérarchies ;
- aide à l'instauration du dialogue avec les utilisateurs non informaticiens participant à sa conception, ou à sa maintenance, pour détecter les anomalies ;
- aide à la navigation hypertexte dans la terminologie pour indexer des contenus ;
- aide à la valorisation de l'ontologie ;
- aide pour susciter du *feedback* de la part des utilisateurs externes en cas de mise à disposition ;

Les usagers sont de plus en plus habitués à visualiser des données. Les outils utilisés dans le contexte des activités professionnelles doivent également être attrayants. Il est préférable que ceux-ci correspondent aux habitudes des utilisateurs. Si ces interfaces font sens pour eux, il est permis de penser qu'elles susciteront une plus grande réactivité de la part des différents participants.

11 Quelques éléments et suggestions pour redémarrer le projet d'ontologie ToxNuc

L'étude d'opportunité remise en fin de stage s'est surtout attachée à dresser le bilan de l'ontologie dans son état actuel avec ses potentialités stratégiques de valorisation pour le CEA dans le domaine de la Toxicologie Nucléaire et Environnementale. La matrice de la **Erreur ! Source du renvoi introuvable.** est une des synthèses extraite de l'étude d'opportunité. Cette dernière a également fait part de l'environnement externe à partir du *benchmark* des RTO disponibles. Y sont également mentionnés des projets similaires pour des disciplines connexes comme la toxicologie prédictive et les nanosciences. L'étude comporte enfin une synthèse des informations recueillies sur les pratiques et les méthodologies adoptées par d'autres établissements de recherche, ainsi que sur les potentialités de mutualisation dans l'environnement du Web sémantique. Cependant, l'environnement interne, avec les motivations mais aussi les résistances, ainsi que les risques en présence mériteraient d'être analysés davantage pour élaborer la revue d'opportunité qui permettra de décider de l'abandon ou de la redéfinition du projet. En cas de reprise, les lignes qui suivent s'efforcent de délimiter le périmètre potentiel de cette ontologie. Elles tentent aussi de lister un certain nombre de préconisations.

11.1 Les prérequis pour pérenniser et valoriser la publication de ToxNuc

11.1.1 Envisager les aspects juridiques avant tout

Parmi les besoins exprimés, celui de la mise à disposition sur le Web de cette terminologie de la toxicologie nucléaire et l'étude de sa réutilisation semble être l'objectif principal. Et cela même avant celui de son utilisation effective au sein d'un outil interne ou de la plateforme ToxNuc. S'assurer de sa réutilisabilité est donc primordial avant même de garantir techniquement sa réutilisation. En effet, comment savoir si d'autres parviennent à l'utiliser techniquement si le droit ne leur en est pas déjà donné ? De plus la licence choisie peut permettre de tracer cette réutilisation. La publication d'une telle ressource accompagnée d'une licence d'utilisation est en outre connotée de sérieux et d'ouverture. La raison en est qu'en plus d'être une sécurité juridique pour les réutilisateurs, elle traduit en effet la réflexion portée sur cet objet au plus haut niveau de l'institution, soit ici du CEA, et sur la prise de conscience des obligations implicites liées. Ce mémoire abordera cet aspect sensible dans le paragraphe suivant. Enfin et surtout, il est nécessaire de s'inquiéter de cet aspect juridique en amont même des aspects techniques, eu égard aux enseignements d'autres projets [54, BERMÈS]. En effet, d'autres expériences de publications de données ou de schémas de données ont montré que celui-ci pouvait être un point de blocage. Et cela alors même qu'un long travail avait par ailleurs été mené pour les rendre techniquement valides. Pour rappel, le programme Transversal Toxicologie Nucléaire a pour tutelle le CEA. Pour cette institution, cela semble être encore davantage

un point dont il faut s'assurer avant d'entamer une autre initiative. Le stage a été l'occasion d'évoquer cet aspect et il semble que le CEA soit favorable à l'adoption d'une licence de type Creative Commons.

11.1.2 Les aspects techniques et communicationnels

Si le souhait est de publier cette ontologie pour observer si d'autres communautés s'en saisissent, une définition fine de la connaissance du domaine augmente son attractivité. Mais selon [61, VANDENBUSSCHE *et al.*], une expressivité trop contrainte et une granularité importante pourraient freiner la réutilisation de l'ontologie. Ne pas la spécifier davantage éviterait aussi de solliciter à nouveau les experts. Cependant, même pour un projet à minima comme celui-ci, l'économie de certains questionnements et investissements serait dommageable pour le succès du projet.

Au niveau technique, certains aspects devront être recensés et chiffrés par l'étude de faisabilité. Il s'agit de ce qui a trait par exemple à l'hébergement sécurisé de l'ontologie et de ce qui garantit son accessibilité de façon pérenne. En outre, la publication de l'ontologie engage l'image de l'institution. Il est alors recommandé de l'éditorialiser, c'est-à-dire de l'accompagner d'une présentation en plus de sa documentation. Dès lors, le référencement naturel serait renforcé et l'ontologie gagnerait en audience par l'accompagnement d'une documentation régulièrement mise à jour. La documentation est par ailleurs générable à partir du fichier OWL (cf. Annexe 2 - Livrable 4). Cette dernière complèterait la documentation davantage destinée aux machines car portée au sein même du fichier OWL par l'utilisation des éléments de métadonnées Dublin Core par exemple. La documentation pourrait présenter cette ontologie au sein des pages Web publiques de la plateforme ToxNuc. Y seraient accessibles les articles scientifiques à propos de cette réalisation dans le cadre du programme Transversal Toxicologie Nucléaire. Par ailleurs, rassurer sur l'actualisation programmée ou régulière peut encourager les potentiels réutilisateurs. Vu l'importance grandissante des *altmetrics*, investir les réseaux sociaux numériques (RSN) serait aussi une forme de valorisation à envisager. En revanche, des outils de *feedback*, comme ne serait-ce qu'un courriel, devraient être envisagés pour que les utilisateurs (futurs collaborateurs ?) puissent signaler des erreurs, des lacunes et suggérer des améliorations.

11.1.3 Les aspects organisationnels et la gouvernance

Pour corriger l'ontologie ou réorienter ces demandes et pour y répondre il est nécessaire d'assigner cette responsabilité à une personne en particulier. De même, assurer la publication des nouvelles versions et documenter les deltas entre elles, en plus d'assurer une veille des usages, nécessite disponibilité et compétences. Cela permet de maintenir le dynamisme de cette ressource et la confiance de ses potentiels réutilisateurs.

Certes l'évolution du Web vers un Web encore plus social et interconnecté renforce les potentialités de mutualisation et de répartition de l'effort pour faire évoluer cette ressource. Mais cela n'en réduit pas les charges de maintenance, bien au contraire. Ainsi même si l'ontologie de la Toxicologie Nucléaire est envisagée à minima comme une terminologie de ce domaine, sa publication nécessite de coordonner plusieurs actions. Ainsi les aspects techniques, informatiques, juridiques,

organisationnels et informationnels sont à piloter avant, pendant mais également après le développement. Pour l'après mise en œuvre, il s'agit de concevoir un ou des scénarios de continuité. Pour reprendre les propos d'Éric Dagiral lors de l'entretien réalisé au sujet de l'ontologie conçue au sein de la Plateforme Maladies Rares : « La technique n'est pas purement technique. C'est une imbrication de social et de technique » [cf Annexe 4, p. 6/6].

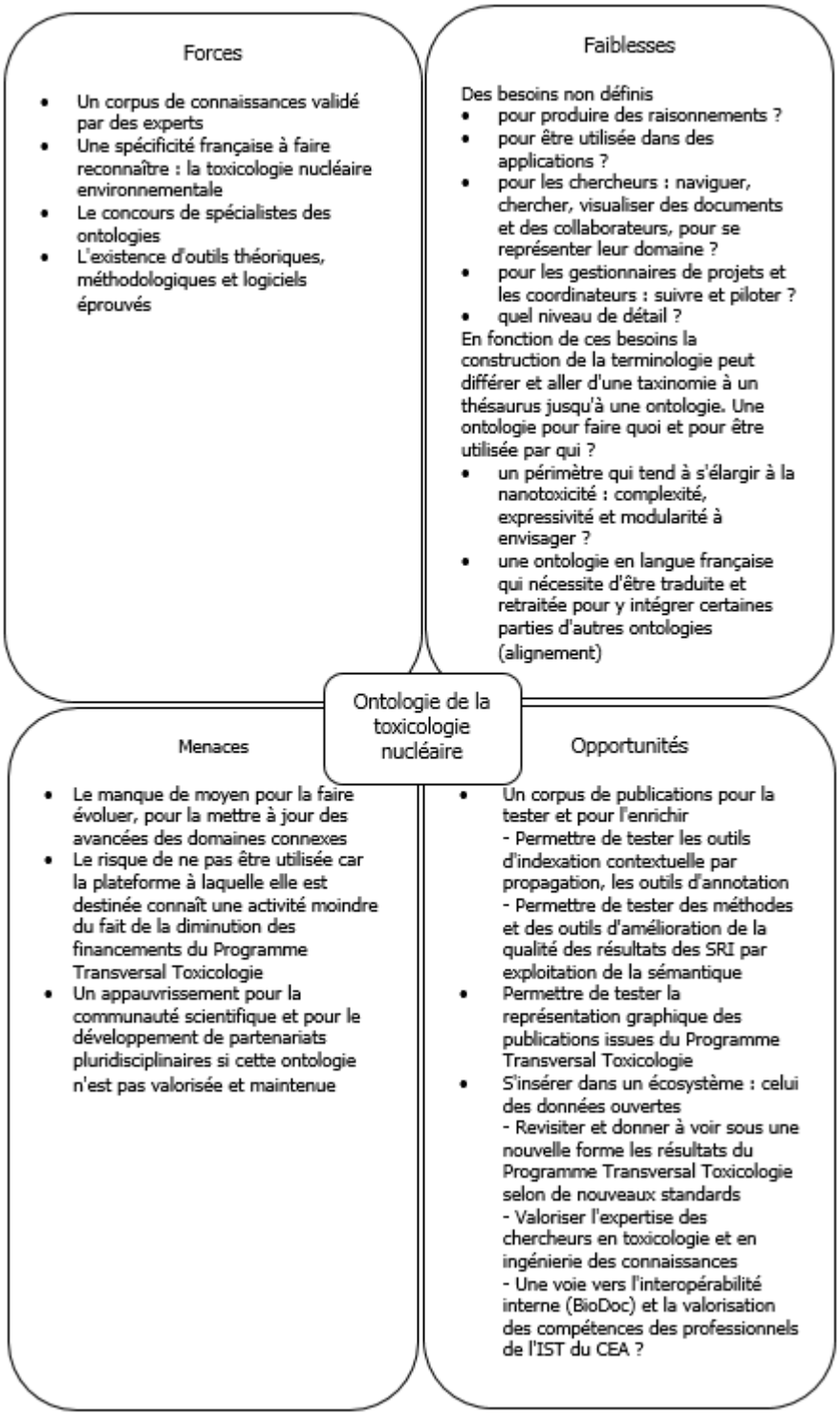


Figure 10 : la matrice SWOT appliquée à l'ontologie de la Toxicologie Nucléaire

11.2 Le scénario orienté publication d'un modèle conceptuel du domaine de la toxicologie nucléaire

11.2.1 Définir les relations entre les entités du domaine

Si la volonté est de modéliser la Toxicologie Nucléaire ainsi que la façon dont est pratiquée l'étude des phénomènes observés et mesurés par cette discipline, l'ontologie actuelle nécessiterait d'être complétée. D'une part, il faudrait définir réellement les entités de ce référentiel sémantico-linguistique, c'est-à-dire distinguer les classes les unes des autres selon le paradigme différentiel développé par B. Bachimont. D'autre part, il faudrait formaliser les relations que les concepts entretiennent les uns avec les autres. Enfin, tout ce qui a trait aux nanosciences et à l'environnement du point de vue de la Toxicologie Nucléaire est encore à ajouter. Cela permettra d'être en adéquation avec les orientations du programme prises ces dernières années. Pour les chercheurs de ce domaine, plus que de partager un vocabulaire, elle aurait la vertu de « systématiser l'expression des connaissances qu'ils manipulent » [27, BACHIMONT]. Ce référentiel serait le témoin du degré de maturation et de consolidation de ce domaine de recherche.

11.2.2 Faire entrer le modèle dans le Web de données

Afin que cette discipline où la France excelle soit visible dans le Web de données à travers la publication de son référentiel, une réflexion supplémentaire devrait être menée. Elle est également nécessaire dans une perspective d'interopérabilité des jeux de données que la communauté ToxNuc souhaiterait éventuellement publier à l'avenir.

D'une part, cette réflexion peut consister à déterminer quels schémas de métadonnées réutiliser. Il pourrait au minimum s'agir du plus répandu comme Dublin Core pour présenter et aider à l'identification de ToxNuc (auteur, éditeur, source…). La réutilisation des schémas de métadonnées existants est à envisager plus particulièrement pour aider à l'identification de ToxNuc par ses potentiels réutilisateurs. Et comme ces schémas existent et sont de plus en plus répandus, il serait dommage de ne pas s'en servir pour indiquer les dates, le n° de version ou encore le type de licence attribuée à ToxNuc par exemple. D'autre part, il peut s'agir de choisir quels concepts appartenant aux « vocabulaires de valeurs » (cf. 3.4 Ontologies versus SKOS) répertoriés durant le stage. Leurs termes seraient à utiliser comme des instances pour ToxNuc. Ou bien alors, après avoir instancié ToxNuc avec ses propres valeurs, il conviendrait de choisir avec quels concepts d'autres ressources établir des liens sémantiques. C'est-à-dire qu'il faudrait choisir avec quels concepts et valeurs d'autres référentiels les aligner (*matching*). Ces choix sont guidés généralement par la confiance accordée aux institutions productrices et gestionnaires de ces ressources en fonction de leur notoriété. Si la communauté ToxNuc, et donc le CEA, utilise des schémas de métadonnées et des « vocabulaires de valeurs » largement répandus, les données publiées par ce biais seront mieux à même d'être interprétées. Enfin, pour être considéré comme en capacité de supporter l'interconnexion des données, le principe de la modularité devrait être envisagé. Ainsi l'étude réalisée dans le cadre du projet DataLift démontre que le caractère réutilisable diminue si la taille et le niveau de complexité augmente [61, VANDENBUSSCHE *et al.*].

11.3 Le scénario orienté recherche d'information et portail sémantique

11.3.1 Enrichir la terminologie : une nécessité

Dans une optique applicative avec pour objectif d'améliorer la recherche des publications stockées dans la plateforme Toxnuc, compléter l'ontologie actuelle semble être une nécessité. Plus la terminologie sera riche et meilleur sera le taux de rappel une fois-celle-ci injectée dans le moteur de recherche. Les termes contenus dans l'ontologie peuvent aussi être utilisés pour guider la recherche grâce à l'autocomplétion au moment de la saisie. Quels pourraient être ces compléments ? L'ajout des synonymes pour chacun des concepts recensés augmenterait son potentiel. De plus, lui ajouter des concepts relatifs aux nanosciences et aux sciences de l'environnement lui permettrait d'être en adéquation avec les publications pour aider à mieux les retrouver. L'enrichir avec les variantes des notations scientifiques servant à désigner les éléments chimiques lui permettrait d'être en concordance avec les différents usages. Par ailleurs, un outil de *Text mining* pour annoter les publications et donc les indexer automatiquement nécessite une ressource sémantico-linguistique riche. Enfin, comme il a été vu précédemment, les classes de l'ontologie pourraient servir à classer les publications et à les présenter sous des catégories ou sous des tags si l'ontologie est combinée avec l'outil *Folksonomies*. Il pourrait également être envisagé de classer avec ces mêmes catégories les références contenues dans la base BioDoc signalant les publications des différents programmes de la DSV du CEA. [45, DELAHOUSSE ; 46, FRANCART.]

11.3.2 Tester le Metadata mining avec la base BioDoc

Comme il a été abordé en 8.3 (Les dimensions organisationnelle et technique), les métadonnées de la plateforme ToxNuc sont difficilement exploitables en raison d'une saisie non normalisée de celles-ci. Plusieurs solutions sont envisageables pour corriger ces métadonnées. Elles vont de la plus basique, comme la reprise de la saisie de façon manuelle et rétrospective, jusqu'aux plus sophistiquées comme l'utilisation d'outils d'analyse textuelle. Certains sont en effet capables de reconnaître la structure d'une publication scientifique et d'en extraire le titre et les auteurs. Par ailleurs, récupérer ces métadonnées via les différentes plateformes sociales allant des LGB aux RSN de chercheurs semble hasardeux. De même, interroger les références recensées par *Google scholar* ou celles des grandes bases de données des sociétés *Thomsom Reuters* et *Elsevier* semble autant délicat techniquement que juridiquement[99]. Mais dans tous les cas, des difficultés liées à l'hétérogénéité des métadonnées sont à prévoir.

[99] LANGLAIS P.-C. *Data mining : l'Europe s'oriente vers une exception* [En ligne]. *Sciences communes*. 13 mai 2014a. Disponible sur : < http://scoms.hypotheses.org/208 >
LANGLAIS P.-C. *Faire du data mining avec Google : comment tromper big brother ?* [En ligne]. *Sciences communes*. 24 mai 2014b. Disponible sur : < http://scoms.hypotheses.org/216 >
LANGLAIS P.-C. *Text mining : vers un nouvel accord avec Elsevier* [En ligne]. *Sciences communes*. 29 octobre 2014c. Disponible sur : < http://scoms.hypotheses.org/276 >

En revanche, utiliser la base BioDoc alimentée par les documentalistes de la DSV du CEA pourrait peut-être garantir une homogénéité de ces métadonnées, contrairement aux précédentes solutions proposées. D'une part la source serait unique. D'autre part cette base est alimentée uniquement par les références de la base bibliométrique *Web of science* et cette alimentation automatique est contrôlée par les documentalistes. Peut-être serait-il intéressant de mettre en balance les différents modes de récupérations des métadonnées précédemment évoqués avec cette tentative de faire de BioDoc un laboratoire de fouille de métadonnées. Il s'agit notamment de profiter des richesses internes que sont la connaissance des documentalistes de la constitution et de la structuration de cette base. Leur maîtrise du signalement pour vérifier les métadonnées ainsi attribuées aux publications pourrait également être précieuse. En effet, quelles que soient les méthodes envisagées, aucune ne peut faire l'économie d'une vérification humaine.

11.4 Le scénario orienté base de connaissance et toxicologie prédictive

11.4.1 La toxicologie nucléaire : quelques éléments de compréhension

Avant d'évoquer ce que pourrait être une base de connaissance de la toxicologie nucléaire, voici une présentation de cette discipline à la croisée de la physique, la chimie, la biologie et la médecine.

Les êtres vivants sont exposés aux agents chimiques et aux rayonnements émis par des éléments radioactifs présents à l'état naturel ou provenant d'activités industrielles. Si la diversité de ces éléments est grande, il en est de même de leurs formes physico-chimiques ainsi que de leur niveau de concentration.

La recherche en toxicologie nucléaire a pour but d'identifier, de quantifier et de comprendre les mécanismes moléculaires et cellulaires de la réponse d'un organisme vivant lors de son exposition chronique, fractionnée ou aiguë à des niveaux de doses variés, mais aussi selon la forme physico-chimique, la voie d'administration ou d'absorption de ces éléments. Il y a toxicité lorsqu'il y a engendrement de troubles, temporaires ou non, de certaines fonctions à divers niveaux d'organisation du vivant. Pour les substances chimiques, on parle de chimiotoxicité et de radiotoxicité pour les substances radioactives.

Quels sont les effets chimio et radiotoxiques de ces divers éléments sur les êtres vivants ? L'interdisciplinarité par la mise en commun des connaissances et des approches méthodologiques et expérimentales des diverses disciplines est nécessaire pour répondre à cette question. La toxicologie nucléaire environnementale et humaine est donc un champ pluridisciplinaire dont les avancées devraient également servir de bases à la toxicologie prédictive pour contribuer à l'évaluation et à la gestion des risques environnementaux et sanitaires. En effet, les champs d'application de la toxicologie et de l'écotoxicologie nucléaire sont multiples et complémentaires entre eux. Ils vont de la détection aux traitements des contaminations, depuis l'environnement jusqu'à l'Homme. [70, MÉNAGER *et al.*]

11.4.2 Les publications du programme ToxNuc pour exprimer des connaissances et l'ontologie ToxNuc pour les représenter

Une base de connaissance est l'une des finalités possibles d'une ontologie. Elle peut être définie comme une base contenant des connaissances, mais aussi comme étant aussi à la fois capable de produire de nouvelles connaissances[100]. L'apprentissage automatique pour l'acquisition de connaissances nécessite des développements complexes et coûteux. Là encore, le recours au TALN est nécessaire. Il sert pour implémenter des règles de raisonnement basées sur l'analyse linguistique et grammaticale. Ce sont elles qui permettent de repérer et d'indexer de nouvelles entités ainsi que des informations à leur propos lorsque des textes sont soumis au système d'information. Ainsi l'ontologie est une brique au sein d'un système complexe. Son rôle est de fournir les faits et les affirmations déclarés par les experts sur les entités introduites dans l'ontologie ToxNuc. C'est ce qui permet aux nouvelles données identifiées de venir correctement la peupler. Au sein de la base de connaissance, ce modèle conceptuel du domaine est alors mis à l'épreuve du réel décrit dans les publications. C'est la raison pour laquelle la définition des relations est délicate pour éviter que trop de contraintes empêchent d'agréger de nouvelles données.

Si, l'ontologie à elle seule ne peut constituer une base de connaissance, en tant qu'ossature elle permet d'élaborer une base de connaissance diversement orientée selon la façon dont elle est conçue. Elle pourrait être orientée KM, c'est-à-dire capitalisation des connaissances produites au cours du programme Transversal Toxicologie Nucléaire. Il faudrait alors lui ajouter et spécifier des entités supplémentaires comme « Projet », « Membre_projet » afin de savoir qui a produit quoi par exemple. Autrement, elle pourrait être orientée vers la veille. Dans ce cas il faudrait l'instancier avec un vocabulaire plus large que celui de la toxicologie nucléaire. Pour être compatible dans l'univers réticulaire du Web de données et ainsi optimiser l'échange de données, il existe deux possibilités. Soit le système aurait avantage à ce que ce vocabulaire soit aligné avec ceux utilisés par d'autres sources de données. Soit il gagnerait à directement utiliser ceux-ci en déclarant les types et les URI de ces derniers. Cette orientation donnée à l'ontologie ToxNuc permettrait de promouvoir le potentiel instrumental de la modélisation d'un domaine et des ontologies informatiques auprès des toxicologues. Elle rendrait aussi possible la mise à jour du modèle conceptuel défini en 2006 afin de restituer les nouvelles orientations prises au cours du temps par les recherches en toxicologie nucléaire. Etant donné la multiplicité et la diversité du champ de connaissances de la toxicologie nucléaire et environnementale, la mise en œuvre ou le déploiement d'une plateforme dotée d'outils hautement performants semble amplement justifié.

[100] HATON J.-P., HATON M.-C. « Systèmes à bases de connaissances ». *Techniques de l'ingénieur. Informatique*. 2012. n°H3740, p. 18.

11.5 Les ontologies informatiques : de nouvelles possibilités et des points de vigilance

Chacun de ces scénarios nécessite des compétences en informatique et en ingénierie des connaissances mais pas seulement. La proximité avec des toxicologues est également nécessaire pour éliciter la connaissance du domaine et valider les raisonnements produits. Un accompagnement du changement est donc nécessaire avec le soutien d'une volonté politique affirmée. En plus de cette volonté, il peut être nécessaire de créer de l'urgence pour reprendre et faire aboutir le projet. Cependant, le rapport coût/efficacité dans une période de resserrement budgétaire doit être mis en regard des gains attendus. Ces derniers sont d'ordre symbolique comme l'image du CEA. Les potentielles synergies entre communautés scientifiques et les diverses mutualisations sont également envisageables. Tous ces éléments font que ce type de projet n'est pas purement technique et informatique. En cas de non réalisation, des démarches pourraient malgré tout être entreprises. D'une part l'ensemble des réflexions et travaux menés autour du projet d'ontologie pourrait être capitalisés pour éclairer des projets similaires à l'avenir. D'autre part, il pourrait s'agir d'améliorer le signalement des ressources de la base documentaire ToxNuc devenue Toxcea. Ceci afin de faciliter les migrations futures et de garantir l'accès pérenne aux publications par de meilleures conditions d'archivage. Des bénéfices pourraient d'ailleurs en être retirés pour améliorer la gestion de contenu d'autres programmes de recherche du CEA à l'avenir.

Conclusion

Le Web et les technologies qui le sous-tendent ont profondément transformé les pratiques des utilisateurs et irrigué les solutions fonctionnelles proposées aux différentes organisations, qu'elles soient ouvertes ou fermées au Web [66, ECONOCOM-OSIATIS.]. Les ontologies qui combinent technologies issues de l'intelligence artificielle (IA) avec aujourd'hui les standards du Web en sont un exemple. Du point de vue des professionnels de l'I&D, gestionnaires d'information et plus largement infomédiaires, une ontologie était structurellement appréhendée sous l'angle terminologique. Dans l'édition de 2004 du *Dictionnaire de l'information*, elle était définie comme « *une représentation des connaissances terminologiques relatives à un domaine, agréée par une communauté de personnes* » puis associée au Web dans la réédition de 2008 [2, CACALY ; 3, LE COADIC]. Fonctionnellement, elle était vue comme « *un outil de recherche permettant de dépasser les limites des systèmes documentaires classiques, en particulier grâce à un enrichissement de la sémantique des relations et de la combinaison des mots clés* » [22, DALBIN].

A la lecture de la section 21 consacrée aux ontologies de la nouvelle norme ISO 25964-2:2013 sur l'interopérabilité des thésaurus avec les autres SOCs, il est manifeste que les professionnels de l'I&D ont pris acte de la définition dont se reconnaissent les spécialistes de l'IC. En effet, c'est la définition de T. Grüber élargie avec celle de W. Borst qui est reprise. Elle y est donc définie comme une *« une spécification formelle et explicite d'une conceptualisation partagée »*. Au niveau fonctionnel, il y est reconnu que les ontologies ont pour but de procéder à des raisonnements et non à de la recherche d'information même si elles peuvent y contribuer. Il est indiqué également que les thésaurus ne sont pas des ontologies bien qu'ils puissent être assimilés, dans certaines situations, à des ontologies légères [8, ISO 25964-2 pp. 72-78].

Face à de telles promesses de recherche plus intelligente et face à l'engouement pour le Web sémantique, il est apparu intéressant de se pencher plus avant sur les difficultés rencontrées par ceux qui tentent l'aventure ontologique et les stratégies mises en place pour dépasser ces difficultés.

Il était donc nécessaire d'observer dans quelle mesure les ontologies qui apportent fluidité et pertinence dans les systèmes d'information peuvent être conçues, maintenues et pérennisées. Mais avant de tenter de vérifier l'hypothèse émise pour trouver une réponse à cette problématique, une première partie a été consacrée à définir et à repositionner les ontologies dans la typologie des SOCs.

La posture induite par cette problématique tendait à aborder les ontologies et les technologies sémantiques dans leur histoire et leur environnement. Plus que sous l'angle de leurs mécanismes internes et de la démonstration du « comment ?», c'est la mise en exergue du positionnement des SOCs les uns par rapport aux autres qui a été visée. Cependant, même si la volonté était de se garder de reproduire les schémas habituels et massivement disponibles sur le sujet, il a été difficile de s'affranchir des aspects structurels et techniques. Et ceci d'autant plus que l'exercice a révélé la difficulté de délimiter l'objet face à une masse énorme de représentations proposées par les différents acteurs concernés. Ainsi le travail pour démêler l'enchevêtrement des définitions a pris une place plus grande qu'imaginée initialement. Cela a amené à s'intéresser également aux aspects économiques des SOCs dans un écosystème en pleine recomposition avec l'émergence de plus en plus prégnante du Web de données.

Après cette première démarche de définition, il apparaît que « *le mot « ontologie » reste un peu intimidant* » [61, VANDENBUSSCHE p. 7]. Au passage, il semble aussi que la définition et l'histoire du Web sémantique ne fassent pas l'unanimité. Et cela bien que diverses initiatives et réalisations se font jour de façon croissante depuis la fin des années 2000. Ces initiatives tirent des bénéfices de l'utilisation d'ontologies informatiques. Et cela quel que soit le secteur, de la bioinformatique à la médecine, jusqu'au domaine des données culturelles en passant aussi par le secteur des services marchands. Cependant, non directement visibles et relativement longues à élaborer, les ontologies sont difficiles à promouvoir. Les diverses raisons en sont la difficulté à mesurer directement leur retour sur investissement car elles s'insèrent dans des dispositifs numériques plus vastes. C'est donc la promotion de ces dispositifs plus vastes et leur soutien par les décideurs qui permettront de bénéficier des ontologies informatiques et des technologies sémantiques. Il serait intéressant de continuer à observer la diversité des projets dans lesquelles ces technologies continueront à s'insérer.

L'étude du cas exposée dans un second temps permet de répondre en partie à la problématique globale pour laquelle l'hypothèse postulant que la dimension organisationnelle d'un tel projet prend une part non négligeable par rapport à celle prise par la dimension technique avait été émise. Ceci a en partie été confirmé par la reconstitution de l'historique du projet d'ontologie du domaine de la toxicologie nucléaire. Les témoignages recueillis lors d'entretiens sollicités auprès de divers spécialistes pour réaliser une comparaison avec d'autres projets (*benchmark*) participent aussi à ce constat.

Les enjeux qui s'imposent aux organisations et autres collectifs sont comme bien souvent stratégiques et politiques avec aujourd'hui un impératif de mise en visibilité accrue de leurs activités et des données qu'ils produisent. Il en est ainsi pour les chercheurs du secteur public dont l'évaluation repose sur la publication et notamment la publication en collaboration avec des chercheurs extérieurs

à leur propre laboratoire. Il est à souligner encore à ce propos que la collaboration est aussi l'une des clés de génération de nouvelles connaissances.

L'approche par l'outillage et les nouvelles technologies de l'information et de la communication sont elles aussi représentatives des choix actuels pour mettre en mouvement les membres d'une organisation. Et cela d'autant plus pour aider au management des équipes dispersées, pluriorganisationnelles, et pour lesquelles la transversalité est souhaitée. Pour autant, les usages collaboratifs, la transformation des comportements n'opèrent pas automatiquement avec la mise en place de plateformes collaboratives.

Ce retour d'expérience sur la création et l'usage d'outils de soutien à la recherche scientifique met en lumière la nécessité d'appréhender la conception d'une ontologie comme un processus. Ainsi son cycle de création, de réadaptation et d'expansion peut être assimilé à celui de tout projet de conception logicielle et en partie à celui de la création et du développement d'un thésaurus. L'idée de processus induit qu'un soin soit apporté pour que chaque étape du cycle de vie des ontologies informatiques permette de passer à l'étape suivante.

A l'échelle du cas d'étude observé, le *benchmark*, tout comme l'étude et l'historique de l'ontologie de toxicologie nucléaire, tendent à démontrer que la démarche projet et sa dimension organisationnelle peuvent, par leur absence, entraver la conception d'une ontologie.

Mais au-delà même de cette confirmation à propos du processus de construction d'une ontologie, retracer l'histoire du dispositif dans lequel elle doit s'insérer témoigne aussi que le manque d'accompagnement peut avoir au moins trois conséquences :

- l'absence de contributions et d'échanges au sein d'une plateforme ayant une composante documentaire ;
- des difficultés d'accès aux publications ;
- des risques pour la pérennisation de celles-ci.

Au sein de la démarche projet donc, le soin et les moyens accordés à la communication, l'accompagnement et l'animation de la communauté à laquelle est destiné ce dispositif sont prégnants. En effet, une médiation humaine semble être nécessaire afin que la médiation numérique, elle, puisse être effective puis optimale. Les risques sont sinon ceux d'un dispositif en inadéquation avec les besoins comme par exemple l'usage subi, l'usage déficient ou bien encore le non-usage. Mais en contrepartie, cette activité de communication et d'animation demande d'y consacrer du temps. Elle devrait être envisagée comme une activité à part entière et une fonction reconnue. Sinon, elle risque de ne pas être réattribuée à la suite d'un départ d'un ou des personnels qui auront pris spontanément et sans réelle affectation cette charge d'animation. C'est un investissement pour l'amélioration des outils et donc aussi pour l'amélioration des services rendus à la communauté.

Etendre la médiation à l'implémentation de l'ontologie et de l'outil *Folksonomies* dans la plateforme ToxNuc devenue Toxcea participerait à l'adhésion à ces technologies pleines de promesses. Une voie d'harmonisation devrait aussi être trouvée entre les usages potentiels et fantasmés par les spécialistes des technologies numériques mais aussi par les usagers d'une part, et les usages réels d'autre part afin de limiter l'effet déceptif. Les aspects basés sur la spécification et la formalisation informatique d'un domaine ne doivent pas laisser perdre de vue la dimension organisationnelle et les usages.

Par ailleurs, l'expérience a montré qu'il est indispensable de faire de même pour les ressources et les données elles-mêmes. Effectivement, ce sont elles qui sont exploitées par les ontologies. « *Les données numériques sont très fragiles* »[101] pour des raisons d'obsolescence rapide des formats. Le collectif a d'ailleurs pu constater que le stockage de documents PDF sur une plateforme ne suffit pas. Garder la mémoire d'un programme scientifique va plus loin que des questions d'espace de stockage, de serveurs ou de choix de formats de fichiers pour les publications et autres ressources stockées.

Plus largement, la stratégie doit englober les questions relatives à l'archivage des publications mais aussi des outils créés pour l'occasion. Ceux-ci sont assimilables à ce que l'on appelle aujourd'hui les données de la recherche[102]. La gestion des évolutions de ces outils numériques et les évolutions du vocabulaire d'un domaine sont aussi des traces qui participent à écrire l'histoire d'une discipline. Encore faut-il que ces traces soient exploitables. La traçabilité doit être assurée par la présence d'informations sur le contexte de création. Ces informations, autrement dit ces métadonnées, doivent être de qualité. Sans cela, les technologies sémantiques ne pourront rien en faire. Il est contreproductif d'insérer des connaissances mal structurées dans les processus de calcul. « *La gestion de l'ensemble de ces ressources n'est pas pensée par les chercheurs au moment de la création* »[103]. En la matière, les professionnels de l'information scientifique et technique (PIST), documentalistes et autres infomédiaires des organismes de recherche peuvent seconder les chercheurs.

[101] D'après les propos d'Esther Dzalé Yeumo Kaboré (INRA) et d'après la diapositive 18 de la présentation à la journée d'étude Médiadix « Mémoires numériques – Publics, ressources et bibliothèques en mutation » le 10 octobre 2014< http://mediadix.u-paris10.fr/brochure/voir_fiche.php?Id_stage=1328&st_niveau=02 >
[102] FAYET S. *« Données » de la recherche, les mal-nommées* [En ligne]. *URFIST Info*. 15 novembre 2013. Disponible sur : < http://urfistinfo.hypotheses.org/2581 >
[103] Cf. note n° 91, diapositive n° 20

Concernant les compétences nécessaires, ces professionnels sont parmi les plus conscients de l'importance de l'usage des métadonnées structurées et normalisées, au sein de structures de recherche comme au sein des entreprises[104]. En effet, ils ont une maîtrise de la conception des vocabulaires, ou du moins de leur gestion. Leurs compétences dans le traitement des données leur permettent de voir comment celles-ci peuvent gagner en visibilité et permettre de nouvelles collaborations grâce à l'utilisation des standards du Web. L'utilisation de ces standards pour la modélisation et la formalisation d'un domaine garantit sa réutilisabilité technique et son expansion. Elle garantit aussi son intégration dans divers dispositifs numériques selon l'évolution des besoins. Elle lui permet aussi d'aller au-delà de l'univers fermé de l'organisation, c'est-à-dire dans le Web de données. Par ailleurs, les PIST peuvent accompagner et soulager les chercheurs pour les tâches d'archivage de leurs données. Mais ils peuvent aussi participer à la recherche de ressources pour l'élaboration d'outils comme les vocabulaires contrôlés et la modélisation des connaissances de leur domaine sous la forme d'ontologies informatiques.

Au niveau global encore, les différents SOCs sont aujourd'hui mobilisés pour valoriser les données car c'est à l'échelle des données que se situent les potentiels de connaissance. D'ailleurs, des projets dans des disciplines connexes à la toxicologie nucléaire sont actuellement menés. Les enjeux de la valorisation des données de la recherche sont communs aux diverses communautés scientifiques du fait des décisions prises au niveau européen[105]. Ces enjeux convergent avec le mouvement en cours au sein du milieu professionnel de l'I&D des secteurs publics. En effet, des institutions de recherche et les institutions patrimoniales opèrent une mutation de leurs outils et pratiques. Ceci se fait dans la continuité de la tradition d'échange et de mutualisation des données qu'ils participent à enrichir collectivement. Ainsi, depuis 2008, ces professionnels lient les différents vocabulaires contrôlés et référentiels dont ils ont la charge et font basculer leurs données dans le Web de données. Ces deux mouvements d'ouverture concourent donc à l'amélioration de la diffusion des connaissances. Ils contribuent par là même à la mutualisation et à la collaboration entre les chercheurs.

[104] D'après les propos de Florence Gicquel lors de sa présentation des résultats de l'enquête « *Transformation* numérique » menée conjointement par l'Ecole des bibliothécaires documentalistes (EBD) et le cabinet *Knowledge Consult*, lors de la journée d'étude Médiadix « *Mémoires numériques – Publics, ressources et bibliothèques en mutatio*n » le 10 octobre 2014.
[105] « Comité scientifique du Conseil européen de la recherche (CER) ». In : *Libre accès à l'information scientifique et technique : actualités, problématiques et perspectives* [En ligne]. 2008. Disponible sur : http://openaccess.inist.fr/?Comite-scientifique-du-Conseil

En dernier lieu, les méthodes et solutions proposées pour la recherche, la sélection et la description des ressources utiles pour décider de la poursuite du travail de formalisation de ToxNuc ont été bien réceptionnées. Ces méthodes peuvent être adaptées aux autres domaines relevant du périmètre des sciences du vivant et des sciences biomédicales. Cependant, d'autres seraient à développer pour des recherches liées aux ontologies relevant des Sciences humaines et sociales (SHS). Entre autres limites à l'étude menée, poursuivre l'observation du processus de réflexion en cours sur ToxNuc excède en temps la durée du stage. Ainsi, à l'issue de cette mission, l'observation de la gouvernance d'une ontologie, et en l'occurrence ici de celle d'une modélisation des connaissances d'un domaine, ne peut pleinement aboutir. La contribution de cette étude proviendra de sa capacité à impulser plusieurs réflexions :

- l'évaluation des besoins de la communauté des toxicologues dans le contexte stratégique et économique actuel ;
- le management de la documentation relative au projet de construction d'ontologie en tant que moyen de pérennisation et de valorisation ;
- la prise de conscience de l'importance à accorder à la standardisation et la structuration des métadonnées en général.

En cas de reprise du projet, poursuivre l'observation permettrait de tester et de réviser les recommandations synthétisées au cours du travail de réflexion mené en parallèle du traitement de l'ontologie lors du stage. L'ontologie ToxNuc gagnerait aussi à faire la preuve de ses potentiels comme cela est le cas dans le domaine de la génomique avec la Gene Ontology (GO). Produire un démonstrateur avec l'ontologie ToxNuc est probablement la meilleure façon de convaincre la communauté de consacrer du temps et des moyens pour l'interopérabilité de ses données. L'interopérabilité permettrait :

- d'assurer la connexion entre diverses applications du CEA ;
- d'assurer l'échange de données avec les institutions partenaires ;
- de porter certaines données de l'institution dans le web de données.

Selon les préconisations des différents experts du liage des données, ce type de projet a également intérêt à être adossé à un autre projet dont les finalités apparaissent plus directement. Ils préconisent aussi de mettre en place une organisation en mode projet entre les partenaires[106]. Afin également de susciter l'adhésion, un autre objectif fort pour les porteurs de projets de ce type est l'exigence portée sur l'ergonomie fonctionnelle et l'esthétisme des interfaces des outils développés, même lorsqu'il s'agit de démonstrateurs.

Par ailleurs si ToxNuc a vocation à perdurer pour contribuer à faire de la toxicologie nucléaire prédictive, il sera nécessaire que l'ontologie soit enrichie et actualisée régulièrement. Le défi sera alors d'en établir la gouvernance : quels curateurs, avec quelles procédures, avec quels outils ? Les quatorze ans de recul sur le maintien de la GO sont riches d'enseignements à ce propos. Ils questionnent tout autant les fondements épistémologiques et méthodologiques des ontologies, qu'ils permettent aussi de dessiner l'histoire de la biologie moléculaire [42, MAYOR *et al.*]. Grâce à cet héritage, des dispositions peuvent être prises, entre autres, pour que les aménagements mais surtout les suppressions de concepts, de relations et avec eux la suppression de leurs annotations, ne disparaissent pas totalement. L'application de la nouvelle norme ISO 25964 assure la bonne gestion de l'ensemble des SOCs, pour peu que la décision soit prise de l'implémenter. Il faudra aussi compter sur le fait que les outils proposés par les éditeurs de logiciels de gestion de ces vocabulaires en permettent aussi une administration intuitive.

Grâce à cette gouvernance, les scientifiques devraient être en capacité de retracer les différentes étapes de la construction de cette ontologie de domaine et donc de cette discipline. En effet, une gouvernance qui omet ou n'anticipe pas l'usage des données de gestion pour ne penser les usages qu'au présent, peut faire disparaître une partie des débats ou du moins leurs traces. Certes, les débats seront autrement et probablement traçables au travers de leur présence dans les wikis, listes de diffusion et échanges de courriels. Ceci traduit d'ailleurs bien l'idée qu'une ontologie est aussi une construction sociale résultant d'une négociation entre experts. Enfin, la gouvernance doit envisager l'infrastructure nécessaire pour porter dans le Web les données structurées par une ontologie. Cela permettra d'enrichir le web de données et d'ajouter à la diversité des points de vue qui y sont déjà présents. Le passage dans ce nouvel environnement favoriserait le dialogue entre les communautés et pourrait faire évoluer leurs disciplines respectives.

[106] Cf. Annexe 2 - Livrable 4 - Etude d'opportunité pour le projet de reprise de l'ontologie ToxNuc à partir des informations recueillies

Dans cet écosystème pour l'instant ouvert, librement favorable au *mash-up*, les fonctions des infomédiaires, dont font partie les gestionnaires d'information comme les documentalistes, seront probablement amenées à elles aussi « mash-uper » avec celles d'autres professions[107]. Tout du moins, ces professionnels seront amenés à collaborer encore davantage en amont pour animer des groupes d'experts afin de désambiguïser, organiser la connaissance et aider à la valoriser. En aval de la chaîne de production de l'information, il s'agit pour eux d'adapter leurs compétences pour faire reconnaître ce qui est en fait dans le prolongement de leurs missions traditionnelles : celles de classer pour organiser l'accès mais aussi veiller pour sélectionner et contrôler la qualité, non plus forcément des données, mais des sources de données déjà décrites par leurs producteurs, et en gérer les flux. Le Web sémantique est une promesse pour éviter les silos de données et mettre en valeur le Web profond, dit Web invisible ou encore Web caché, car inaccessible des robots indexeurs des moteurs de recherche généralistes. Cependant, les nouveaux gisements promis par les mégadonnées[108] risquent en même temps d'opacifier le Web. Dans ce contexte, la captation des données par certains modèles dominants et le repositionnement des différents acteurs de la chaîne de valeur, sont à la source de futurs modèles juridiques et contractuels. Pour les professionnels de l'information et de la documentation (I&D) ainsi que pour ceux des sciences de l'information et des bibliothèques (SIB), il conviendra alors de les surveiller et de les maîtriser autant que de s'approprier les évolutions technologiques.

[107] En référence au commentaire de Stéphane Pouyllau, aujourd'hui directeur-adjoint technique la TGIR Huma-num et co-directeur de projet de la plateforme Isidore, à l'un des billets d'Emmanuelle Bermès sur son blog *Figoblog*
BERMÈS E. *Le problème avec le catalogue...* [En ligne]. *Figoblog, un blog sur Internet, la bibliothéconomie et la confiture de figues*. 4 novembre 2010. Disponible sur :
< http://www.figoblog.org/node/1982 >

[108] La délégation générale à la langue française et aux langues de France (DGLFLF) recommande d'employer cette expression à la place de Big data
< http://www.culture.fr/layout/set/print/franceterme/terme/COGE874 >

Bibliographie

La bibliographie qui suit a été arrêtée le 18 novembre 2014. Tous les liens d'accès aux documents référencés étaient valides à cette date. Seuls les liens pérennes vers les ressources accessibles gratuitement ont été retenus.

Il s'agit d'une bibliographie analytique, organisée autour des grands thèmes traités dans ce mémoire :

Les systèmes d'organisations des connaissances (SOCs)

Les folksonomies

L'indexation automatique

L'organisation des connaissances

Les ontologies informatiques

Le programme Transversal Toxicologie Nucléaire

Ces thématiques principales sont elles-mêmes divisées en sous-thématiques, afin de permettre au lecteur un accès rapide aux ressources les plus pertinentes selon sa problématique.

Les références comportent une numérotation correspondant à celle reprise dans le corps du mémoire sous la forme [n°, NOM de l'auteur]. Dans chaque sous-thématique, les documents sont classés par ordre alphabétique des noms d'auteur, à l'exception de la norme ISO qui a volontairement été placée en tête des références.

La rédaction des références bibliographiques est conforme aux normes :

- Z44-005. décembre 1987. Documentation. Références bibliographiques : contenu, forme et structure et à la norme
- NF ISO 690-2 Février 1998 Information et documentation. Références bibliographiques Documents électroniques, documents complets et parties de documents

L'encyclopédie en ligne Wikipédia, qui n'est pas citée dans cette bibliographie, a également été utilisée régulièrement, comme première approche pour cerner certaines notions, ou pour obtenir des informations factuelles.

1 Les systèmes d'organisations des connaissances (SOCs)

[1] BOULOGNE A., INSTITUT NATIONAL DES TECHNIQUES DE LA DOCUMENTATION. Vocabulaire de la documentation. Paris, France : ADBS éditions, DL 2004, 2004. 334 p. ISBN : 2-84365-071-2.
[2] CACALY S., LE COADIC Y.-F., POMART P.-D. Dictionnaire de l'information. Paris : A. Colin, 2004. ISBN : 2-200-26682-0 ; 978-2-200-26682-0.
[3] LE COADIC Y.-F., POMART P.-D., SUTTER E. Dictionnaire de l'information. Paris, France : A. Colin, DL 2008, 2008. VI-295 p. ISBN : 978-2-200-35132-8.

Ces trois dictionnaires spécialisés ont été utiles pour observer l'évolution de la définition des ontologies émises à l'attention des professionnels de l'I&D.

Les langages documentaires et indexation

[4] ADBS. Langages documentaires et outils linguistiques. *Documentaliste-Sciences de l'Information* [En ligne]. ADBS, 2007, Vol. 44, n°1, p. 66-74. ISSN 0012-4508. Disponible sur : < http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1.htm >

Le dossier de ce numéro de la revue professionnelle Documentaliste-Sciences de l'Information dresse un bilan rétrospectif et prospectif sur les langages documentaires notamment concernant leur complémentarité avec les méthodes et technologies sémantiques tout en abordant les questions relatives à l'interopérabilité. Plusieurs des articles contenus dans ce dossier sont cités dans cette bibliographie.

[5] DALBIN, Sylvie. Descripteurs : site dédié aux thésaurus et autres vocabulaires contrôlés pour l'accès à l'information, [En ligne]. < http://dossierdoc.typepad.com/descripteurs >

Il s'agit d'un blog dédié aux thésaurus et autres vocabulaires contrôlés pour l'accès à l'information. Son contenu est régulièrement actualisé. L'auteur y publie également certains des articles parus dans différentes publication.

[6] MENON B. « Les langages documentaires : un panorama, quelques remarques critiques et un essai de bilan ». *Documentaliste-Sciences de l'Information* [En ligne]. 2007. Vol. 44, n°1, p. 18. DOI : 10.3917/docsi.441.0018. Disponible sur : < http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-18.htm >

Cet article a été particulièrement structurant pour tenter d'établir une filiation entre les différents langages documentaires et clarifier les termes parfois approximatifs qui servent à les désigner.

Thesaurus

[7] Norme internationale ISO 25964-1 : *Information et documentation, Thésaurus et interopérabilité avec d'autres vocabulaires Partie 1, Thésaurus pour la recherche documentaire.* 1e édition. Genève, ISO/AFNOR, 2011. 152 p.

Le report à la norme s'est avéré nécessaire pour poser la définition de ce qu'est un thesaurus dont l'orientation « concept » a été renforcé et formalisé. Y sont abordés les aspects allant du développement à la maintenance des thésaurus monolingues et multilingues. La norme inclut également les questions relatives aux formats et aux protocoles d'échanges de données.

[8] Norme internationale ISO 25964-2 : *Information and documentation – Thesauri and interoperability with other vocabularies – Part 2 : Interoperability with other vocabularies*. 1ère édition. Suisse, ISO, Mars 2013. 99 p.

Le second volume de la norme publié près d'un an et demi après le volume premier n'a pas encore été traduit. Il est consacré à l'interopérabilité des thésaurus avec les autres langages documentaires dans le contexte du Web sémantique. « Il propose des solutions pour l'utilisation simultanée de plusieurs langages contrôlés pour accéder à de très larges collections de ressources distribuées sur de multiples réseaux, rapidement, efficacement, et dans la langue choisie par le chercheur d'information. » Sa consultation a été particulièrement utile pour s'assurer des définitions des taxinomies et des ontologies établies et reconnues par l'ISO.

Voir aussi : ISO 25964-1 Schema and Data Model. [en ligne]. Disponible sur :
< http://www.niso.org/schemas/iso25964/#schema >

Cette page Web offre un aperçu des deux volumes de la norme ainsi que des liens vers des ressources complémentaires, notamment le schéma de données RDF de la norme ainsi que le tableau de correspondances entre la norme et SKOS.

[9] DALBIN S., YAKOVLEFF N., ZYSMAN H, et al. Livre blanc sur la norme ISO 25964-1 « Thésaurus pour la recherche documentaire » [En ligne]. AFNOR, BiVi, mis en ligne le 29 janvier 2013. Disponible sur : < http://www.bivi.fonctions-documentaires.afnor.org/livres-blancs/livre-blanc-sur-la-norme-iso-25964-1-thesaurus-pour-la-recherche-documentaire-parue-en-version-anglaise-en-aout-2011 >

Après être revenu sur les enjeux et le contexte de la parution d'une nouvelle norme concernant les thésaurus, ce document en français présente une version synthétique de la norme. Cette synthèse est destinée aux professionnels de l'I&D, aux informaticiens en charge de dispositifs d'accès à l'information, et aux architectes de l'information.

[10] DALBIN S. La norme « ISO 25964-1(2011) - Thésaurus pour la recherche documentaire » est publiée [En ligne]. Descripteurs. 21 août 2011. Disponible sur :
< http://dossierdoc.typepad.com/descripteurs/2011/08/norme-iso-25964-1-thesaurus-publication-officielle.html >

L'auteur de ce billet de blog présente succinctement la norme ISO 25964-1:2011 en donnant également un lien vers la représentation sous la forme d'un diagramme UML du modèle de données métier de la norme qui devrait en faciliter l'implémentation dans les logiciels et faciliter l'interopérabilité entre les différents langages et systèmes.

Historique

[11] DALBIN S. « Thésaurus et informatique documentaires: Des Noces d'Or ». *Documentaliste-Sciences de l'Information* [En ligne]. 2007. Vol. 44, n°1, p. 76-80. DOI : 10.3917/docsi.441.0076 Disponible sur :
< http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-76.htm >

Dans cet article, l'auteur revient sur cinquante années d'informatisation des thésaurus. L'auteur déplore la faiblesse des différents logiciels documentaires concernant l'indexation, la recherche, la maintenance terminologique.

2 Les folksonomies

[12] AMAR M. Nouvelles pratiques d'indexation, nouveaux enjeux documentaires ? [En ligne]. 14 avril 2008. Disponible sur :
< http://urfist.enc.sorbonne.fr/sites/default/files/file/traitementdoc/Pratiques-d'indexation-support.pdf >

Il s'agit du support conçu par l'auteur pour l'animation d'une formation dispensée par l'Urfist de Paris visant à situer la diversité des modes d'indexation des ressources numériques. L'auteur revient sur la notion d'indexation, son histoire, ses problématiques notamment avec les avancées technologiques et l'indexation automatique. L'auteur présente aussi les enjeux de l'indexation sémantique utilisant les ontologies informatiques avant d'aborder l'indexation sociale.

[13] AMAR M. *Taxonomies, ontologies et folksonomies...* [En ligne]. 16 juin 2008. Disponible sur : < http://urfist.enc.sorbonne.fr/sites/default/files/file/traitementdoc/TaxonomiesAndCies-support-de-formation.pdf >

Il s'agit du support conçu par l'auteur pour l'animation d'une formation dispensée par l'Urfist de Paris visant à dresser une typologie des langages documentaires, à les définir et les comparer. L'auteur aborde aussi les enjeux de l'interopérabilité dans l'environnement du Web et s'attache à identifier les usages.

[14] ERTZSCHEID O. To tag or not to tag [En ligne]. *Affordance.info, ISSN 2260-1856*. 30 octobre 2007. Disponible sur :
< http://affordance.typepad.com/mon_weblog/2007/10/to-tag-or-not-t.html >

L'auteur de ce blog pose la question de la convergence des systèmes d'indexation sociale pour faire évoluer l'indexation plus classique ou bien automatisée. Pour cela l'auteur se réfère à un article dont il propose un résumé et à deux exemples d'hybridation de l'indexation classique avec l'indexation sociale.

[15] POUPEAU G. *La folksonomie, c'est limité* [En ligne]. *Les petites cases*. 8 mars 2006. Disponible sur : < http://www.lespetitescases.net/la-folksonomie-c-est-limite >

Dans ce billet, l'auteur revient sur l'opposition entre la folksonomie et les vocabulaires contrôlés. Il pose aussi la question des usages et en propose de nouvelle combinaison entre vocabulaires contrôlés et indexation sociale.

[16] POUPEAU G. *L'ontologie est-elle vraiment surfaite ?* [En ligne]. *Les petites cases*. 18 avril 2006. Disponible sur : < http://www.lespetitescases.net/l-ontologie-est-elle-vraiment-surfaite-y >

Ce billet aborde la question des ontologies informatiques, en développant un argumentaire contre ses détracteurs. Mais c'est surtout la proposition d'associer l'indexation sociale et les ontologies informatiques pour organiser l'information qui a été intéressante pour réfléchir à la combinaison possible entre l'ontologie ToxNuc et l'outil Folksonomies du LGI2P.

[17] VAN DAMME C., HEPP M., SIORPAES K. « Folksontology: An integrated approach for turning folksonomies into ontologies ». *Bridging the Gap between Semantic Web and Web*. 2007. Vol. 2, n°2, p. 57–70.

Dans cet article académique datant de 2007, les auteurs font le constat qu'il est encore loin d'exister des ontologies pour représenter certains domaines alors que l'indexation sociale très largement utilisée peut couvrir de nombreux domaines. Ils proposent alors d'appréhender la résultante de l'indexation sociale que sont les tags comme une forme de négociation entre les individus à propos des contenus indexés. Par conséquent ils suggèrent de prendre appui sur l'indexation sociale pour concevoir des ontologies et décrivent alors les ressources et les techniques utilisées pour valider leur hypothèse.

3 L'indexation automatique

[18] CHARTRON G., DALBIN S., MONTEIL M.-G., VERILLON M. « Indexation manuelle et indexation automatique : dépasser les oppositions ». *Documentaliste-Sciences de l'information*. juillet 1989. Vol. 26, n°4-5, p. 181-187.

Cet article datant de 1989 relate une expérience de comparaison entre l'indexation manuelle et l'indexation automatique d'un même corpus de documents selon trois critères. Les conclusions invitent à l'introduction de la sémantique grâce à des méthodes statistiques. Les auteurs de l'article conseillent également d'associer une supervision par un expert humain à l'indexation automatique réalisée avec le système LEX/NET car leur complémentarité semble être le gage d'une indexation de meilleure qualité. Il est par ailleurs consultable en ligne sur le site professionnel de l'un de ses auteurs à l'adresse suivante :
<http://www.atd-doc.com/xmedia/publications/IndexationManuelleAutoDepasser-les-oppositions-1989.pdf>

4 L'organisation des connaissances

[19] CAUSSANEL J., CAHIER J.-P., ZACKLAD M., CHARLET J. « Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web Sémantique ». In : *Conférence Ingénierie des Connaissances IC2002, Rouen Mai*. [s.l.] : [s.n.], 2002.]

Dans cet article datant de 2002, les auteurs présentaient le formalisme Topic Maps pour la représentation des connaissances, qu'ils considéraient comme prometteur pour l'intégration des ontologies dans le Web et pour répondre aux objectifs du Web sémantique. Même s'il semble que depuis les recherches sur les Topic Maps soient ralenties et que cela soit le formalisme OWL qui se soit répandu, il intéressant de signaler ces travaux dans une perspective historique.

[20] PEPPER S. « The TAO of Topic Maps ». In : *Ontopia* [En ligne]. [s.l.] : [s.n.], [s.d.]. Disponible sur : < http://www.ontopia.net/topicmaps/materials/tao.html >

Cet article est probablement une version postérieure à 2002 d'un article dont la version originale date de 2000. Il explique la structure du modèle Topic Maps et non sa syntaxe XTM. Le but des Topic Maps est moins de faire des inférences que de permettre la navigation dans de vastes corpus interconnectés.

[21] ISKO ; POLITY, Yolla, HENNERON, Gérard, PALERMITI, Rosalba, éd. *L'organisation des connaissances: approches conceptuelles*. Paris : Harmattan, 2005. 266 p. (La librairie des humanités). ISBN 2-7475-8274-4

La lecture des communications de ce congrès de l'International Society for Knowledge Organization a été l'une des premières références pour aborder ce travail de réflexion sur les ontologies conçues non comme un outil du Web sémantique, mais comme une forme de SOC. Cette lecture a permis de dresser une première typologie des SOC en les replaçant dans une perspective documentaire de classification et d'indexation.

***Voir aussi** :*

[22] DALBIN S. « Quatrième congrès d'ISKO-France: L'organisation des connaissances : approches conceptuelles ». *Documentaliste-Sciences de l'Information* [En ligne]. 2003. Vol. 40, n°6, p. 380. DOI : 10.3917/docsi.406.0380. Disponible sur : < http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2003-6-page-380.htm >

Le compte-rendu rédigé par Sylvie Dalbin pour la revue professionnelle Documentaliste-Sciences de l'Information est accessible en ligne et donne une définition des ontologies du point de vue des professionnels de l'I&D.

[23] MONDECA blog. Les taxonomies de navigation - La recherche à facettes : Définition, utilisation, objectifs, mise en œuvre. [En ligne]. *Mondeca - Leçons de Choses*. Disponible sur : < https://mondeca.wordpress.com/2007/10/07/les-taxonomies-de-navigation-la-recherche-a-facettes-definition-utilisation-objectifs-mise-en-oeuvre >

La présence de contenus numériques et la recherche informatisée actualisent la notion de taxonomie. L'auteur expose comment la société éditrice de logiciel de gestion de vocabulaires contrôlés Mondeca utilise ce qu'il est maintenant courant d'appeler des « taxonomies de navigation » pour guider les usagers dans la découverte des contenus d'un site Web. Celles-ci se distinguent des taxinomies classiques car elles sont construites à partir de critères relevant du marketing.

[24] REMILLIEUX A. *Explicitation et modélisation des connaissances de conduite de changement à la SNCF : vers une gestion des connaissances préréfléchies*. Business administration. Institut National des Télécommunications, 2010. NNT : 2010TELE0013. tel-00693957

Cette référence a été retenue car y est illustré un cas concret de gestion des connaissances avec la mise en place d'un dispositif socio-technique. Y est aussi expliquée l'articulation entre l'ontologie de la base de connaissance et la « taxonomie de navigation » de l'interface utilisateur.

[25] ZACKLAD M. « Évaluation des systèmes d'organisation des connaissances ». *Les Cahiers du numérique*. 2010. Vol. 6, n°3, p. 133–166.

L'auteur traite des différents SOCs adaptés à la gestion de l'information documentaire qui est aujourd'hui de plus en plus numérique. Il constate qu'il est difficile de les discerner les uns des autres dans ce contexte. Il élabore une grille de critères pour tenter de les distinguer et d'appréhender leur complémentarité.

5 Les ontologies informatiques

Pour une première approche du sujet

[26] RICHY H., DESPRÉS S. Métadonnées, ontologies et documents numériques. *Techniques de l'ingénieur*. 2007. p. h7155v2.

Les auteurs de cet article commencent par distinguer les métadonnées, qui historiquement, sont destinées à classer et donc à donner accès à la documentation sous forme papier, des ontologies, permettant d'organiser, de structurer les connaissances appliquées à certains domaines. Le recours aux métadonnées est ensuite analysé dans le contexte du Web en faisant un rappel sur les formats informatiques. Enfin, dans une dernière partie, sont présentées les nouvelles solutions visant à rendre les métadonnées compréhensibles par les machines, notamment les perspectives offertes par l'usage des ontologies.

[27] BACHIMONT B. *Qu'est-ce qu'une ontologie ?* [En ligne]. 2004. Disponible sur : < http://www.technolangue.net/imprimer.php3?id_article=280 >

Cette interview de B. Bachimont est conseillée à toute personne souhaitant rapidement comprendre ce qu'est une ontologie. Très accessibles, les réponses apportées aux questions de son intervieweuse permettent d'appréhender concrètement la construction mais aussi les problèmes posés lorsqu'il s'agit de faire évoluer une ontologie. Par ailleurs sa réponse à la question sur le risque d'uniformisation culturelle et sur le classement du patrimoine ouvre des pistes de réflexion.

[28] GANDON F. *Ontology In A Nutshell (version 2)* [En ligne]. 27 novembre 2007. Disponible sur : < http://fr.slideshare.net/fabien_gandon/ontology-in-a-nutshell-version-2 >

Cette présentation de 81 diapositives explique de façon très illustrée ce qu'est une ontologie informatique. Plus que pour les aspects relatifs aux langages RDF et OWL, cette présentation a été retenue pour ces définitions des différents SOCs.

Les références incontournables

[29] BORST W. *Construction of Engineering Ontologies. PhD thesis,*. [s.l.] : University of Tweente, Enschede, NL–Centre for Telematica and Information Technology, 1997.

Cette référence a été sélectionnée pour cette bibliographie bien que cette thèse n'ait pas été consultée. La raison en est que de nombreuses publications scientifiques s'y réfèrent car l'auteur développe la notion de consensus qui a été ajoutée plus tard par [31, STUDER et al] à la définition de l'ontologie informatique publiée auparavant par [32, GRÜBER].

[30] GUARINO N., OBERLE D., STAAB S. « What Is an Ontology? ». In : STAAB S., STUDER R., éd. *Handbook on Ontologies*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2009. p. 1-17. ISBN : 978-3-540-70999-2 ; 978-3-540-92673-3.

Bien que cet article apparaisse très technique pour certaines de ses parties, il est intéressant car il explique les notions de conceptualisation, de spécification, ainsi que la notion de consensus. Il rappelle également la fusion entre les définitions de [32, GRÜBER].et [29, BORST]. La figure 4 de ce chapitre de livre a participé en partie à inspirer la figure 7 sur l'expressivité des SOCs p. 38 de ce mémoire.

[31] STUDER R., BENJAMINS V. R., FENSEL D. « Knowledge engineering: Principles and methods ». *Data & Knowledge Engineering*. 1998. Vol. 25, no. 1-2, p. 161-197.

Il s'agit ici de l'article à l'origine de la fusion de la définition donnée par [32, GRÜBER].et celle de [29, BORST].

[32] GRUBER T. R. « A translation approach to portable ontology specifications ». *Knowledge acquisition*. 1993. Vol. 5, n°2, p. 199–220.

Cette référence n'a été que très peu consultée pour la raison suivante : presque toutes les publications scientifiques consacrées aux ontologies reprennent les arguments de T. Grüber. C'est également pour cette raison qu'il paraît indispensable de la faire figurer dans cette bibliographie sélective et analytique.

La conception des ontologies

[33] AUSSENAC-GILLES N., CHARLET J., RAYNAUD-DELAÎTE C. « Ingénierie des connaissances ». In : MARQUIS P, PAPINI O, PRADE H, éd. *Panorama de l'intelligence artificielle : ses bases méthodologiques, ses développements. Vol. 1 : Représentation des connaissances et formalisation des raisonnements*. Toulouse : Cépaduès, 2014. p. 615-649.ISBN : 978-2-36493-0414.

Cet article a été utile pour comprendre la filiation de l'ingénierie des connaissances avec l'intelligence artificielle. Ses auteurs retracent l'histoire de l'IC et balayent un panorama des méthodes, des logiciels et des interfaces d'aide à la modélisation. Enfin, les auteurs ouvrent sur les perspectives offertes par les nouveaux usages permis par le Web social. Ils préfigurent aussi que le Web de données conduit à orienter la recherche sur la question des alignements pour typer et organiser les données.

[34] AUSSENAC-GILLES N., CONDAMINES A. « Corpus et terminologie ». In : *Roger T. Pédauque. La redocumentarisation du monde*. Toulouse : Cépaduès, 2007. p. 131-147.ISBN : 978-2-85428-728-8.

Cette contribution a permis d'éclairer la vision des différentes disciplines concernées par la constitution de produits terminologiques à partir de textes et en lien avec différents types d'applications : la linguistique de corpus, la terminologie, l'informatique et les sciences de l'information. Cette synthèse de travaux interdisciplinaires est intéressante dans le cadre de ce mémoire car elle prend en compte les questions d'usage et de maintenance ainsi que les modalités d'évaluation des ressources terminologiques ou ontologiques (RTO). Par ailleurs, il a permis de souligner que le TALN peut être à la fois « producteur et utilisateur de RTO ».

[35] BACHIMONT B. « Engagement sémantique et engagement ontologique: conception et réalisation d'ontologies en ingénierie des connaissances ». In : *Ingénierie des connaissances: évolutions récentes et nouveaux défis*. Paris : Eyrolles, 2000. p. 305–323.

Dans cet article, l'auteur présente sa méthode pour construire des ontologies à partir de l'expression linguistique des connaissances présente dans des corpus. Il développe les notions d'engagement sémantique, d'engagement ontologique, de sémantique formelle et de sémantique différentielle. Il était important de comprendre cette méthode car les concepteurs de l'ontologie ToxNuc s'en sont inspirés.

[36] BALMISSE G. *Guide des outils du knowledge management: panorama, choix et mise en œuvre*. Paris : Vuibert, 2005. 315 p. (Collection Entreprendre informatique). ISBN : 2711748294

Cet ouvrage très abordable tant pour les aspects sociaux que pour les aspects techniques de la gestion des connaissances au moyen d'outils socio-techniques a été consulté à de nombreuses reprises. Il a notamment permis de comprendre l'articulation des différentes techniques du traitement automatique du langage naturel (TALN) et il a inspiré la figure 4 p. 30 de ce mémoire.

[37] CHAUMIER J. « Les ontologies : antécédents, aspects techniques et limites ». *Documentaliste-Sciences de l'Information*. 2007. Vol. 44, n°1, p. 81. DOI : 10.3917/docsi.441.0081. Disponible sur : < http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-81.htm >

Cette lecture issue de la revue professionnelle Documentaliste-Sciences de l'Information est intéressante pour quiconque souhaite découvrir les ontologies sans posséder de connaissances en informatique et pour aussi avoir accès à une sélection bibliographique. Y sont succinctement exposés un historique, la nature et la structure technique des ontologies ainsi que les outils méthodologiques et techniques pour les concevoir et les développer. Enfin, l'auteur aborde leur lien avec les thésaurus et fait part aussi de leurs limites notamment concernant le coût et les délais pour concevoir ce type de dispositif.

[38] DECLERCK G., CHARLET J. « Intelligence artificielle, ontologies et connaissances en médecine : les limites de la mécanisation de la pensée ». *Revue d'intelligence artificielle* [En ligne]. 30 août 2011. Vol. 25, n°4, p. 445-472. Disponible sur : < http://dx.doi.org/10.3166/ria.25.445-472 >

Il s'agit ici d'un article théorique qui interroge sur la neutralité ou plutôt sur l'intentionnalité de la technologie, intentionnalité qui se limite à celle que les hommes peuvent encoder dans les machines.

[39] DJAMBIAN C. « Le métier : son savoir, son parler ». In : INSTITUT PORPHYRE SAVOIR ET CONNAISSANCE (ANNECY)., SOCIÉTÉ FRANÇAISE DE TERMINOLOGIE., UNIVERSITÉ DE SAVOIE., UNIVERSITÉ SORBONNE NOUVELLE (PARIS)., éd. *Actes de la conférence TOTh, Annecy, 26-27 mai 2011* [En ligne]. Terminologie & Ontologie : Théories et applications. Annecy : Institut Porphyre, savoir et connaissance, 2012. p. 75–91. ISBN : 978-2-9536168-4. Disponible sur : < http://www.porphyre.org/toth/files/actes/TOTh-2011-actes.pdf >

Cette référence est intéressante car son auteur expose sa démarche pour construire une terminologie puis une ontologie en vue de constituer une base de connaissances pour les ingénieurs de la Division Ingénierie Nucléaire d'EDF.

[40] GANDON F. « Ontologies informatiques ». In : *Interstices* [En ligne]. [s.l.] : [s.n.], 2006. Disponible sur : < https://interstices.info/ontologie >

Cet article librement accessible en ligne et dont la lecture est par ailleurs abordable permet de prendre connaissance de ce qu'est une ontologie informatique. L'auteur propose également de nouvelles applications des ontologies autres que l'accès à l'information.

[41] GUARINO N. « Formal ontology in information systems : proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy ». In : *Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998*. Amsterdam : IOS Press, 1998. p. 3-15

L'auteur revient sur les « ontologies fondationnelles », c'est-à-dire les ontologies de haut niveau (Top ontologies) qui sont considérées comme nécessaires pour l'interopérabilité sémantique entre différents systèmes d'information et pour permettre l'interdisciplinarité.

[42] MAYOR C., ROBINSON L. « Ontological realism, concepts and classification in molecular biology: Development and application of the gene ontology ». *Journal of Documentation*. 2014. Vol. 70, n°1, p. 173-193.

Les auteurs de cet article se penchent sur le développement du projet Gene Ontology qui a été lancé en 2000. Il est intéressant au titre de la gouvernance mise en place pour maintenir et enrichir cette ressource, sujet auquel s'intéresse ce mémoire. Les auteurs mettent l'accent sur le dispositif social avec l'étude de la gestion des désaccords et la subjectivité d'un groupe restreint malgré le souhait de consensus formulé par les principes méthodologiques ayant trait à la construction d'ontologie.

Les ontologies et leurs applications

[43] CHARLET J., BACHIMONT B., TRONCY R. « Ontologies pour le Web sémantique ». *Revue Information, Interaction, Intelligence I3* [En ligne]. 2004. Disponible sur : < http://www.eurecom.fr/~troncy/Publications/Troncy-revue_i304.pdf >

Cet article revient lui aussi sur la généricité des ontologies. S'il fait le point sur les problèmes que rencontre le Web sémantique par rapport aux ontologies et les axes de réflexion ou de recherche dans ce domaine, c'est plus la section consacrée à l'opposition entre ontologie et thésaurus qui a retenu l'attention quant à la typologie établie dans ce mémoire.

[44] CHARLET J., BANEYX A., STEICHEN O., ALECU I., DANIEL-LE BOZEC C., BOUSQUET C., JAULENT M.-C. « Utiliser et construire des ontologies en médecine. Le primat de la terminologie. » *Technique et science informatiques*. 2009. Vol. 28, n°2, p. 145–171.

Cet article a l'avantage d'exposer trois cas d'usage assez différents dans le domaine médical. Dans certains cas, il est possible d'avoir recours aux ressources linguistiques comme les classifications et les thésaurus particulièrement développés dans ce domaine pour concevoir des ontologies. Mais il est aussi nécessaire de recourir à des corpus bien particuliers comme les textes générés par les professionnels de santé lors de l'exercice de leur activité.

[45] DELAHOUSSE J. « Apports d'une ontologie de domaine aux services d'accès aux contenus ». In : GRIVEL L. éd. *La recherche d'information en contexte outils et usages applicatifs*. Paris : Hermès science publications : Lavoisier, 2011. p. 179-190.ISBN : 9782746225817 2746225816.

Les exemples donnés dans cette contribution illustrent les apports des ontologies dans les services d'accès aux contenus sur les sites Internet : plus grande efficacité des moteurs de recherche ; navigation et recherche à facettes grâce aux taxinomies dynamiques, à l'annotation sémantique et la classification supervisée par les ontologies. Ceci permet d'imaginer des usages de l'ontologie ToxNuc au sein de la plateforme du programme Transversal Toxicologie Nucléaire. Par ailleurs l'auteur évoque une autre finalité comme les bases de connaissances et pose aussi les principes du maintien de la qualité. Il traite enfin de l'utilisation du standard SKOS.

[46] FRANCART T. *Recherche d'informations : du plein-texte aux ontologies* [En ligne]. *Sparna blog*. Disponible sur : < http://blog.sparna.fr/recherche-informations-moteur-plein-texte-ontologies >

Dans ce billet, l'auteur de ce blog met en garde contre l'idée qu'une ontologie à elle seule permette d'accéder aux informations contenues dans un corpus. Il résume, par ordre croissant de complexité de mise en œuvre, les différentes possibilités de faire de la recherche d'information dans ce corpus en partant de la recherche non-structurée plein texte jusqu'à celle utilisant les vocabulaires métiers. Selon l'auteur, les ontologies sont plus complexes que les vocabulaires métiers et sont plus orientées vers la constitution d'une base de connaissance que destinées à être couplées avec un moteur de recherche dans l'optique d'accéder aux corpus.

[47] ISAAC A., BOUCHET T. « RAMEAU et SKOS » [En ligne]. Arabesques. avril - mai - juin 2009. n°54, p. 13-14. . Disponible sur : < http://www.abes.fr/Arabesques/Arabesques-n-54 >

Dans ce court article, les auteurs relatent les réalisations du projet à titre expérimental de conversion au format RDF selon le modèle de représentation SKOS. Ils préviennent que l'objectif n'est pas de remplacer le format des vocabulaires contrôlés utilisé dans leur contexte habituel d'usage, c'est-à-dire le format MARC, mais d'en faciliter la publication, l'échange et l'interconnexion dans le contexte du Web sémantique. Ils exposent les limites de cette conversion ainsi que ses perspectives.

[48] ISAAC A. *Sémantique et interopérabilité* [En ligne]. In : *Journée d'études AFNOR/BNF CG46*. 28 mars 2008. 28 p. Disponible sur : < http://www.bnf.fr/documents/isaac.pdf >

L'auteur de cette présentation explique que l'intérêt de convertir les référentiels de type vocabulaire contrôlé dans les formats du Web sémantique est de faciliter l'interopérabilité au niveau sémantique de ces différents référentiels en procédant ensuite à des alignements.

[49] MERZEAU L. « Traces captées traces éditorialisées ». In : *Mémoire numérique. Publics, ressources et bibliothèques en mutation, Journée d'étude organisée par Médiadix et l'URFIST de Paris*. Médiadix, Saint-Cloud, 10 octobre 2014. Disponible sur : < http://merzeau.net/traces-captees-editorialisees >

La présentation a eu lieu dans le cadre d'une journée d'études dont le propos était que ce n'est plus sur leur seul enregistrement ou leur indexation que doit s'exercer une médiation des contenus, devenus aujourd'hui massivement numériques, mais sur leur éditorialisation.

[50] MCGUINNESS, Deborah L. « Ontologies Come of Age ». In : FENSEL, Dieter et al. dir. *Spinning the Semantic Web: Bringing the World Wide Web to its Full Potentia*l. Cambridge (Mass.) : MIT Press, 2003. P. 171-194

Comme dans de nombreux autres articles sur les ontologies, l'auteur propose un état de l'art avant d'exposer en quoi le Web sémantique et les usages actualisent les ontologies. Mais cette référence a retenu l'attention en raison d'une figure illustrant les différents niveaux de conceptualisation et de formalisation, qui a elle-même inspiré la figure 7 sur l'expressivité des SOCs p. 38 de ce mémoire.

[51] KEMBELLEC G., CHARTRON G. « Introduction générale aux systèmes de recommandation ». In : KEMBELLEC G, CHARTRON G, SALEH I, éd. *Les moteurs et systèmes de recommandation*. [S.l.] : ISTE Editions, 2014. p. 1– 24.ISBN : 978-1-78405-041-2 (papier) ISBN : 978-1-78406-041-1 (ebook).

Cette introduction sur les systèmes de recommandation montre comment la prescription se réactualise dans le contexte numérique où le nombre d'informations est toujours plus croissant. Elle est intéressante car elle expose l'une des applications de l'interconnexion des référentiels de données hétérogènes couplée avec les statistiques et des moteurs de règles. La conception de ces systèmes nécessite des compétences issues de plusieurs disciplines. Très développés dans le domaine de l'e-commerce, ces systèmes peuvent inspirer la contextualisation des données sur la plateforme ToxNuc.

6 Le Web sémantique / le Web de données

[52] LAUBLET P. Web sémantique - Principes, représentations sémantiques et ontologies [En ligne]. Encyclopédie Techniques de l'Ingénieur. 2010. p. h7502.

L'auteur revient sur les principes du Web pour montrer que le Web sémantique est une extension du Web actuel mais avec une sémantique interprétable par les machines et non plus seulement des humains. Cet article est intéressant car il revient sur les technologies du Web et montre les rapports entre Web sémantique, vocabulaires contrôlés et ontologies.

Les standards, les métadonnées et les données publiques

[53] ADBS. Web sémantique, web de données... Quelle nouvelle donne ? *Documentaliste-Sciences de l'information*, ADBS, 2011, Vol. 48, n°4, 80 p. ISSN 0012-4508

Le dossier de ce numéro de la revue professionnelle Documentaliste-Sciences de l'Information consacré au Web sémantique et au Web de données donne des clés de compréhension à propos des nouvelles technologies. Il aborde la question des transformations professionnelles pour les producteurs de métadonnées dont le domaine de l'I&D. Plusieurs articles issus de ce numéro sont présents dans cette bibliographie.

[54] BERMÈS E., ISAAC A., POUPEAU G. *Le Web sémantique en bibliothèque*. Paris : Éd. du Cercle de la librairie, 2013. (Collection Bibliothèques). ISBN : 978-2-7654-1417-9.

Cet ouvrage synthétique est abordable bien que technique. Il permet d'appréhender les intérêts et les enjeux pour une institution d'appliquer les principes du Web sémantique pour ses données. Le lecteur est guidé pas à pas à travers quatre cas pratiques pour appréhender les questionnements et démarches à appliquer.

[55] DALBIN S. « Faire vivre les données. 2 : référentiels et terminologies ». *Documentaliste-Sciences de l'Information*. 2013. Vol. 50, n°3, p. 38-39.

L'auteur aborde la question de la conversion des vocabulaires contrôlés dans l'univers ouvert du Web et celle de la croissance des métadonnées nécessaires à leur exploitation plus efficiente. Elle encourage les professionnels à réinvestir la gestion des référentiels terminologiques. L'auteur apporte une distinction entre trois types de projets liés aux modèles. Enfin, elle plaide pour que la gestion des données n'occulte pas les référentiels terminologiques.

[56] DELAHOUSSE J. « Faire vivre les données. 1 : formats et modèles ». *Documentaliste-Sciences de l'Information*. 2013. Vol. 50, n°3, p. 36-37.

L'auteur de cet article plaide pour le partage des ontologies, qu'il nomme modèles, afin de gagner en efficacité lors de leur construction. Il explique aussi l'impact du Web sémantique sur leur construction. Il constate par ailleurs que la multiplicité des modèles oblige ceux qui désirent publier des données à mener des arbitrages. Il aborde enfin la question des compétences nécessaires pour comprendre les jeux de données ainsi que celles nécessaires pour choisir et combiner les ontologies. Ces diverses entreprises comme celle aussi de leur création sont à envisager sous une forme collaborative.

[57] GANDON F., FARON-ZUCKER C., CORBY O. Le Web sémantique comment lier les données et les schémas sur le web ?. Paris : Dunod, 2012. 1 vol. (XIII-206 p.) p. (InfoPro). ISBN : 978-2-10-057294-6.

Cet ouvrage a été à de très nombreuses reprises consulté, tant pour comprendre l'histoire et les enjeux du Web sémantique et de son corollaire le Web de données, que pour consulter des définitions. Les trois auteurs, tous les trois informaticiens, s'adressent aussi aux développeurs web et illustrent leurs propos avec beaucoup d'exemples commentés de code informatique.

[58] ILLIEN G. « Décrire les objets du savoir, les nouveaux paradigmes du catalogage ». *Documentaliste-Sciences de l'Information*. 2013. Vol. 50, n°3, p. 26-29.

L'auteur fait le point sur les évolutions du travail de catalogage et sur le catalogue en tant qu'application au sein des bibliothèques. Il fait le constat que le résultat de ce travail reste caché dans le Web profond et invite à faire le pari du Web sémantique. Il dresse une liste des défis auxquels les professionnels doivent se préparer et pour lesquels il est nécessaire de faire évoluer leurs pratiques dans le sens des opportunités ainsi offertes.

[59] TRONCY R. « Owl, un « chouette » langage pour représenter des ontologies ». Documentaliste-Sciences de l'Information. 2011. Vol. 48, n°4, p. 34.

Un article synthétique où l'auteur explique rapidement la différence entre RDFS et OWL. Il constate la multiplicité des ontologies qui se sont répandues sur le Web et évoque enfin le vocabulaire des géants des moteurs de recherche : schema.org.

[60] VENTRESQUE V. Les mutations des collections numériques à l'heure du Web de données [En ligne].. Mémoire de Master 2 : Archives Numériques. [S.l.] : ENSSIB, 2013. 86 p. Disponible sur : < http://www.enssib.fr/bibliotheque-numerique/documents/64112-les-mutations-des-collections-numeriques-a-l-heure-du-web-de-donnees.pdf >

Ce mémoire a participé à la réflexion sur le Web sémantique auquel est en grande partie consacré ce mémoire. Certaines des pistes proposées en conclusion par son auteur ont fait l'objet d'une réflexion dans le cadre de cette étude de l'ontologie ToxNuc, à savoir celles des facteurs organisationnels, sociaux et économiques, mais aussi la question de la confiance dans les données en fonction du contexte et de la provenance sans rompre pour autant avec le principe de l'ouverture.

[61] VANDENBUSSCHE P.-Y., VATANT B. *Projet DataLift : de la donnée brute publiée vers la donnée sémantique interconnectée. Datalift D2.1 : méthodes et indicateurs pour la sélection d'ontologies fiables et utilisables* [En ligne]. 28 décembre 2011. Disponible sur : < https://gforge.inria.fr/docman/view.php/2935/7547/DataLiftD2.1-v1.1-2011-07-12.pdf >

Ce document est un support méthodologique destiné aux fournisseurs de données pour les conseiller dans leur choix d'une ontologie existante, que les auteurs appellent ici vocabulaire, ou pour sa création. Ils exposent ainsi ce qu'ils reconnaissent comme des critères fonctionnels de qualité des ontologies de domaine destinées à porter des jeux de données dans le Web de données. Ces réflexions sont issues du projet mené pour établir un catalogue d'ontologies, le LOV (Linked Open Vocabularies, http://lov.okfn.org). Ce catalogue est l'une des briques du projet Datalift développé pour élever les données (conversion, publication et interconnexion), afin qu'elles puissent intégrer le Web de données.

[62] ZACKLAD M. « Quelle formalisation pour les contenus culturels ? ». *Documentaliste-Sciences de l'Information*. Vol. 48, n°4, p. 40-41

L'auteur remet d'abord en question la posture qu'il qualifie de techno-centrée des groupes de normalisation qui promeuvent le recours aux ontologies. Il considère que les usages et les enjeux sociologiques, économiques et politiques ne sont pas assez pris en compte. Il considère les ontologies inadaptées pour décrire la sémiotique des contenus culturels. Il conclut que le couplage des ontologies aux autres SOCs alors en mutation et auxquels, contrairement à d'autres auteurs, il associe l'indexation sociale, pourrait être favorable pour l'accès aux contenus.

Le Web sémantique et les entreprises

[63] BARBAUX A. « L'industrie s'empare du web sémantique ». In : *usinenouvelle.com* [En ligne]. [s.l.] : [s.n.], 2010. Disponible sur : < http://www.usinenouvelle.com/article/l-industrie-s-empare-du-web-semantique.N128850 >

Cet article est intéressant car son auteur rappelle le fonctionnement des standards du Web sémantique. Elle montre comment les technologies du Web sémantique peuvent servir aux entreprises pour aligner leurs différents référentiels terminologiques et améliorer l'exploitation des connaissances enregistrées dans leurs différents systèmes d'information.

[64] DALBIN S. « Le web sémantique en entreprise : quelques cas d'usages ». *Documentaliste-Sciences de l'Information* [En ligne]. 2011. Vol. 48, n°4, p. 42-44. Disponible sur : < http://dx.doi.org/10.3917/docsi.484.0042 >

L'auteur de cet article s'appuie sur plusieurs exemples de projets dans les entreprises qui s'appuient sur des ontologies. Ces projets concernent les espaces numériques de travail par exemple. Ainsi, il y a de plus en plus d'activités de rédaction collaborative au sein de plateformes utilisant des feuilles de styles pour encoder et représenter l'information selon des référentiels. Les référentiels terminologiques eux, ont de plus en plus vocation à devenir communs à plusieurs applications ou services et à être déconnectés de ceux-ci. Cet article est intéressant pour comprendre les obstacles que peuvent rencontrer certains de ces projets et propose d'utiliser les standards du Web sémantique qui s'inscrivent dans une économie de la réutilisation pour diminuer les coûts.

[65] ECONOCOM-OSIATIS. *Les technologies sémantiques : quel avenir pour l'entreprise ? Etat des lieux et apports fonctionnels* [En ligne]. octobre 2013. 60 p. Disponible sur : < http://www.osiatis.com/medias/upload/files/fr_FR/livre-blanc-tech-semantiques-vfweb_1381992554.pdf >

Les auteurs de ce document reviennent sur l'histoire du web et sur les technologies du Web sémantique tout en expliquant aussi les techniques et les usages du TALN. Ce document a aussi pour avantage de proposer un glossaire. Un nouveau livre blanc devrait paraître avant la fin de de l'année 2014 et sera consacré à la démocratisation des usages du Web sémantique avec des études de cas et se penchera sur les apports de valeurs pour les entreprises.

[66] ECONOCOM-OSIATIS. *La valeur des technologies du Web Sémantique* [En ligne]. *Le blog des digital experts*. Disponible sur : < http://www.osiatis.com/blog/la-valeur-des-technologies-du-web-semantique >

Cet article récent sur le blog de la société Osiatis constate l'émergence des technologies sémantiques au sein des entreprises. Après une explication succincte des principes du Web sémantique mais aussi du TALN, deux cas clients sont exposés.

[67] SCIANDRA D., TÉMOIGNAGE RECUEILL PAR DOMINIQUE COTTE. « Fournir une vision agrégée et cohérente de l'information ». *Documentaliste-Sciences de l'Information*. 2013. Vol. 50, n°3, p. 57-58.

Dans cette interview, Diane Sciandra expose comment, au sein d'une grande entreprise où doivent interagir de nombreux métiers et spécialités, la « donnée produit » peut devenir le pivot d'un référentiel autrement composé des données « clients/fournisseurs » et des données financières, pour s'interconnecter avec les référentiels des autres systèmes et qu'ainsi chacun puisse conserver son intégrité et respecter les différentes vues des métiers sur les données. Pour harmoniser et contextualiser les données, il est nécessaire de spécifier les vocabulaires métiers. Là, il y a donc création d'ontologie non pour modéliser la connaissance mais pour modéliser les données produits au sein du cycle de production. Ce référentiel de données nécessite d'être maintenu pour rester en capacité de contrôler les données produits. Cet article est aussi intéressant car il conclut sur la nécessité « d'enrôler » les utilisateurs opérationnels pour assurer la bonne structuration des données.

7 Le programme Transversal Toxicologie Nucléaire

[68] DSV/DIR-BIODOC-CEA. *Bilan bibliométrique - programme Toxicologie Nucléaire 2001-2014 présenté lors de la Journée scientifique Toxicologie NeuroSpin/CEA Saclay*. 10 avril 2014.

Il s'agit d'un poster montrant les résultats bibliométriques du programme Transversal Toxicologie Nucléaire depuis ses origine en 2001 jusqu'à l'année en cours, 2014.

[69] GACHET V., RICHARD V. *Quelle pertinence pour un organisme public de recherche d'intégrer les réseaux sociaux numériques à sa stratégie d'information et de communication ?*. Mémoire de master professionnel « communication et technologie numérique ». S.l. : Paris Sorbonne - Paris IV (Celsa), Ecole des Mines d'Alès, 2011. 149 p.

Ce mémoire s'intéresse à l'opportunité quant à sa notoriété pour un organisme d'intégrer les réseaux sociaux numériques à la stratégie globale de communication, ainsi qu'aux risques qui y sont associés. Il a fourni des éléments d'information sur les différentes plateformes collaboratives et de capitalisation des connaissances mis en place au CEA.

[70] MÉNAGER M.-T., GARNIER-LAPLACE J., GOYFFON M., éd. *Toxicologie nucléaire environnementale et humaine*. Paris : Éd. Tec & Doc. : Lavoisier, 2009. XLIX-748 p.-[8] p. de pl. ISBN : 978-2-7430-1174-1.

Ce livre est une référence en toxicologie nucléaire et il fait état des premiers résultats du programme transversal piloté par le CEA. La préface et les pages d'introduction de cet ouvrage ont permis de réaliser une synthèse sur l'étude des modes d'action des toxiques sur le vivant et sur les enjeux de cette discipline afin de la présenter dans une des sections du mémoire. Par ailleurs, cet ouvrage a été utilisé au cours de la mission de stage pour vérifier les notations scientifiques des éléments chimiques de même que pour prendre connaissance de certaines définitions à partir de son glossaire. L'articulation des chapitres a permis de réfléchir à de potentielles relations à intégrer dans l'ontologie ToxNuc.

[71] Programme Transversal Toxicologie. Toxcea, [en ligne] < http://www.toxcea.org >

Il s'agit de la plateforme actuelle du programme. Ses contenus sont en partie publics et librement accessibles. Un lien est fait vers l'ancienne plateforme multi-organismes : http://www.toxnuc-e.org

[72] RICCIO P.-M., COMMANDRE M. « Une approche systémique et non-linéaire de l'émergence d'une communauté scientifique ». In : *TIC'IS 2009. Journées d'Etudes Dynamique de Réseaux, Information, Complexité et Non-Linéarité*. Marseille : [s.n.], 2009.

Les auteurs de cette communication expliquent comment, dans le cadre d'une recherche-action, ils ont mis en place une plateforme de travail collaboratif destinée à faciliter le travail des chercheurs du programme Transversal Toxicologie Nucléaire. Ils expliquent qu'il est nécessaire de conjointement élaborer un dispositif sociotechnique pour accompagner et encourager le travail collaboratif.

[73] RICCIO P.-M. « Les réseaux sémantiques étendus, une alternative aux ontologies dans la médiation entre parcours de recherche ou d'apprentissage et données préexistantes sur les réseaux numériques ». In : *Actes des journées d'étude du groupe de travail TIC-IS de la SFSIC* [En ligne]. *(TIC'IS 2010*. [s.l.] : [s.n.], 2010. Disponible sur :
< http://halshs.archives-ouvertes.fr/hal-00812616/ >

La lecture de cet article de colloque a été très utile pour remonter aux origines du projet d'ontologie et tracer la chronologie de sa conception.

8 Les mémoires INTD

Les thèmes de ces deux mémoires, issus de cette même formation, peuvent apporter un complément au présent mémoire.

[74] KELLER L. *Encadrer la réingénierie d'un thesaurus : méthode, enjeux et impacts pour l'équipe d'un service de veille et documentation en entreprise*. Titre professionnel « Chef de projet en ingénierie documentaire ». S.l. : Conservatoire National Des Arts Et Metiers (Cnam) - Institut National des Techniques de la Documentation (INTD), 2014.

[75] PIERRE B. *L'avenir des langages documentaires dans le cadre du Web sémantique : conception d'un thésaurus iconographique pour le Petit Palais*. Titre professionnel « Chef de projet en ingénierie documentaire ». S.l. : Conservatoire National Des Arts Et Metiers (Cnam) - Institut National des Techniques de la Documentation (INTD), 2010.

9 Les cours INTD

[76] ARRUABARRENA B. *Datavisualisation*, cours Titre 1 INTD 2013

[77] CARMES M. *Introduction générale*, cours Titre 1 INTD. 2013

[78] CARMES M. *Changement organisationnel et TIC : introduction à la sociologie des usages et synthèse sur les transformations technico-organisationnelles*, cours Titre 1 INTD. 2014

[79] CHARLET, J. *Ontologies, terminologie, Web sémantique (3)*, cours Titre 1 INTD. 2014.

[80] KEMBELLEC G., CHARTRON G. *Le système de recommandation : un nouveau médiateur des contenus ?* Cours Titre 1 INTD. 2014. 20 février 2014.

[81] NIANG-KEITA N. *Introduction au data mining*, cours n°1 Titre 1 INTD. 2014.

[82] NIANG-KEITA N. *Analyses multidimensionnelles de données numériques*, cours n°2 Titre 1 INTD. 2014.

[83] RAIS N. *Web, standard et enjeux techniques*, cours Titre 1 INTD. 2014

[84] RAIS N. *Internet : Les bases du Web de données*, cours Titre 1 INTD. 2014

[85] SAPORTA, Gilbert. *De l'analyse des données au Data mining et aux Big data*, cours Titre 1 INTD. 2014

[86] SAINT-LÉGER M. DE. *Fouille de données textuelles*, cours Titre 1 INTD. 2013

Annexes

Annexe 1
Livrable 3 - Etat de l'art : répertoire des ressources termino-ontologiques disponibles dans le domaine

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Plan

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

L'ontologie de la toxicologie nucléaire			
Acronyme	ToxNuc		
Editeur	Non éditée – en cours de développement		
Domaines couverts	toxicité nucléaire ; toxicité des éléments radioactifs ; toxiques nucléaires ; chimiotoxicité ; radiotoxicité ; toxicocinétique ; toxicodynamique ; écotoxicité ; toxicologie environnementale		
Date de création	2006	*Date de mise à jour*	août 2014
Nombre de classes	649	*Nombre d'instances*	
Nombre de propriétés	0	*Niveau de profondeur*	5
Format	OWL		
Utilisation de la recommandation SKOS	Oui, pour dénoter les concepts avec des termes préférentiels et des synonymes en français et en anglais.		
Documentation	Historique du projet d'ontologie de la toxicologie nucléaire / Véronique Gachet. – Novembre 2011. – 7 p.		
Licence d'utilisation	En cours de réflexion : Creative Commons, CeCILL…		

Remarques :

- Aucun lien avec d'autres ontologies
- Pas de relations sémantiques spécifiques au domaine
- Une structure arborescente du vocabulaire du domaine caractérisée par des relations hiérarchiques de subsomption de type « is_a », c'est-à-dire des relations génériques spécifiques. Cependant, il reste à vérifier s'il n'y a pas de concepts liés par méronymie, soit vérifier la présence de relations de subsomption de type « part-of » dites aussi partitives. Dans ce cas des corrections seraient nécessaires.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

L'ontologie de la toxicologie nucléaire			
Acronyme	ToxNuc		
Editeur	Non éditée – en cours de développement		
Domaines couverts	toxicité nucléaire ; toxicité des éléments radioactifs ; toxiques nucléaires ; chimiotoxicité ; radiotoxicité ; toxicocinétique ; toxicodynamique ; écotoxicité ; toxicologie environnementale		
Date de création	2006	*Date de mise à jour*	août 2014
Nombre de classes	649	*Nombre d'instances*	
Nombre de propriétés	0	*Niveau de profondeur*	5
Format	OWL		
Utilisation de la recommandation SKOS	Oui, pour dénoter les concepts avec des termes préférentiels et des synonymes en français et en anglais.		
Documentation	Historique du projet d'ontologie de la toxicologie nucléaire / Véronique Gachet. – Novembre 2011. – 7 p.		
Licence d'utilisation	En cours de réflexion : Creative Commons, CeCILL…		

Remarques :

- Aucun lien avec d'autres ontologies
- Pas de relations sémantiques spécifiques au domaine
- Une structure arborescente du vocabulaire du domaine caractérisée par des relations hiérarchiques de subsomption de type « is_a », c'est-à-dire des relations génériques spécifiques. Cependant, il reste à vérifier s'il n'y a pas de concepts liés par méronymie, soit vérifier la présence de relations de subsomption de type « part-of » dites aussi partitives. Dans ce cas des corrections seraient nécessaires.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Les ressources termino-ontologiques dans le domaine

Il n'existe pas d'ontologie à l'heure actuelle ni même de terminologie (taxinomie, thésaurus, classification) consacrée à la toxicologie nucléaire environnementale.

Dix ressources ont été sélectionnées. Elles sont classées en fonction de leur proximité avec le domaine de la toxicologie nucléaire : toxicologie, environnement, nanosciences, biomédical, puis par ordre alphabétique pour chacun de ces domaines.

Toxicologie :

Environnement :

Nanosciences :

Biomédical :

Pour chacune, il s'agit de présenter des éléments d'identification de ces ressources. Il s'agit aussi d'indiquer des éléments qui puissent contribuer à l'évaluation de leur intérêt pour l'ontologie du collectif ToxNuc.

Deux autres ressources non ontologiques référencées sur BioPortal sont également présentées :

La méthode de recherche et le choix des éléments de description sont présentés dans le chapitre « La démarche et les critères de sélection » p. 18.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

OpenTox Ontology			
Recommandation SKOS	Non (les synonymes font l'objet d'une classe)	*Format*	OWL
Editeur	OpenTox FP7 EU project Project coordinator Dr. Barry Hardy : barry.hardy@douglasconnect.com, +41 61 851 0170		
Domaines couverts	toxicologie prédictive ; effets toxicologiques ; anatomie ; modèles animaux		
Documentation	http://www.opentox.org		
Date des fichiers			

Fichiers	*Nombre de classes*	*Nombre de propriétés*
toxicological_endpoints_ontology.owl	175	7**
all_organ_systems_ontology.owl	898	9***
effects_respiratory_tract_ontology.owl	547	11****
Nombre de projets 1	*Nombre de publications**	10
Visualisation	La visualisation de l'ontologie nécessite d'installer l'extension Collaborative Protégé et la création d'un compte.	
Licence d'utilisation		

* Nombre de publications signalées dans PubMed, 11 sur TOXLINE

**	***	****
has_species_sex	has_example	is_a_synonym_of
has_species_strain	has_part	has_components_of
has_test_condition	has_the_abreviation	has_limitations
has_test_result	is_a_synonym_of	has_part
has_test_species	is_an_example_for	hasRuleGroup
has_tumor_site	is_part_of	is_also_in
objectProperty_1	is_produced_by	is_characterized_by
	is_the_abreviation_of	is_described_in
	poduces	is_effect_for_species
		is_effect_in
		is_part_of

Remarques :

- La communauté OpenTox s'efforce de suivre les principes de l'OBO Foundry.
- L'ontologie est téléchargeable en plusieurs fichiers.
- L'ontologie est composée des classes suivantes :
  - "ToxicityStudyType" (avec pour concepts définis « aquatic toxicity studies », « bacterial mutagenis studies », « carcinogenicity studies », « in vivo micronucleus studies », « repeated dose toxicity studies »)
  - "TestSystem" (avec pour sous-classes « strain »s, « species », « sex », « route of explosions », etc.)
  - "TestResult" (avec pour sous-classes « toxicity measure », « test call », « mode of action », etc.)
  - "Organs" (décrivant jusqu'aux composants histologiques, aucun alignement avec des ontologies consacrées à la description anatomique)
- Elle a pour objectif d'être une ontologie de haut niveau, compatible avec l'ontologie formelle Basic Formal Ontology (BFO).
- Elle est conçue avec l'idée d'être modulaire.
- Plusieurs modèles animaux présents dans ToxNuc sont présents dans cette ontologie.
- La version OWL est téléchargeable depuis le site consacré à cette ontologie à l'adresse suivante : http://www.opentox.org/dev/Ontology.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Environment Ontology					
Acronyme	EnvO	*Format*	OBO et OWL	*Recommandation SKOS*	Oui
Editeur	Communauté universitaire ouverte obo-envo@lists.sourceforge.net				
Domaines couverts	environnement ; sciences de la vie				
*Date du 1er dépôt**	2013		*Date du dernier dépôt*		février 2014
Nombre de classes	1 397		*Nombre d'instances*		1
Nombre de propriétés	21		*Niveaux de profondeur*		11
Documentation	http://environmentontology.org Buttigieg P. L., Morrison N., Smith B., Mungall C. J., Lewis S. E. « The environment ontology: contextualising biological and biomedical entities ». Journal of Biomedical Semantics [En ligne]. 11 décembre 2013. Vol. 4, n°1, p. 43. Disponible sur : < http://dx.doi.org/10.1186/2041-1480-4-43 > (consulté le 27 août 2014). PMID: 24330602				
*Nombre de projets***	7		*Nombre de publications****		4
Description BioPortal	http://bioportal.bioontology.org/ontologies/ENVO				
Visualisation BioPortal	http://purl.bioontology.org/ontology/ENVO				
Licence d'utilisation					

* sur BioPortal
** Nombre de projets s'y référant d'après les déclarations faites sur BioPortal
*** Nombre de publications signalées dans PubMed

Remarques :

- 27% des concepts n'ont pas de définitions associées.
- Un rapide sondage a permis de ne trouver que très peu de termes propres à ToxNuc qui soient en commun avec ceux présents dans cette ontologie.
- Elle est en partie alignée avec l'ontologie formelle Basic Formal Ontology (BFO).
- La version OWL est téléchargeable depuis le site consacré à cette ontologie à l'adresse suivante : http://www.environmentontology.org/downloads.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

GEneral Multilingual Environmental Thesaurus					
Acronyme	gemet_oaei	*Format*	OWL	*Recommandation SKOS*	Non
Editeur	Fichier mis à disposition des participants de l'OAEI2007 de l'Ontology AligmentEvaluation Initiative (OAEI). http://wrvhage.nl/ w.r.van.hage@vu.nl				
Domaines couverts	environnement ; écologie				
*Date du fichier repéré**	2007				
Nombre de classes			*Nombre d'instances*		
Nombre de propriétés			*Niveaux de profondeur*		
Documentation	« The GEMET thesaurus has three types of top concepts: themes, groups, and supergroups. In the OWL version these are represented as owl:Class, like normal concepts. In addition to this, they also have an rdf:type gemet:Theme, gemet:Group, or gemet:SuperGroup. » http://oaei.ontologymatching.org/2007/environment/				
Licence d'utilisation					

* Fichier repéré lors d'une recherche avec le moteur généraliste Google

Remarques :

- Le fichier n'a pas été étudié car il n'a pu être ouvert avec Protégé 4.3.0.
- D'après l'interrogation du moteur Scholar Google, plusieurs articles font état de projet s'inspirant du GEMET pour constituer des ontologies spécifiques.
- GEMET est un thésaurus multilingue de plus de 5 500 termes. Il est constitué par l'European Environment Information and Observation Network de l'European Environment Agency et consultable à l'adresse : http://www.eionet.europa.eu/gemet/fr/themes

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Semantic Web for Earth and Environment Technology Ontology					
Acronyme	SWEET	*Format*	OWL	*Recommandation SKOS*	Oui
Editeur	Développée par le Jet Propulsion Laboratory de la NASA et sous la responsabilité de la Earth Science Information Partners foundation (ESIP foundation). http://wiki.esipfed.org/index.php/SWEET_Governance Thomas.Huang@jpl.nasa.gov				
Domaines couverts	sciences de la terre ; sciences de l'environnement				
*Date du 1er dépôt**	2012		*Date du dernier dépôt*		
Nombre de classes	4 549		*Nombre d'instances*	2 152	
Nombre de propriétés	359		*Niveaux de profondeur*	11	
Documentation	http://sweet.jpl.nasa.gov/				
*Nombre de projets***	0		*Nombre de publications****	1[1]	
Description BioPortal	http://bioportal.bioontology.org/ontologies/SWEET				
Visualisation BioPortal	http://purl.bioontology.org/ontology/SWEET				
Licence d'utilisation					

*sur BioPortal
** Nombre de projets s'y référant d'après les déclarations faites sur BioPortal
*** Nombre de publications signalées dans PubMed

Remarques :

- 63% des concepts n'ont pas de définitions associées.
- Certains termes utilisés pour désigner les concepts de la classe « ModeleBiologique » de ToxNuc sont présents dans SWEET.

[1] Cette publication fait référence à la SWEET ontology en la nommant « Semantic Web for Earth and Environmental Terminology ». Google scholar référence 194 publications.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

NanoParticle Ontology					
Acronyme	NPO	*Format*	OWL	*Recommandation SKOS*	Non
Editeur	Soutenue initialement par le National Institute Of Health (US NIH) et actuellement utilisé et mise à jour par le NCI caBIG Nanotechnology Working Group.				
Domaines couverts	Caractéristiques physiques, chimiques et fonctionnelles de la nanotechnologie utilisée dans le diagnostic et le traitement du cancer.				
*Date du 1er dépôt**	2008		*Date du dernier dépôt*	2012	
Nombre de classes	1 904		*Nombre d'instances*		
Nombre de propriétés	81		*Niveaux de profondeur*	16	
Documentation	http://www.nano-ontology.org				
*Nombre de projets***	3		*Nombre de publications****	3	
Description BioPortal	http://bioportal.bioontology.org/ontologies/NPO				
Visualisation BioPortal	http://purl.bioontology.org/ontology/NPO				
Licence d'utilisation					

*sur BioPortal
** Nombre de projets s'y référant d'après les déclarations faites sur BioPortal
*** Nombre de publications signalées dans PubMed

Remarques :

- 63% des concepts n'ont pas de définitions associées.
- Elle est basée sur l'ontologie formelle Basic Formal Ontology (BFO).
- C'est principalement les termes qui désignent les éléments chimiques dans ToxNuc qui sont retrouvés dans NPO.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Bilingual Ontology of Alzheimer's Disease and Related Diseases (English-French)					
Acronyme	OntoAD	*Format*	OWL	*Recommandation SKOS*	Oui
Editeur	Khadim Dramé, Khadim.Drame@isped.u-bordeaux2.fr				
Domaines couverts	maladie d'Alzheimer ; maladies neurodégénératives				
*Date du 1er dépôt**	Octobre 2009		*Date du dernier dépôt*	Octobre 2009	
Nombre de classes	5 899		*Nombre d'instances*	2 465	
Nombre de propriétés	182		*Niveaux de profondeur*	24	
Documentation	Voir la publication signalée dans PubMed[2]				
*Nombre de projets***	0		*Nombre de publications****	1[6]	
Description BioPortal	http://bioportal.bioontology.org/ontologies/ONTOAD				
Visualisation BioPortal	http://purl.bioontology.org/ontology/ONTOAD				
Licence d'utilisation					

*sur BioPortal
** Nombre de projets s'y référant d'après les déclarations faites sur BioPortal
*** Nombre de publications signalées dans PubMed

Remarques :

- Taxonomie du domaine et des relations formelles non-taxonomiques plus spécifiques.
- Version BETA.
- 46 % des classes sont sans définition.
- Cette ontologie doit être implémentée dans un portail sémantique, Semantic BiblioDem Portal (SemBiP), qui permettra l'interrogation de la base de données bibliographique BiblioDem[6].
- Un sondage rapide montre que des termes désignant des molécules sont communs entre cette ressource et ToxNuc.

[2] Dramé K., Diallo G., Delva F., Dartigues J. F., Mouillet E., Salamon R., Mougin F. « Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology: an application to Alzheimer's disease ». *J Biomed Inform* [En ligne]. avril 2014. Vol. 48, p. 171‑182. Disponible sur : < http://dx.doi.org/10.1016/j.jbi.2013.12.013 > (consulté le 19 août 2014). PMID 2438242

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Computer Retrieval of Information on Scientific Projects Thesaurus					
Acronyme	CRISP	*Format*	UMLS[3]	*Recommandation SKOS*	Oui
Editeur	Anita Ghebeles, af8d@nih.gov				
Domaine couvert	Sciences biomédicales				
*Date du 1er dépôt**	2013		*Date du dernier dépôt*		mai 2014
Nombre de classes	9 045		*Nombre d'instances*		0
Nombre de propriétés	0		*Niveaux de profondeur*		12
Documentation					
*Nombre de projets***	0		*Nombre de publications****		0 (6 avec Google scholar)
Description BioPortal	http://bioportal.bioontology.org/ontologies/CRISP				
Visualisation BioPortal	http://purl.bioontology.org/ontology/CRISP				
Licence d'utilisation	« This ontology is made available via the UMLS. Users of all UMLS ontologies must abide by the terms of the UMLS license, available at https://uts.nlm.nih.gov/license.html »				

*sur BioPortal
** Nombre de projets s'y référant d'après les déclarations faites sur BioPortal
*** Nombre de publications signalées dans PubMed

Remarques :

- 78% des concepts n'ont pas de définitions associées.
- C'est une ontologie avec une expressivité formelle de type « is_a » uniquement.
- Nombre de termes utilisés pour désigner les concepts de ToxNuc sont présents dans CRISP avec des définitions, des synonymes, comme ceux des classes « ElementChimique », « ModeleVegetal », « MoleculeOrganique », « Organisme ».

[3] Un fichier au format CSV est signalé sur BioPortal pour la dernière version. Cependant, il semble impossible de le télécharger.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Ontologie pour le National Cancer Institute Thesaurus					
Acronyme	NCIt	*Format*	OWL DL	*Recommandation SKOS*	Non
Editeur	Enterprise Vocabulary Services (EVS) – US National Cancer Informatics Program - Center for Biomedical Informatics and Information Technology (CBIIT) http://cbiit.nci.nih.gov/ncip/biomedical-informatics-resources/interoperability-and-semantics/terminology				
Domaines couverts	soin clinique ; recherche fondamentale ; recherche appliquée ; information publique ; activité administrative				
*Date du 1er dépôt**	2007		*Date du dernier dépôt**		juillet 2014
Nombre de classes	108 256		*Nombre d'instances*		34782
Nombre de propriétés	173		*Niveaux de profondeur*		16
Documentation	http://nciterms.nci.nih.gov/ncitbrowser/pages/home.jsf?version=14.06e				
*Nombre de projets***	14		*Nombre de publications****		11
Description BioPortal	http://bioportal.bioontology.org/ontologies/NCIT				
Visualisation BioPortal	http://purl.bioontology.org/ontology/NCIT				
Licence d'utilisation	« The version of the NCI Thesaurus (NCIt) available in BioPortal has been modified by reformatting some property values so that they can be more easily browsed (replacing or removing embedded XML). The original, unmodified NCIt, as well as NCIt license information, is available at http://ncit.nci.nih.gov. »				

*sur BioPortal
** Nombre de projets s'y référant d'après les déclarations faites sur BioPortal
*** Nombre de publications signalées dans PubMed

Remarques :

- Certaines propriétés pourraient inspirer des relations à établir entre les instances de classes de ToxNuc comme par exemple « Biological_Process_Has_Associated_Location ».
- 35% des classes sont sans définition.
- De nombreux concepts de ToxNuc sont présents dans la sous-classe ayant pour label « Element », elle-même composée des sous-classes ayant pour labels « Radioisotope », « Metal », « Radioactive Element », « Trace Element ».
  Il existe aussi la classe « Nanoparticle » qui est également subdivisée ainsi qu'une classe « Radiopharmaceutical Compound ».
  Pour la grande majorité de ces concepts, il existe des annotations : la formule chimique ; plusieurs labels et définitions ; les identifiants uniques de la terminologie UMLS[4], de la banque de données CAS[5] et de la base de données ChEBI[6].

[4] Conçu en 1986 et maintenu à jour par la National Library of Medicine (US NLM), l'Unified Medical Language System (UMLS) fournit une mise en correspondance entre les termes des différents vocabulaires contrôlés pour les sciences biomédicales. http://www.nlm.nih.gov/research/umls/
[5] L'American Chemical Society (ACS) maintient et commercialise une base de données CAS Registry dans laquelle sont répertoriées et décrites des substances chimiques. http://www.cas.org
[6] Chemical Entities of Biological Interest (ChEBI) est une base de données et une ontologie des entités moléculaires. C'est une ressource produite par l'European Bioinformatics Institute qui dépend de l'European Molecular Biology Laboratory (UK EMBL-EBI). http://www.ebi.ac.uk/chebi/

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Neuroscience Information Framework (NIF) Standard Ontology					
Acronyme	NIFSTD	*Format*	UMLS[7]	*Recommandation SKOS*	Oui
Editeur	Fahim Imam, mimam@ucsd.edu The Neuroscience Information Framework (NIF) is a dynamic inventory of Web-based neuroscience resources: data, materials, and tools accessible via any computer connected to the Internet. An initiative of the US National Institutes of Health (NIH) Blueprint for Neuroscience Research. http://neuinfo.org				
Domaine couvert	Neuroscience				
*Date du 1er dépôt**	2012		*Date du dernier dépôt*		mars 2013
Nombre de classes	108 427		*Nombre d'instances*		553
Nombre de propriétés	625		*Niveaux de profondeur*		30
Documentation	http://neuinfo.org/#vocab				
*Nombre de projets***	5		*Nombre de publications****		3 (voir aussi http://neuinfo.org/about/publications.shtml)
Description BioPortal	http://bioportal.bioontology.org/ontologies/NIFSTD				
Visualisation BioPortal	http://purl.bioontology.org/ontology/NIFSTD				
Licence d'utilisation					

*sur BioPortal
** Nombre de projets s'y référant d'après les déclarations faites sur BioPortal
*** Nombre de publications signalées dans PubMed

Remarques :

- 63% des concepts n'ont pas de définitions associées.
- Des termes présents pour désigner les concepts des sous-classes de « MoleculeOrganique » de ToxNuc sont présents dans NIFSTD.

[7] Un fichier au format CSV est signalé sur BioPortal pour la dernière version. Cependant, il semble impossible de le télécharger.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Systematized Nomenclature of Medicine - Clinical Terms (ontologie multilingue[8])					
Acronyme	SNOMED CT	*Format*	UMLS	*Recommandation SKOS*	Oui (partiellement)
Editeur	International Health Terminology Standards Development Organisation (IHTSDO). http://ihtsdo.org				
Domaines couverts	soin clinique ; domaine vétérinaire ; systèmes de santé ; pharmacologie ; physique ; chimie ; géographie ; ethnographie				
*Date du 1er dépôt**	2009		*Date du dernier dépôt*		février 2014
Nombre de classes	401 214		*Nombre d'instances*		Plus de 800 000
Nombre de propriétés			*Niveaux de profondeur*		28
Documentation	http://www.ihtsdo.org/snomed-ct/snomed-docs/				
*Nombre de projets***	20		*Nombre de publications****		123
Description BioPortal	http://bioportal.bioontology.org/ontologies/SNOMEDCT				
Visualisation BioPortal	http://purl.bioontology.org/ontology/SNOMEDCT				
Licence d'utilisation	« This ontology is made available via the UMLS. Users of all UMLS ontologies must abide by the terms of the UMLS license, available at https://uts.nlm.nih.gov/license.html »				

*sur BioPortal
** Nombre de projets s'y référant d'après les déclarations faites sur BioPortal
*** Nombre de publications signalées dans PubMed

Remarques :

- C'est la plus grande des ontologies médicales et il s'agit de la version ontologique de la terminologie de référence SNOMED[9].
- Expressivité formelle de type « is_a » uniquement
- 99,8 % des classes sont sans définition
- Nombreux doublons[10].

[8] Vandenbussche P.-Y., Charlet J. « Méta-modèle général de description de ressources terminologiques et ontologiques ». In : Actes des 20es Journées Francophones d'Ingénierie des Connaissances-IC2009 [En ligne]. Journées francophones d'ingénierie des connaissances. Grenoble : Presses universitaires de Grenoble, 2009. Disponible sur : < http://hal.archives-ouvertes.fr/hal-00379935/ > (consulté le 25 juin 2014). ISBN : 978-2-7061-1538-7.
« Today, SNOMED CT is available in US English, UK English, Spanish, Danish and Swedish. Translations into French, Lithuanian, and several other languages are currently taking place. ». Disponible sur : < http://www.ihtsdo.org/snomed-ct/snomed-ct0/different-languages/ > (consulté le 19 août 2014)
[9] Mazuel L., Charlet J. « Alignement entre des ontologies de domaine et la Snomed: trois études de cas. » In : Actes des 20es Journées Francophones d'Ingénierie des Connaissances-IC2009 [En ligne]. Journées francophones d'ingénierie des connaissances. Grenoble : Presses universitaires de Grenoble, 2009. Disponible sur : < http://ic2009.inria.fr/docs/papers/MazuelCharlet_IC2009_16.pdf > (consulté le 16 juin 2014). ISBN : 978-2-7061-1538-7.
[10] « The SNOMED CT medical terminology ontology alone contains 370,000 class names, and existing technology has not yet been able to eliminate all semantically duplicated terms. » Semantic Web [En ligne]. Wikipedia, the free encyclopedia. 13 août 2014. Disponible sur : < http://en.wikipedia.org/w/index.php?title=Semantic_Web&oldid=620118637 > (consulté le 19 août 2014)

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

International Nuclear Information System / Energy Technology Data Exchange Thesaurus (ressource non ontologique)					
Acronyme	INIS/ETDE Thesaurus	*Format*	PDF et interactif	*Recommandation SKOS*	Non
Editeur	International Atomic Energy Agency (AIEA) http://www.iaea.org				
Domaines couverts	physique (en particulier, la physique des plasmas, la physique atomique et moléculaire, et en particulier la physique nucléaire et des hautes énergies) ; chimie ; sciences des matériaux ; sciences de la terre ; biologie de rayonnement ; effets de radio-isotopes et cinétique ; sciences de la vie appliquées, de radiologie et de médecine nucléaire ; isotope et technologie de source de rayonnement ; radioprotection ; applications des rayonnements, de l'ingénierie, de l'instrumentation ; combustibles fossiles ; combustibles synthétiques ; sources d'énergie renouvelables ; systèmes énergétiques de pointe ; technologie de fission ; technologie de réacteur de fusion ; gestion des déchets ; aspects environnementaux de la production et de la consommation d'énergie à partir de sources nucléaires et non-nucléaires ; efficacité énergétique ; conservation de l'énergie ; économie et sociologie de la production et l'utilisation d'énergie ; politique énergétique ; droit nucléaire				
Date de création	1966	*Date de la dernière édition*	août 2014		
Nombre de termes	30 000				
Langue	Multilingue (allemand, anglais, arabe, chinois, espagnol, français, japonais, russe)				
Documentation	http://www.iaea.org/inis/products-services/thesaurus				
Interrogation en ligne	https://nkp.iaea.org/INISMLThesaurus https://nkp.iaea.org/Thesaurus				
Versions PDF consultées pendant le stage	Dictionnaire français-anglais sans hiérarchie du thésaurus (2013) : http://www.iaea.org/inis/products-services/publications/INIS-ETDE-01/2013/IAEA-INIS-ETDE-01-fr-en-2013-01.pdf Thésaurus en français (2014) : http://www.iaea.org/inis/products-services/publications/etde_inis_2_7_fr_rev_2_4.pdf				

Remarque :

- Thésaurus non utilisé par le BioDoc, Direction des sciences du vivant du CEA (CEA-DSV). Indexation des publications des programmes selon une classification transmise par la Direction de la stratégie et des programmes.
- Les professionnels de l'IST du CEA qui ont participé à chacune des rencontres annuelles des professionnels de l'IST à Nancy tenues entre 2002 et 2009 n'ont pas fait part de projet de mise à disposition de ce thésaurus dans un format du Web de sémantique[11].

[11] http://www.isore.cnrs.fr/spip.php?rubrique51

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Medical Subject Headings (ressource non ontologique)					
Acronyme	MESH	*Format*	UMLS	*Recommandation SKOS*	Non
Editeur	National Library of Medicine (NLM), US National Institutes of Health (NIH) NLM Customer Service, custserv@nlm.nih.gov				
Domaines couverts	médecine ; santé ; sciences de la vie ; sciences humaines ; sciences sociales				
*Date du 1er dépôt**	2009	*Date du dernier dépôt*	juillet 2014		
Nombre de classes	245 871	*Nombre d'instances*			
Nombre de propriétés	0	*Niveaux de profondeur*			
Documentation	http://www.nlm.nih.gov/mesh/meshhome.html				
*Nombre de projets***	12				
Description BioPortal	http://bioportal.bioontology.org/ontologies/MESH				
Visualisation BioPortal	http://purl.bioontology.org/ontology/MESH				
Licence d'utilisation	« This ontology is made available via the UMLS. Users of all UMLS ontologies must abide by the terms of the UMLS license, available at https://uts.nlm.nih.gov/license.html »				

*sur BioPortal
** Nombre de projets s'y référant d'après les déclarations faites sur BioPortal

Remarques :

- Vocabulaire contrôlé du thésaurus créé et maintenu par la National Library of Medicine (NLM).
- Les relations hiérarchiques de subsomption telles que les donne à visualiser BioPortal sont de type « is_a », or comme il s'agit d'un thésaurus les relations peuvent aussi être en réalité de type « part_of », ce qui peut conduire à des interprétations étranges comme « Un doigt est une main ».
- 89% des concepts n'ont pas de définitions associées.
- Depuis 2013, Claudie Hasenfuss de l'Inserm (hasenfus@vjf.inserm.fr) dépose des fichiers de versions anglais/français en UMLS, et depuis 2014 en CSV.
- Version ALPHA.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Psychology Ontology (ressource non ontologique)					
Acronyme	APAONTO	*Format*	OWL	*Recommandation SKOS*	Non
Editeur	American Psychological Association (APA) alexander garcia, alexgarciac@gmail.com Ian Galloway, igalloway@apa.org				
Domaines couverts	psychologie ; psychiatrie				
*Date du 1er dépôt**	février 2014		*Date du dernier dépôt*	juillet 2014	
Nombre de classes	8 947		*Nombre d'instances*		
Nombre de propriétés	0		*Niveaux de profondeur*	1	
Documentation					
*Nombre de projets***	0		*Nombre de publications****	0	
Description BioPortal	http://bioportal.bioontology.org/ontologies/APAONTO				
Visualisation BioPortal	http://purl.bioontology.org/ontology/APAONTO				
Licence d'utilisation					

*sur BioPortal
** Nombre de projets s'y référant d'après les déclarations faites sur BioPortal
*** Nombre de publications signalées dans PubMed

Remarques :

- Terminologie non structurée.
- Terminologie à plat du thésaurus de l'APA : Thesaurus of Psychological Index Terms®.
- Version ALPHA.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

La démarche et les critères de sélection des ressources termino-ontologiques

Des sources ontologiques peuvent être trouvées en faisant des requêtes dans les moteurs de recherche généralistes avec les combinaisons suivantes :

"nuclear toxicology" filetype:owl
+toxic* +ontolog*
"termes des classes principales ou sous-classes de ToxNuc" + ontolog*
"termes des classes principales ou sous-classes de ToxNuc" + owl

Les ontologies ainsi trouvées sont difficiles à identifier : auteur, date, objectifs qui ont participé de leur création, métriques et popularité sont autant d'éléments difficiles à rassembler.
C'est la raison pour laquelle le choix a été de conduire les principales recherches depuis un site estimé par la communauté des chercheurs en ingénierie des connaissances du domaine biomédical comme **BioPortal**. Il compte à ce jour le plus grand nombre d'ontologies biomédicales référencées. Ainsi les 379 ontologies qu'il comptabilise recouvrent les 149 ontologies répertoriées par le site **Ontobee**. Il en est de même pour les 93 ontologies interrogeables depuis le portail **Ontology Lookup Service** (août 2014).

Repérer une ontologie de la toxicologie ou de la toxicologie nucléaire sur BioPortal

La première démarche a été d'interroger le moteur d'ontologie qui effectue la recherche sur le nom des ontologies répertoriées. Aucune ontologie se rapportant à la toxicologie ou à la notion de toxicité n'est répertoriée. Cependant, la liste des projets référencés par le portail mentionne celui nommé « **OpenTox Ontology** » enregistré par OpenTox FP7 EU project. D'après la déclaration dont il fait l'objet, ce projet ne s'inspire d'aucune ontologie existante. Un lien renvoie vers la page Web de ce projet financé par l'Union européenne.

Sélectionner des ontologies à partir des caractéristiques de l'ontologie ToxNuc sur BioPortal

L'outil de recommandation de BioPortal a également été utilisé afin de repérer des ontologies proposant des concepts en commun avec ToxNuc.

Pour chaque recherche, cet outil suggère une liste de 25 ontologies en fonction d'une liste de mots-clés ou d'un texte d'un minimum de 50 mots soumis par l'utilisateur. Les critères de sélection sont au nombre de trois[12] :

- le taux de recouvrement entre le texte soumis et les ontologies référencées sur le portail
- le taux de connectivité des ontologies, c'est-à-dire celles qui utilisent le plus les concepts des autres ontologies
- la taille des ontologies, soit celles qui ont le plus grand nombre de concepts

Recherche par liste de mots-clés :

Les termes contenus dans le fichier Excel de l'ontologie validée en 2006 sont exclusivement en langue française. Or les ontologies référencées sur BioPortal sont en anglais. Les termes de ToxNuc ont donc été exportés puis traduit en anglais à l'aide du service gratuit de traduction en ligne de Google[13].

[12] Jonquet C., Musen M. A., Shah N. H. « Building a biomedical ontology recommender web service ». *J Biomed Semantics* [En ligne]. 2010. Vol. 1 Suppl 1, p. S1. Disponible sur : < http://dx.doi.org/10.1186/2041-1480-1-S1-S1 > (consulté le 12 août 2014)

[13] Ce service de traduction, comme tous les services gratuits de ce genre n'est pas totalement fiable. Il a cependant l'avantage de fournir très rapidement une traduction. Il a été préféré au service gratuit disponible sur le site Reverso (http://www.reverso.net/text_translation.aspx?lang=FR). En effet ce dernier se révèle limité par le nombre de caractères qu'il est possible de lui soumettre.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Recherche par texte d'un minimum de 50 mots :

Plusieurs textes ont été soumis au moteur de recommandation. Tous ont auparavant été traduits avec le service gratuit de traduction en ligne de Google :

- l'avant-propos et l'introduction générale de l'ouvrage de référence français sur la toxicologie nucléaire environnementale et humaine paru en 2009[14]. En effet, un identifiant de connexion à la plateforme Toxcea.org a permis d'accéder facilement à ces contenus dans une forme numérique exploitable.
- l'article « Toxicologie nucléaire » de l'encyclopédie collaborative Wikipédia. En effet, à cette date, il n'existe pas d'article sur la toxicologie nucléaire dans les pages en langue anglaise de Wikipédia.

Sont listées ici les 10 premières ontologies sur les 25 du classement généré par l'outil de recommandation de BioPortal :

1. Medical Subject Headings (MESH)
2. National Cancer Institute Thesaurus (NCIT)
3. Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT)
4. Bilingual Ontology of Alzheimer's Disease and Related Diseases (ONTOAD)
5. Psychology Ontology (APAONTO)
6. Computer Retrieval of Information on Scientific Projects Thesaurus (CRISP)
7. Neuroscience Information Framework (NIF) Standard Ontology (NIFSTD)
8. Semantic Web for Earth and Environment Technology Ontology (SWEET)
9. NanoParticle Ontology (NPO)
10. Read Codes, Clinical Terms Version 3 (CTV3) (RCD) : ressource non présentée dans le chapitre « **Les ressources termino-ontologiques dans le domaine** », pour laisser place à la présentation de trois autres ontologies. Il s'agit de celles de l'« OpenTox FP7 EU project » repérées grâce à la liste des projets enregistrés sur BioPortal, ainsi que deux ontologies du domaine de l'environnement, **Environment Ontology (ENVO)** et **GEneral Multilingual Environmental Thesaurus (GEMET)**.

Repérer une ou des ontologies du domaine de l'environnement

Lors de la recherche sur BioPortal d'ontologie présentant des similarités avec ToxNuc, l'ontologie **Semantic Web for Earth and Environment Technology Ontology (SWEET)** a été repérée. Mais c'est aussi le terme « environment » pour interroger le moteur d'ontologie qui a permis la sélection de l'ontologie **Environment Ontology (ENVO)**. Une recherche via les moteurs généralistes a permis de récupérer un fichier OWL pour le thésaurus **GEneral Multilingual Environmental Thesaurus (GEMET)**. Il est développé par l'European Environment Agency (EEA) et les membres du réseau European Environment Information and Observation Network (Eionet). Il faut ensuite les évaluer selon les critères présentés précédemment, comme le recoupement de concepts entre ces terminologies et celle du collectif ToxNuc.

Il est à noter que l'US Environmental Protection Agency (EPA) consacre une partie de ces activités au développement de services de terminologie au sein de l'une de ses directions : Office of Environmental Information's Data Standards Branch (ETSS_Admin@epamail.epa.gov)[15]. Ce service de l'US Environmental Protection Agency (EPA) rend disponible à la consultation glossaire, thésaurus et taxonomies. Il est possible d'exporter et d'importer ces ressources. Les formats proposés sont les suivants : RTF, XML, HTML, PDF, SKOS. Des Web services sont également mis à disposition. Le site annonce que des ontologies sont en cours de développement et seront bientôt disponibles[16].

[14] Ménager M.-T., Garnier-Laplace J., Goyffon M. (éd.). Toxicologie nucléaire environnementale et humaine. Paris, France : Éd. Tec & Doc. : Lavoisier, 2009. XLIX-748 p.-[8] p. de pl. p. ISBN : 978-2-7430-1174-1.

[15] http://ofmpub.epa.gov/sor_internet/registry/termreg/home/overview/home.do ou http://iaspub.epa.gov/sor_internet/registry/termreg/home/overview/home.do

[16] « Ontologies (terms connected by a rich web of relationships) are not currently available but will be included as they are developed. » http://ofmpub.epa.gov/sor_internet/registry/termreg/home/whatisterminology/

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Les outils de recherche de sources ontologiques

Les portails, entrepôts et moteurs de recherche d'ontologies dans les disciplines biomédicales

BioPortal http://bioportal.bioontology.org

Pays : Etats-Unis

Langue : anglais

Responsabilité éditoriale : National Center for Biomedical Ontology (NCBO), one of the National Centers for Biomedical Computing supported by the NHGRI, the NHLBI, and the NIH

Le portail de cette archive ouverte d'ontologies biomédicales existe depuis 2005 et compte 379 ontologies (août 2014). BioPortal offre des fonctionnalités de recherche, de navigation et de visualisation des ontologies signalées et déposées par leurs auteurs. Outre la description de leurs caractéristiques, il présente aussi le degré de similitude entre les différentes ontologies sous forme de tableau ou d'une visualisation dynamique. Il met à disposition un moteur de recommandation et un annotateur pour aider à la sélection d'ontologies à partir de ressources textuelles. L'interface collaborative permet de commenter les ontologies et de signaler des erreurs. L'interrogation des ontologies est possible soit depuis l'interface Web soit via des Web services. Le portail est parfois instable ou indisponible.

Open Biological and Biomedical Ontologies (OBO) http://www.obofoundry.org

Pays : Etats-Unis

Langue : anglais

Responsabilité éditoriale : « The coordinating editors of the OBO Foundry are Michael Ashburner (GO), Chris Mungall (GO, BIRN, modENCODE, BBOP), Suzanna Lewis (GO, BIRN, modENCODE, BBOP), Alan Ruttenberg (ORG, Immport), Richard H. Scheuermann (IRD, ViPR, Immport, CTSA), Barry Smith (NCBO, ORG, Immport), and Melissa Haendel (OHSU). Susanna Assunta-Sansone (University of Oxford e-Research Centre, NPG Scientific Data) is industry liason. »

Une liste de dix ontologies respectant les principes de cette collaboration pour le développement d'une *core* ontologie dans le domaine biomédical est proposée. A celle-ci s'ajoute la liste de 122 ontologies (août 2014) d'autres collectifs et projets qui suivent les principes développés par OBO Foundry.

Ontobee http://www.ontobee.org/index.php

Pays : Etats-Unis

Langue : anglais

Responsabilité éditoriale : He Group, University of Michigan Medical School

Ce moteur permet de rechercher par concept au sein des 150 ontologies biomédicales référencées (août 2014), de les visualiser au sein de leur hiérarchie et de connaître si ces concepts sont reliés à ceux d'autres ontologies. Y est associé le programme Ontobeep qui fournit les statistiques d'alignement et de comparaison entre les ontologies indexée par Ontobee.

Ontology Lookup Service (OLS) http://www.ebi.ac.uk/ontology-lookup

Pays : Organisation intergouvernementale située au Royaume-Uni

Langue : anglais

Responsabilité éditoriale : European Bioinformatics Institute is part of EMBL (European Molecular Biology Laboratory)

Ce moteur, créé en 2006, permet d'interroger par concept 91 ontologies (août 2014) du domaine biomédical via une interface HTML, soit via des web services. Toutes les informations concernant ces concepts sont automatiquement mises à jour. Des sondages sont réalisés avec une fréquence quotidienne afin de détecter les nouvelles versions chez les fournisseurs d'ontologies, de les télécharger et de les analyser afin de mettre à jour OLS.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Les outils de recherche de sources ontologiques

Les portails, entrepôts et moteurs généralistes de recherche d'ontologies

DAML Ontology Library http://www.daml.org/ontologies/ontologies.html
Pays : Etats-Unis
Langue : anglais
Responsabilité éditoriale : Defense Advanced Research Projects Agency (DARPA)
Annuaire des ontologies conçu à l'occasion du programme de recherche sur le web sémantique DARPA Agent Markup Language (DAML) mené de 1999 à 2006. L'annuaire maintenu jusqu'en 2004 est toujours consultable. Il référence 282 ontologies.

Linked Open Vocabularies (LOV) http://lov.okfn.org/dataset/lov/index.html
Pays : France
Langue : anglais
Responsabilité éditoriale : Développé en 2011 par l'Inserm et Mondeca dans le cadre du programme DataLift soutenu par l'ANR[17] et hébergé depuis 2012 par la Fondation Open Knowledge (OKFN).
Il s'agit d'un catalogue de 458 ontologies au format RDFS ou OWL (août 2014) sélectionnées pour le programme DataLift, et nommées vocabulaires par le projet. Ils sont interrogeables par son moteur via leur nom mais aussi par les éléments présents dans les classes et les propriétés. Ces vocabulaires peuvent aussi être retrouvés via une classification visuelle par grands domaines. Pour chacun, LOV présente les métadonnées officiellement déclarées par leurs éditeurs ou ajoutées par les curateurs du catalogue. Leur description est datée et commentée par les curateurs du LOV. Une représentation graphique permet d'appréhender le degré d'interconnexion des vocabulaires entre eux ainsi que de visualiser l'historique de leurs versions. LOV a pour vocation d'être un outil de gouvernance pour encourager la qualité des vocabulaires sémantiques disponibles sur le Web. Ceci afin de faciliter leur réutilisation et lier les données. La réflexion sur les critères d'évaluation est donc en cours. Leur élaboration est menée en collaboration avec le projet OOPS.

Ontohub (BETA) https://ontohub.org
Pays : Allemagne
Langue : anglais
Responsabilité éditoriale : Universität Bremen
Moteur de recherche interrogeant 2 542 ontologies et 42 entrepôts (août 2014). La recherche peut être également faite par sélection dans une liste de domaines. Ce projet prend en charge l'organisation, la publication, la récupération, le développement, la cartographie, la traduction. Il évalue, avec l'outil OOPS! - OntOlogy Pitfall Scanner!, un large éventail d'ontologies formelles dans différentes langues et différents langages (OWL, CASL, Propositional, SoftFOL, CommonLogic, DOL). Il s'appuie sur le projet OntoIOp (Ontology, Model and Specification Integration and Interoperability - http://ontoiop.org).

[17] Datalift est une plateforme dédiée à l'exploitation des données liées (conversion, publication et interconnexion des jeux de données). En sortie, Datalift produit des données liées : des "linked data", celles qu'on dit aussi sémantiques et interconnectées. http://datalift.org/

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Les portails, entrepôts et moteurs généralistes de recherche d'ontologies (suite)

Ontology Design Patterns (ODP) http://ontologydesignpatterns.org
Pays : Communauté internationale
Langue : anglais
Responsabilité éditoriale :Un collectif composé de trente universitaires
Ce wiki est consacré au design d'ontologies reposant sur des patrons de conception. Depuis celui-ci, il est possible de consulter la liste d'ontologies référencées par le collectif qui les valide selon neuf critères. Des tris peuvent être réalisés sur cette liste selon le domaine d'appartenance des ontologies par exemple. Le nombre de téléchargement des ontologies est également indiqué. De même, les ontologies ou éléments d'ontologies soumis à examen ainsi que l'appréciation par le collectif sont aussi consultables.

Protege Ontology Library http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library
Pays : Communauté internationale
Langue : anglais
Responsabilité éditoriale : Les utilisateurs de Protégé participent à alimenter les pages du Wiki
Protégé est un éditeur Open Source pour la création d'ontologies. Il est développé par le Stanford Center for Biomedical Informatics Research de la Stanford University School of Medicine. La communauté des utilisateurs de Protégé alimente la liste des ontologies créées avec cet outil sur cette page du Wiki de Protégé. Y sont référencées des ontologies aux formats OWL, DAML+OIL, RDFS. Une section de cette page est aussi consacrée aux ontologies crées avec la version Protege-Frames. Les mises à jour de cette page sont manifestes d'après l'historique de la page.

Schemapedia http://schemapedia.com
Pays : Royaume-Uni
Langue : anglais
Responsabilité éditoriale : Ian Davis
Répertoire collaboratif d'ontologies. En cours de migration (août 2014).

Swoogle http://swoogle.umbc.edu/
Pays : Etats-Unis
Langue : anglais
Responsabilité éditoriale : UMBC Ebiquity Groupe, University of Maryland.
Swoogle est le premier moteur de recherche sémantique. Il a été mis en service en 2004 et maintenu jusqu'en 2007. Il est encore interrogeable aujourd'hui pour retrouver des URI de documents sémantiques ou d'ontologies du Web sémantique. Les statistiques des pages de l'aide portent à croire que celui-ci est encore consulté. 1,3 million de documents du Web sémantique contenant près de 240 millions de triplets sont indexés par Swoogle.

Watson http://watson.kmi.open.ac.uk/
Pays : Royaume-Uni
Langue : anglais
Responsabilité éditoriale :Le Knowledge Media Institute (KMi) de l'Open university
Watson est un moteur de recherche sémantique qui indexe des documents sémantiques et des ontologies. Il fournit une liste d'URI et quelques métadonnées les concernant : poids du fichier auxquelles ces URI appartiennent, format. Pour chaque URI, il est indiqué si elle correspond à la description d'une classe, d'une relation ou d'une instance. Des options de recherche permettent de limiter la recherche au terme exact recherché ou de combiner les filtres sur les labels, commentaires, certaines entités…

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Les outils de recherche de terminologies non-ontologiques

Listes et moteurs de recherche de thésaurus

BAsel Register of Thesauri, Ontologies & Classifications (BARTOC) http://www.bartoc.org
Editeur : Basel University Library
Pays : Suisse
Langue : multilingue (65 langues)
Ce registre répertorie 741 ressources dont 408 thésaurus et 71 ontologies mais aucune dans le domaine de la toxicologie. Il existe une fonction recherche simple à laquelle peuvent être associés des filtres, de même qu'il existe des facettes pour filtrer les résultats de la recherche. Les ressources sont également présentées sur une carte géographique. Chaque ressource est décrite avec un résumé et une grille d'identification composée entre autres de l'indice Dewey, de l'identifiant VIAF[18], des conditions d'accès, des formats disponibles, des langues qui la compose et d'un lien vers la description dans l'encyclopédie collaborative Wikipédia.

La liste de thésaurus de l'annuaire collaboratif de l'Open Directory Project (ODP) DMOZ
http://www.dmoz.org/World/Français/Références/Thésaurus
Editeur : curateurs bénévoles
Pays : communauté internationale, probablement de langue française pour cette section
Langue : français
« Cette catégorie contient [78] thésaurus documentaires (ou thésaurus de descripteurs [ou des liens vers des listes de thésaurus]) [...]. Ces thésaurus sont nécessairement en version française, consultables en ligne, téléchargeables ou a minima pour lesquelles il existe sur le Web une présentation, même succincte. »

[18] Fichier d'autorité international virtuel : «projet commun de plusieurs bibliothèques nationales, mis en œuvre et hébergé par OCLC. L'objectif du projet est de faire baisser le coût et de valoriser les fichiers d'autorité des bibliothèques par l'appariement et l'établissement de liens entre les fichiers d'autorité des bibliothèques nationales, et en rendant cette information disponible sur le web » http://viaf.org

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Quelques ressources pour affiner les concepts de l'ontologie ToxNuc[19]

Les bases de données spécialisées

Enzyme database - BRENDA http://www.brenda-enzymes.info/
Editeur : Institut de biochimie et de bioinformatique de l'Université technique de Braunschweig,
Pays : Allemagne
Langue : anglais
Cette base de données gratuite créée en 1987 par le Centre national allemand de recherche pour la biotechnologie à Braunschweig (GBF). Les données sont extraites de la littérature et d'une évaluation critique par des scientifiques qualifiés.

Protein du NCBI http://www.uniprot.org/
Editeur : National Center for Biotechnology Information, U.S. National Library of Medicine (NLM)
Pays : Etats-Unis
Langue : anglais
La base de données de protéines du NCIBI est une collection enrichie de plusieurs sources dont GenBank, RefSeq et TPA, ainsi que par les données de SwissProt, PIR, PRF, et l'APB.

Les glossaires du domaine de la toxicologie

Les glossaires du collectif **ToxNuc** :
sur la plateforme Toxnuc-e : http://www.toxnuc-e.org/vulgarisation/site/glossaire.php
sur la plateforme Toxce : http://toxcea.fr/index.php?pagendx=63
Il est accessible au moyen d'un identifiant et d'un mot de passe. Il est incomplet[20].

Les définitions des termes de l'ontologie de la toxicologie du la communauté **OpenTox** :
sur le wiki de la communauté : http://www.opentoxipedia.org/index.php/Main_Page
Les sources des définitions sont citées, certaines renvoient aux définitions du site de l'US Environmental Protection Agency (EPA) qui dispose d'un service dédié à la terminologie (voir aussi p. 19).

Les glossaires du domaine de l'environnement

Les définitions du service de terminologie de l'US Environmental Protection Agency (EPA) sont interrogeables via un moteur et un Abécédaire cliquable (voir aussi p. 19).
http://ofmpub.epa.gov/sor_internet/registry/termreg/searchandretrieve/termsandacronyms/search.do

Les glossaires du domaine des nanosciences

L'International Institute for Nanotechnology, fondé en 2000 et basé aux Etats-Unis, propose un glossaire sur son nouveau site à l'adresse suivante : http://www.iinano.org/full-glossary[21]

[19] Voir aussi la liste d'outils sur le site du projet eTOX http://cadd.imim.es/etox-library/library/links/tools-links
[20] Problème signalé à Marie-Thérèse Ménager par courriel le 11 juin 2014.
[21] Ancienne adresse Internet de l'IIN : http://discovernano.org

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Quelques ressources pour affiner les concepts de l'ontologie ToxNuc[22] (suite)

Les portails terminologiques

HeTOP (Health Terminology / Ontology Portal) http://www.hetop.eu/hetop

Pays : France

Langue : Français, interrogation multilingue

Responsabilité éditoriale : Ce Portail Terminologique de Santé est proposé par l'équipe CISMeF du CHU de Rouen (LITIS EA 4108, Université de Rouen), en collaboration avec l'INSA de Rouen et la société MONDECA.

Le Portail Terminologique de Santé (PTS) inclut les principales terminologies de santé disponibles en français. Il a pour objectif de permettre un accès centralisé aux terminologies suivantes :
ACR, ADICAP, ATC, BHN, BNCI, BNPC, Bon Usage Radio., CCAM, CIF, CIM-9, CIM-10, CIM-O, CISMeF, CISP-2, Cladimed, DRC, FMA, Gene Ontology, Gènes & Protéines, HPO, ICNP, IUPAC, LOINC, LPP, MedlinePlus, MeSH, MIDAS, NABM, NCCMERP, NCIT, OMIM, HRDO, PATHOS, PSIP Taxonomy, RADLEX, SNOMED int., UMLS (réseau sémantique), VCM, WHO-ART et WHO-ICPS.

Termsciences, portail terminologique multidisciplinaire http://www.termsciences.fr

Pays : France

Langue : Interface en français et en anglais, interrogation multilingue (français, anglais, allemand, espagnol)

Responsabilité éditoriale : Ce portail terminologique a été développé par l'Institut de l'information scientifique et technique (INIST) en association avec le Laboratoire lorrain de recherche en informatique et ses applications (LORIA) et le Laboratoire analyse et traitement informatique de la langue française (ATILF).

Ce portail terminologique permet d'interroger en un point d'accès unique les ressources terminologiques (lexiques, dictionnaires, thesaurus) des organismes publics de recherche et d'enseignement supérieur. Il propose différents modes d'interrogation, notamment par web services. Une interface visuelle de recherche permet également d'interroger la liste des concepts qui sont la racine d'au moins une arborescence des terminologies mises en correspondance pour ce service. L'interrogation de plusieurs bases bibliographiques, annuaires et moteurs est possible depuis l'interface de Termsciences. En outre la « recherche simple » est également possible avec le plug-in TermSciences Quick Search For Firefox. Cependant, il n'est pas fait mention de la dernière date de mise à jour des versions des terminologies utilisées par ce service. Ce service est parfois instable ou indisponible.

[22] Voir aussi la liste d'outils sur le site du projet eTOX http://cadd.imim.es/etox-library/library/links/tools-links

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Ressources pour la recherche d'information et la veille en toxicologie

	Toxicologie	**Toxicologie nucléaire et médecine nucléaire**
Bases de données factuelles ou bibliographiques institutionnelles librement consultables	Agritox http://www.agritox.anses.fr/ ToxNet http://toxnet.nlm.nih.gov/index.html DSSTox http://www.epa.gov/ncct/dsstox/ ACToR http://actor.epa.gov/actor/faces/Download.jsp RTECS http://www.cdc.gov/niosh/rtecs/default.html CTD http://ctd.mdibl.org GENETOX http://www.nlm.nih.gov/pubs/factsheets/genetxfs.html DART http://www.nlm.nih.gov/pubs/factsheets/dartfs.html KEGG http://www.kegg.jp/kegg/drug/ DITOP http://bioinf.xmu.edu.cn/databases/ADR/search.php T3DB http://www.t3db.org	INIS http://www.iaea.org/inis/
Sites de projets et d'outils librement consultables	CAESAR http://www.caesar-project.eu/index.php?page=index CompTox http://www.epa.gov/ncct/ ECOSAR http://www.epa.gov/oppt/newchems/tools/21ecosar.htm eTOX http://www.etoxproject.eu/ OASIS http://toolbox.oasis-lmc.org/?section=download OECD toolbox http://www.qsartoolbox.org/ OpenTox http://www.opentox.org/toxicity-prediction RepDose http://www.fraunhofer-repdose.de/ SEURAT http://www.seurat-1.eu/, http://www.seurat-1.eu//pages/the-cluster-projects/cosmos.php ToxCast http://www.epa.gov/ncct/toxcast/ TOXMAP http://toxmap.nlm.nih.gov/toxmap/ Tox21 http://www.epa.gov/ncct/Tox21/	National Toxicology Program http://ntp-server.niehs.nih.gov/ Accel-RT http://www.accel-rt.org/ Computational Radiotherapy http://www.comprt.org/ GHOST http://www.ghost-project.org/ VoxTox http://www.voxtox.org/
Systèmes industriels	DEREK https://www.lhasalimited.org/derek_nexus/ HazardExpert http://www.compudrug.com/?q=node/35 Multicase http://www.multicase.com QSAR Workbench http://accelrys.com/services/qsar-workbench.html TOPKAT http://accelrys.com/products/discovery-studio/predictive-toxicology.html	

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Ressources pour la recherche d'information et la veille dans les disciplines biomédicales et informatiques

ACM Digital Library http://dl.acm.org
Banque de données Santé Publique (BDSP)[23] http://www.bdsp.ehesp.fr/Base
Biomed Central http://www.biomedcentral.com/
CiteSeer[x] http://citeseerx.ist.psu.edu/index
Embase.com (sur abonnement)[24] http://www.biusante.parisdescartes.fr
IEEE Xplore http://ieeexplore.ieee.org/Xplore/home.jsp
PubMed (dont MEDLINE et liens vers PubMed Central) http://www.ncbi.nlm.nih.gov/pubmed

Une alerte **PubMed** a été paramétrée le 10 juin 2014.
Voici les identifiants afin de changer l'adresse mail pour la réception des alertes :
NCBI username : aclp
Password : Nhy456tgb,

A noter : le mot-clé « Biological ontologies » a été introduit dans le MeSH seulement en 2014. Pour trouver des documents sur les ontologies biomédicales indexées avant 2014, des recherches à l'aide d'autres mots-clés du MeSH sont nécessaires, comme par exemple « Vocabulary, Controlled ».

Les équations de recherche mises en alerte sont les suivantes :
'("Vocabulary, Controlled"[Mesh]) AND ("Toxicology"[Mesh] OR "Ecotoxicology"[Mesh])'
'(tox[sb] AND ontolog*[title])'
Quatre alertes ont été reçues durant le stage. Parmi les références réceptionnées, une paraît intéressante pour réfléchir aux relations qui permettraient d'enrichir l'ontologie au niveau formel :

> YOUNESI E., ANSARI S., GUENDEL M., AHMADI S., COGGINS C., HOENG J., HOFMANN-APITIUS M., PEITSCH M. C. « CSEO - the Cigarette Smoke Exposure Ontology. » *J Biomed Semantics* [En ligne]. 2014. Vol. 5. Disponible sur : < http://dx.doi.org/10.1186/2041-1480-5-31 >
> PMID: 25093069 (voir aussi : https://publicwiki-01.fraunhofer.de/CSEO-Wiki/index.php ; http://purl.bioontology.org/ontology/CSEO)

La revue **Journal of Biomedical Semantics** semble aussi une source à surveiller sans attendre l'indexation de ses articles dans PubMed. En effet, le référencement et l'indexation sont effectifs avec parfois un certain délai après la parution d'articles intéressants pour le projet d'ontologie.

La revue **Applied Ontology** est également une revue référencée par PubMed et dont il est possible de consulter les sommaires sur le site de l'éditeur à l'adresse suivante :
http://www.iospress.nl/journal/applied-ontology

[23] Le thésaurus de la BDSP ne s'est pas révélé d'une grande aide lors de l'étude de la terminologie ToxNuc.
[24] Surveiller l'accès sur le site de la Biu Santé (http://www.biusante.parisdescartes.fr/) pour consultation sur place en cas de réabonnement

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Annexe : l'historique du projet d'ontologie de la toxicologie nucléaire

Historique du projet d'ontologie de la toxicologie nucléaire / Véronique Gachet. – Novembre 2011. – 7 p

Historique du projet d'ontologie de la toxicologie nucléaire

Véronique Gachet

Dans cet historique, nous voulons reprendre la genèse d'un projet d'ontologie de la toxicologie nucléaire initié en 2005 par le Programme Transversal Toxicologie du CEA et le LGI²P de l'Ecole des Mines d'Alès.

L'objectif de ce document est de permettre à de nouveaux contributeurs de reprendre la totalité du dossier dans les meilleures conditions afin d'amener ce projet à son terme.

Véronique Gachet – 7 novembre 2011

I. Introduction

Le Programme Transversal Toxicologie[1] du CEA (Commissariat à l'énergie atomique et aux énergies alternatives) représenté par Marie-Thérèse Ménager et le LGI²P (Laboratoire de Génie Informatique et Ingénierie de production) de l'Ecole des Mines d'Alès travaillent ensemble à l'élaboration d'une ontologie du domaine de la toxicologie nucléaire.
Le projet initié en 2005 avait pour objectif, parallèlement aux travaux de l'atelier de travail collaboratif, la construction d'une **Ontologie de la Toxicologie Nucléaire,** dans le cadre du projet de recherche Intelligence collective et travail collaboratif CYCLONE (LGI²P de l'Ecole des Mines d'Alès). Il s'agissait de décrire par des concepts et mots-clefs les quinze projets du programme ToxNuc-E, selon plusieurs types de propriétés (modèles biologiques, toxiques chimiques, techniques…), dans le but de faciliter la recherche par mots-clefs au sein de la plate-forme de travail collaboratif, et de constituer un réseau sémantique propre à la toxicologie nucléaire. Les termes ont été sélectionnés à partir d'un corpus en français pour ce qui concerne les projets et les exposés et en anglais pour ce qui concerne les publications.

Ce document s'appuie sur les quatre étapes de la méthodologie de construction d'une ontologie telles que Jean Charlet le propose dans son article « L'ingénierie des connaissances, entre science de l'information et science de gestion »[2] : la primauté du corpus et son analyse, la normalisation sémantique, l'engagement ontologique et l'opérationnalisation.

Les participants au projet :

Anoir Imane, doctorante au LGI²P
Gachet Véronique, stagiaire M2 Communication et technologie numérique – CEA/EMA – 2011 - veronique.gachet@cea.fr -01 69 08 00 70
Montméjean Emilie, stagiaire post Master 2 au LGI²P 2008
emilie.montmejean@wanadoo.fr – 06 80 84 45 15 – 05 59 40 79 62

[1] Programme ToxNuc (2001-2003) puis ToxNuc-E (2004-2007) puis le Programme Transversal Toxicologie en 2009
[2] Version allongée d'un chapitre de livre coordonnée par R. Teulier et Ph. Lorino [2005] et faisant suite au colloque de Cerisy « Activité, connaissance, organisation ».

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Shetty Reena, doctorante au LGI²P – Calcul de proximité / sémantique
Villerd Jean, doctorant au LGI²P - cartographie

Charlet Jean, expert en ontologie, INSERM / AP-HP
Vandenbussche Pierre-Yves, thésard, spécialiste en ontologie, INSERM / AP-HP
Ansoborlo Eric, CEA Marcoule – Direction du programme Toxicologie
Martin Laffon Jacqueline, documentaliste au CEA Grenoble
Ménager Marie-Thérèse, CEA, Direction de programme Toxicologie

---

Dray Gérard, LGI²P
Ranwez Sylvie, expert en ontologie, LGI²P
Riccio Pierre-Michel, LGI²P

II. Etape 1 : le corpus et son analyse

L'ontologie a été mise au point à partir des quinze projets du programme ToxNuc-E :

Projets	Coordonnateurs	Correspondants référentiel
Méthodologie et spéciation en milieu biologique et environnemental	Christophe Moulin (CEA) Ryszard Lobinsky (CNRS)	Denis Doizi
Chélation Biologique	Michel Ferrand (CEA)	Michel Ferrand
Cibles moléculaires des actinides	Eric Quémeneur (CEA)	Frédéric Berenguer
Décorporation des actinides	F. Taran CEA (CEA) B. Le Gall (CEA) J.R. Deverre (CEA)	
Iode	Thierry Pourcher (CEA) Philippe Potin (CNRS)	Yves Ambroise
Résistance des bactéries aux métaux lourds et radionucléides	David Pignol (CEA) Marie-Andrée Mandrand(CNRS)	David Pignol

Régulation de la réponse aux toxiques chez la levure	Jean Labarre (CEA)	Stéphane Chedin
Régulation de la réponse aux ML et radionucléides chez Arabidopsis thaliana	Jacques Bourguignon (CEA)	Jacqueline Martin-Laffon
Altération-réparation de l'ADN	Bernard Lopez (CNRS) Nicolas Foray (Inserm)	Yannick Saintigny
Toxico-génomique de la réponse et métaux lourds	Franck Chauvat (CEA) Nathalie Leonhardt	Philippe Ortet
Stress Oxydant (SIDDERE)	Michel Toledano (CEA) Alain Vavasseur Jean Luc Montillet	Alain Vavasseur
Transporteurs membranaires	Cyrille Forestier (CEA)	Cyrille Forestier
Néphrotoxicologie et Toxico-cancérogénèse	Sylvie Chevillard (CEA) Gabriel Baverel (Inserm)	Marie Carrière
Génotoxicité	Pierre Fouchet (CEA)	Yannick Saintigny (correspondant du projet Altération réparation)
Transfert Sol-Plantes	Michel Lebrun (Professeur) Laurence Denaix (Inra)	Brice Bouyssiere

En 2005 et 2006, sous l'égide de **Pierre-Michel Riccio**, **Imane Anoir,** doctorante au LGIPP a recensé, à partir des quinze projets du programme scientifique, le vocabulaire scientifique pertinent relatif au programme Toxicologie Nucléaire pour pouvoir construire l'ontologie du domaine. Cette première analyse a été complétée par le document « ToxNuc-E, Bilan scientifique 2004-2007 ».
Ce travail a permis de recenser plus de 900 termes représentant le domaine de la toxicologie nucléaire environnementale : 60 disciplines scientifiques, 91 toxiques potentiels d'intérêt, 196 molécules, 290 outils et plus de 200 revues scientifiques dans lesquelles les chercheurs publient.

Jacqueline Martin Laffon a contribué avec **Imane Anoir** à la mise en place de l'ontologie pour le projet « Arabidopsis » (description des concepts avec des termes (mots-clefs) génériques (exemple : métal lourd) et spécifiques (exemple : cadmium) + construction d'une liste de revues représentative du domaine et dans lesquelles les chercheurs du projet publient).

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

La description du projet « Arabidopsis » a servi de base à la description des autres projets, pour lesquels les coordonateurs ont apporté leurs éléments spécifiques.
Des fiches par projet ont été élaborées puis validées par les coordonnateurs de chaque projet. A partir de ces fiches, un document de travail de 21 pages (CD ROM joint) sous forme de tableau général a été construit et remis aux coordonnateurs pour validation de leur section le 25/10/2006.
Les fiches « Arabidobsis » et « MSBE » ont été traduite en anglais par **Imane Anoir** et **Jacqueline Laffon.**

Eric Ansoborlo a également contribué à certaines traductions.

Marie-Thérèse Ménager et Pierre Michel Riccio sont intervenus dans l'élaboration de la méthodologie et sa validation.

III. Etape 2 : la normalisation sémantique

Pour chaque terme et/ou mots-clefs, les coordonnateurs de projet[3] ont attribué un poids en fonction de l'importance accordée dans le cadre du projet, ce qui a permis de construire des cartographies dynamiques permettant de représenter les relations entre les projets du Programme ToxNuc-E en fonction de chaque concept. Cette première représentation dynamique a permis de visualiser et donc de mettre en évidence les collaborations scientifiques qui ont émergé pendant le programme. (**Imane Anoir, Reena Sheety, Jean Villerd**).

IV. Etape 3 : l'engagement ontologique

Sous l'égide de **Gérard Dray, Sylvie Ranwez** et **Marie-Thérèse Ménager**, **Véronique Gachet** a repris, en 2011, le travail initié par **Imane Anoir** et **Emilie Montméjean**.
Sylvie Ranwez a contribué à l'aboutissement de l'ontologie en terme de formation, conseil, apport, modifications et corrections.

[3] Ou l'équipe du LGIPP : à vérifier

Auteur : Anne-Claire Le Picard	Date de création : 06-2014	Page **32** sur **34**
Nom du fichier : 2014_stage_intd_livrable3_repertoire_ressources_termino-ontologiques_ontotoxnuc.pdf	Date de modification : 08-2014	N° de version : 01

Les actions de Véronique Gachet :

- Reprendre l'ensemble du dossier et se l'approprier,
  - Corriger les incohérences
  - Supprimer les doublons
  - Supprimer les pluriels

- vérifier le corpus dans le tableau Excel global, la modifier et corriger si besoin (CD ROM joint)

- transférer le tableau structuré en termes français dans le logiciel Protégé :
  - avec le concours de **Jean Charlet** et **Pierre-Yves Vandenbussche**

Protégé *est un système auteur pour la création d'ontologies. Il a été créé à l'université Stanford et est très populaire dans le domaine du Web Sémantique et au niveau de la recherche en informatique. Protégé est développé en Java. Il est gratuit et à code source libre. Protégé peut lire et sauvegarder des ontologies dans la plupart des formats d'ontologies : RDF, RDFS, OWL, etc.*

(http://fr.wikipedia.org/wiki/Prot%C3%A9g%C3%A9_%28logiciel%29 – (consulté le 12 octobre2011)

- reprendre l'ontologie dans Protégé,
  - vérifier l'orthographe, la saisie, la pertinence

- traduire les termes manquants et les saisir dans Protégé
  - la plupart des publications scientifiques étant en anglais, l'ontologie doit être bilingue

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

V. Etape 4 : l'opérationnalisation

Phase à venir :

L'ontologie est exploitable dans Protégé. Elle nécessite quelques corrections sur les traductions. Certains termes doivent être vérifiés.

VI. Bibliographie

Imane ANOIR, (2009), Emergence de communautés scientifiques fondée sur le travail collaboratif, thèse en science de l'information et de la communication, Université Paul Cezanne, Aix-en-Provence.

Jean CHARLET, (2005), Version allongée d'un chapitre de livre coordonnée par R. Teulier et Ph. Lorino [2005] et faisant suite au colloque de Cerisy « Activité, connaissance, organisation ».

Reena T.N. SHETTY, (2008), Enrichissement de réseaux sémantiques par la proximité de concepts, thèse en informatique, École des Mines de Paris.

Jean VILLERD, (2008), Représentations visuelles adaptatives de connaissances associant projection multidimensionnelle (MDS) et analyse de concepts formels (FCA), thèse en informatique, École des Mines de Paris.

Auteur : Anne-Claire Le Picard	Date de création : 06-2014	Page **34** sur **34**
Nom du fichier : 2014_stage_intd_livrable3_repertoire_ressources_termino-ontologiques_ontotoxnuc.pdf	Date de modification : 08-2014	N° de version : 01

Annexe 2
Livrable 4 - Etude d'opportunité pour le projet de reprise de l'ontologie ToxNuc à partir des informations recueillies

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	04

Commanditaires	Elément de la commande n°1
• Direction des sciences du vivant – DSV, Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) • Inserm UMR_S 1142 • Laboratoire de Génie Informatique et d'Ingénierie de Production – LGI2P, Centre de recherche de l'Ecole des Mines d'Alès – EMA (Nîmes)	**Etude d'opportunité pour le projet de reprise de l'ontologie ToxNuc à partir des informations recueillies**

Rappel de la commande

Objectif général	**Une étude d'opportunité est une évaluation de la pertinence d'un projet. Elle vise à valider ou non son intérêt et à éclairer la définition des objectifs généraux.**
Contexte de la demande	
Environnement interne	Plateforme collaborative créée par le LGI2P et maintenue par la société ID-Alizée pour le collectif ToxNuc coordonné par le Commissariat à l'Energie Atomique et aux énergies alternatives (CEA)
Environnement externe	Expérimentations de nouvelles méthodes pour accompagner l'opérateur humain dans son investigation de vastes corpus de connaissances
Résultats et documents attendus	
Etat de l'art	Répertoire des ressources termino-ontologiques disponibles dans le domaine
Synthèse devant permettre	- d'identifier les manques éventuels comblés par l'ontologie du collectif ToxNuc - d'identifier quelles ressources doivent être intégrées pour compléter l'aspect nanotoxicité - d'évaluer l'impact d'une telle ressource au niveau international - d'étudier l'usage des ontologies pour la recherche d'information
Calendrier	Du 2 juin 2014 au 22 août 2014
Personnes mobilisables	J. Charlet (AP-HP & INSERM UMR_S 1142) Marie Thérèse Ménager (CEA) G. Dray, S. Ranwez (LGI2P – EMA Nîmes)
Projet de rattachement	
Equipe de recherche	Knowledge and Image analysis for Decision - KID
Laboratoire	Laboratoire de Génie Informatique et d'Ingénierie de Production – LGI2P, Centre de recherche de l'Ecole des Mines d'Alès – EMA (Nîmes)
Projet	Automatisation des différentes phases d'analyse, d'indexation et d'exploitation des informations
Entre autres objectifs	Etude des apports des standards du Web sémantique et de la modélisation des connaissances à l'aide d'ontologies de domaine à ce type de projet
Mots-clés	Terminologie. Ontologies. Indexation. Analyse de corpus. Acquisition des connaissances à partir de textes

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	04

Plan

L'origine du projet de reprise de l'ontologie de la toxicologie nucléaire

Y avait-il et y a-t-il encore un problème à résoudre ? Quelle est son urgence ? En quoi la conception d'une ontologie est-elle une solution à ce problème ? Il semble que ce projet ne réponde à aucun problème. Aucune urgence ne presse à concevoir une terminologie hiérarchisée, voire une ontologie avec une expressivité plus importante. Cependant une volonté d'action est manifestée par les acteurs contributeurs à cette première ébauche : direction des programmes de la Direction des sciences du vivant du CEA, spécialistes du TAL et des ontologies du Laboratoire de Génie Informatique et d'Ingénierie de Production (LGI2P) à Nîmes et de l'Inserm à Paris.
Les évènements qui motivent la réflexion sur les suites à donner à l'ontologie initiée en 2005 et opérationnalisée en OWL en 2011 :

- la thèse en cours portant sur l'utilisation d'ontologies comme support à la recherche d'informations complémentaires (Nicolas Fiorini au LGI2P)
- le travail d'une étudiante du Cnam-INTD : affinage de l'ontologie actuelle et recherche de ressources termino-ontologiques concurrentes et/ou complémentaires de juin à septembre 2014 (Anne-Claire Le Picard à l'Inserm)

La finalité et l'adéquation avec l'idée d'origine

L'idée d'ontologie a pour origine la constitution d'une équipe de chercheurs au sein d'un programme de recherche sur la toxicologie nucléaire en 2001. Elle accompagne celle qui s'est concrétisée par la création d'une plateforme collaborative de communication, de diffusion et de stockage : http://www.toxnuc.org puis http://www.toxnuc-e.org, devenue enfin http://www.toxcea.org. Au travers des différents documents stockés sur cette plateforme, il est fait état des connaissances que le programme transversal ToxNuc mené par le CEA a continué à étendre.

Dès son origine, la finalité d'une ontologie de la toxicologie nucléaire ainsi que son périmètre semblent flous. Et c'est là en revanche un problème à résoudre pour être en mesure de décider ou non de reprendre sa conception. Quelle est la vision ? A cette première question, suivent de nombreuses questions dont les enjeux comportent des dimensions stratégiques, techniques et organisationnelles fortes. Qui est porteur de cette vision ? Qui est décisionnaire ? Quelle est la valeur potentielle d'une ontologie ? Et de quel type d'ontologie parle-t-on ? Quels services peut-elle rendre ? Quelle priorisation des services attendus ? Dans quel écosystème technique et organisationnel ? Quels sont les types d'usagers réels ou potentiels des dispositifs techniques dans lesquels elle pourrait être implémentée ? Avec quel rapport coûts/bénéfices ?

Concernant la finalité, on s'accorde sur une vision. Sa conception est un moyen d'établir un consensus sur l'état des connaissances dans un domaine où la France occupe la meilleure place. Elle rejoint justement la définition consensuelle de ce qu'est une ontologie. Il s'agit d'une spécification formelle et explicite d'une conceptualisation partagée d'un domaine de connaissance[1]. Cela signifie fixer, avec les termes du domaine, les définitions des concepts (entités, attributs, processus) ainsi que leurs interrelations, convenant aux différents acteurs de la toxicologie nucléaire : géochimistes, biologistes, médecins... Et ceci de façon formelle, afin d'être compréhensible par la machine, pour être exécutée lors d'opérations informatiques. Mais une ontologie n'est pas produite pour elle-même, elle est toujours un moyen et non une fin. Sa formalisation est fonction d'usages plus ou moins élaborés, des plus génériques au plus spécifiques.

[1] "An ontology is a formal, explicit specification of a shared conceptualization." [STUDER R. *et al.* 1998, p. 184] STUDER R., BENJAMINS V. R., FENSEL D. « Knowledge engineering: Principles and methods ». Data & Knowledge Engineering. 1998. Vol. 25, p. 161-197.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	04

Le périmètre ou plutôt les usages qui peuvent être faits d'une ontologie de la toxicologie nucléaire restent aussi pertinents aujourd'hui qu'hier. Ces grandes fonctionnalités sont autant de scénarios possibles. Non exclusifs les uns des autres, ils peuvent être menés de façon progressives, comme les étapes d'un seul et même projet. Quel que soit le choix, il doit être envisagé un ou des scénarios de continuité pour chacun à savoir mettre en place un mode de gouvernance. Ceci afin de pérenniser et actualiser la valeur produite. Les usages et évolutions de l'ontologie actuelle sont fonction de besoins qui restent encore à définir. Il peut s'agir des besoins des utilisateurs directs et indirects de l'ontologie comme les toxicologues, comme des besoins de ceux chargés de son maintien. Et ce sont aussi les attentes des utilisateurs de la plateforme de Toxcea dont il convient de tenir compte. Une ontologie est-elle la solution la plus efficace et la plus économique pour y répondre ?

Avec une expressivité poussée permettant des raisonnements, une ontologie peut participer à alimenter une base de connaissance. C'est-à-dire que les règles informatiques, en d'autres termes l'ontologie reliant les connaissances expertes d'un domaine, permettent l'exploitation de celles-ci par les machines. Ainsi des raisonnements logiques, informatiques peuvent produire de nouvelles connaissances, pour faire de la veille et peupler l'ontologie. Des éléments de pilotage pourraient être ajoutés à la terminologie actuelle afin de se rapprocher des objectifs de départ liés à la collaboration : annuaire, affiliations, projets, co-signatures. Mettre en œuvre la toxicologie prédictive grâce à une ontologie, qui serait le squelette d'une base de connaissance, pourrait être l'étape finale ou le scénario le plus ambitieux et le plus complexe. Ce potentiel est-il perçu et souhaité par les toxicologues ? Quelle communication a été faite à ce propos auprès d'eux ?

Dans une moindre mesure, assimilée à une terminologie structurée, une ontologie peut contribuer à améliorer la navigation et la recherche documentaire sur une plateforme comme Toxcea. En effet, implémentée à la manière d'un thésaurus, elle peut améliorer le taux de rappel du moteur de recherche d'une application grâce aux relations sémantiques. Une taxinomie peut être réalisée à partir des classes principales de l'ontologie pour structurer l'information diffusée sur un portail par exemple. Les classes principales de l'ontologie peuvent contribuer à la classification automatique de documents lors des opérations de traitement automatique du langage naturel (TALN) pour améliorer la catégorisation de ces derniers. Et pour les utilisateurs de la plateforme, certaines classes peuvent alors s'apparenter à des facettes par catégories ou des facettes additionnelles. Autre cas d'usage, intégrée au processus d'annotations, la gestion des ressources numériques peut en partie être automatisée. Elle peut aussi être implémentée à la manière d'un thésaurus à la recherche pour améliorer le taux de rappel du moteur de recherche d'une application. La terminologie se doit alors d'être complète en ce qui concerne la déclaration des équivalences entre les termes, les acronymes, les syntagmes nominaux.

Réaliser une terminologie riche et structurée pour tester son soutien à la recherche d'informations semble être un résultat concret, réalisable et évaluable. Mais l'amélioration de la recherche d'informations est-elle souhaitée explicitement par des toxicologues ? Cela représente-t-il vraiment une valeur ajoutée pour eux ? Dans cette perspective, ce projet reste toujours scientifiquement pertinent et en adéquation avec les usages envisagés à l'origine. Cependant, les usagers finaux de Toxcea sont-ils encore présents et potentiellement mobilisables ? Souhaite-t-on répondre uniquement aux besoins encore non sollicités d'un nombre restreints d'usagers ? Ou alors souhaite-t-on encourager voire même prescrire de nouveaux usages et promouvoir les potentialités des ontologies pour ce faire ? Comme hier, les réponses à ces questions restent déterminantes. Par ailleurs, de nouvelles contraintes liées à l'environnement du projet sont à envisager aujourd'hui.

L'environnement du projet

Avant d'envisager de financer un tel projet, d'en décider le budget et d'établir la liste priorisée et évaluée des actions à entreprendre ainsi que la méthode, il est nécessaire d'en évaluer les coûts. C'est le rôle d'une étude de faisabilité qui s'appuie sur le périmètre préalablement délimité dans une étude d'opportunité. Pour étudier la faisabilité technique et organisationnelle des fonctionnalités, l'environnement doit être pris en compte également.

Ce projet, tel qu'envisagé jusqu'à présent, n'est lié ni au système d'information, ni à la politique documentaire du CEA. Il peut éventuellement être lié à la politique de communication de la DSV. Concernant les autres parties prenantes, il n'entre pas dans le périmètre stratégique de l'un ou l'autre organisme. Cependant, il est profitable pour le LGI2P et l'Inserm de présenter des collaborations avec d'autres organismes. Un transfert de technologie en plus de d'expérimentation est aussi favorable.

Qui conçoit et qui opérationnalise cette terminologie actuellement ?

Actuellement l'ontologie n'est pas prise en charge. Elle a fait l'objet d'une révision à minima courant 2014 afin d'en corriger les erreurs et d'assurer au moins un label anglais et un label français pour chaque concept.

Qui utilise et exploite l'ontologie actuelle ainsi que le dispositif dans lequel il serait intéressant de tester son apport ?

L'ontologie n'est pas utilisée, la plateforme Toxcea peu fréquentée. La plateforme n'est pas utilisée pour rechercher de l'information contenue dans les publications. Et ceci, pas même par le service de veille et de documentation de la DSV, BioDoc.

Qui pourrait l'utiliser et la maintenir, pour quelles raisons et dans quel environnement ?

En ce qui concerne la plateforme Toxcea, les compétences présentes dans le service BioDoc, serait un gage de meilleure saisie des métadonnées descriptives des publications dans Toxcea. Une saisie de qualité, respectant les normes, est garante d'index opérationnels et de l'homogénéité de la présentation des résultats de recherche. Cela a de la valeur car cela permet d'éviter de rebuter les utilisateurs de Toxcea. Cependant, il s'agirait d'une charge de travail importante pour le service BioDoc et sans aucune contrepartie directe pour ce service. Ceci n'est donc pas souhaitable.
En revanche, il serait intéressant que les métadonnées des publications du programme ToxNuc référencées dans BioDoc puissent être reversées de façon automatique dans la base documentaire de Toxcea. Cela garantirait à minima une saisie correcte des métadonnées descriptives. A celles-ci s'ajouterait l'indexation thématique. Elle pourrait être semi-automatique grâce à l'ontologie, c'est-à-dire grâce à l'injection de la terminologie hiérarchisée enrichie des relations d'équivalence explicites dans la chaîne de TALN. Mais cette opération doit au final être contrôlée par un opérateur humain. Seuls des tests permettraient d'en évaluer la charge et au LGI2P d'expérimenter des recherches visant à la réduire au minimum.

Il conviendrait aussi d'étudier dans quelle mesure l'ontologie de la toxicologie nucléaire pourrait en partie être alignée ou enrichie avec la terminologie utilisée par le service BioDoc. Il est à noter que cette dernière émane de la Direction de la stratégie et des programmes du CEA. BioDoc n'utilise pas le INIS/ETDE thésaurus de l'AIEA contrairement au service IST du CEA. Il conviendrait aussi de vérifier si d'autres synergies seraient envisageables avec des partenaires internes et/ou externes.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	04

Fonctionnellement et techniquement, la plateforme Toxcea n'est reliée à aucun réseau interne ou externe du CEA. Cela semble a priori éliminer une contrainte. Cependant, cela peut constituer un frein à son évolution et pour celle des outils qui pourraient lui être adjoints comme l'ontologie. En effet, garantir les compatibilités, pour les échanges de données ou la participation à un système d'information commun, est une réalisation qui s'avère complexe. Elle implique de concevoir des interfaçages spécifiques entre les applications. Mais pour autant, sans cet effort, la plateforme peut perdre de son utilité et devenir archaïque par l'absence d'utilisateurs, directs ou indirects, demandeurs d'améliorations. Par ailleurs, la chaîne de TALN, telle qu'elle a été mise en place par le LGI2P et dans laquelle il a été prévu d'intégrer l'ontologie, est reliée à cette plateforme. Et c'est donc depuis Toxcea que les publications et les descriptions bibliographiques liées sont récupérées. A défaut de rester un outil vivant et en croissance, la plateforme devrait garantir à minima l'accès à la mémoire du projet : versions électroniques pérennes des documents, références ou texte intégral des publications générées par le programme. L'ontologie ne serait-elle pas un outil trop sophistiqué pour une plateforme peu utilisée ? En cas de tests utilisateurs, un certain nombre exprimeraient probablement la nécessité d'améliorer la recherche. Pour autant, utiliseraient-ils plus Toxcea en cas d'amélioration ?

Qui finance ou qui financerait ?

L'organisme qui a le pouvoir de décision final. On peut présumer qu'il s'agisse du CEA. Il semble cependant qu'un budget ait été envisagé pour un stage opérationnel de master 2 informatique de 4 mois minimum, rattaché au Knowledge and Image analysis for Decision (KID) avec accueil au LGI2P ou à l'Inserm UMR_S 1142.

De la précision des objectifs pour l'ontologie et des attentes pour la plateforme Toxcea, dépendent les personnels et les outils impactés, l'évaluation de la faisabilité et l'adhésion des parties prenantes.

L'importance stratégique et économique du projet

L'idée actuelle est toujours en cohérence avec l'objectif général et initial de mise en visibilité de la spécificité française au travers d'une ontologie pour représenter ce domaine. Cependant, il convient de s'assurer qu'aucun autre collectif n'ait déjà assumé la charge de cette entreprise. Le travail de recherche effectué durant le stage Cnam-INTD permet de constater qu'il n'existe pas d'ontologie concurrente pour ce domaine (cf. Livrable 3). Néanmoins le paysage technologique a changé. Les technologies du web sémantique basées sur des standards, garantissant l'interopérabilité des données et des vocabulaires dans lesquels elles sont encapsulées, arrivent à maturité. Elles sont à la fois une chance pour faire évoluer et publiciser l'ontologie actuelle tout en étant un cadre dont les contraintes sont à envisager et évaluer.

Les avantages

Les avantages ne sont pas directement d'ordre économique. Faire évoluer l'ontologie, c'est tendre à renforcer la visibilité de la spécificité française en toxicologie nucléaire. C'est une mission pour laquelle le CEA semble être l'acteur le mieux placé pour faire autorité en la matière. Pour lui, c'est aussi contribuer à l'innovation dans le traitement de l'information et l'informatisation des connaissances. Solliciter le TALN pour indexer et améliorer la recherche d'information, n'est pas une innovation en elle-même. Le TALN a déjà fait ses preuves en la matière. Cependant, ces traitements sont lourds et des recherches visent constamment à en améliorer les procédés. Actuellement, le LGI2P cherche aussi à combiner traitement statistique, foksonomie et richesse sémantique des ontologies pour augmenter le poids de certains termes quand ils correspondent à des labels.

Aux côtés de la documentation du programme, la documentation ayant trait à l'ontologie peut devenir un outil de gestion et de promotion des connaissances en toxicologie nucléaire. La documentation de son évolution, c'est-à-dire les raisons des choix successifs, est aussi un élément d'information supplémentaire sur les orientations de l'institution. En effet, l'ontologie se doit d'être adaptée pour rester en adéquation avec le domaine qu'elle est censée représenter. Ainsi son extension aux nanosciences est le marqueur de cette évolution. En cas d'extension, la modularité de l'ontologie doit être réfléchie et documentée. Ici le projet, surtout s'il s'oriente vers une terminologie plutôt que vers une ontologie orientée base de connaissance, peut être une charge sans réelle opportunité de gain direct. Mais il peut participer à promouvoir l'image du CEA non seulement dans le domaine des connaissances en toxicologie nucléaire, mais aussi dans un domaine qui s'inscrit dans les missions d'Allistene. Le CEA est un des membres fondateurs de cette alliance des sciences et des technologies du numérique innovant dans ces domaines.[2]

En outre, ce projet a l'avantage de s'appuyer sur une terminologie existante, validée et hiérarchisée de 649 termes. Il conviendrait de la compléter par des labels supplémentaires. Ceci par exemple à l'aide du INIS/ETDE thésaurus dont des spécialistes sont présents au CEA. Ce projet pourrait profiter de l'intérêt des différentes communautés pour le Linked Data. Les langages du Web sémantique ont une dimension sociale. Ils offrent l'opportunité de mutualiser les efforts de réalisation et d'obtenir des retours sur l'ontologie à une échelle plus importante. Ainsi, il existe des sources qui peuvent inspirer ou être utilisées pour enrichir et interconnecter l'ontologie actuelle afin de la rendre plus visible. Ces ressources font l'objet du livrable 3. En retour, d'autres communautés pourraient réutiliser ce vocabulaire de la toxicologie nucléaire car celui-ci est déjà encodé en OWL et respecte quelques éléments de la recommandation SKOS. A défaut de lier les données produites par le programme ToxNuc, sa terminologie peut alors être partagée grâce à l'utilisation des standards du Web sémantique. Mise à disposition sur des outils reconnus et proposant des services de retours (outils de feedback), les utilisateurs pourraient la commenter, signaler des dysfonctionnements et des suggestions. Il s'agirait aussi grâce à l'association d'une licence d'utilisation à une publication d'autoriser des contributeurs à faire fructifier ce référentiel. Mais cela contribuerait également à prendre connaissance des autres communautés travaillant sur le sujet et à favoriser de nouvelles collaborations scientifiques. Cette dimension est alors en concordance avec l'esprit initial du projet orienté vers la transversalité et la mutualisation. Par ailleurs, les compétences du LGI2P permettent de concevoir des cas d'usages innovants pour ce type de vocabulaire intégré dans des dispositifs numériques plus larges comme Toxcea et l'application Folksonomies.

Ce projet a donc des atouts. Il contribue aux objectifs stratégiques généraux du CEA en apportant un élément novateur au domaine de la toxicologie nucléaire.

La mesure du résultat

Quelle fiabilité de l'ontologie pour la classification automatique ?

Les indicateurs seront à établir avec les spécialistes du TALN. Les indicateurs devront déterminer s'il faut encore affiner les calculs nécessaires à cette aide à l'indexation. Ils devront également permettre d'évaluer les améliorations à apporter à la taxinomie à laquelle peut être assimilée aujourd'hui l'ontologie en question. Il est probable qu'il faille solliciter des opérateurs humains comme des experts du domaine, des professionnels de l'IST familiers de l'indexation dans ce domaine. Une communication est donc à prévoir en amont afin de les solliciter.

[2] http://www.enseignementsup-recherche.gouv.fr/cid50054/allistene-l-alliance-des-sciences-et-technologies-du-numerique.html et https://www.allistene.fr/ (pages consultées le 24 juillet 2014)

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	04

Quelle utilité pour la communauté scientifique ?

Il est possible de publier l'ontologie en y associant une licence garantissant l'utilisation de celle-ci avec obligation de citation. Ainsi, les cas de réutilisation sont des éléments de mesure du service rendu et de l'intérêt porté à cet objet. Les publications relatant l'appui sur la taxinomie sont aussi un élément de mesure de la visibilité souhaitée.

Les contraintes

Les délais pour effectuer les premiers tests

Aujourd'hui, contrairement aux origines du projet, le nombre de publications issues du programme est suffisant pour pouvoir y appliquer l'ontologie. Cependant, celles-ci nécessitent plusieurs traitements pour être exploitables, dont une reprise de leur signalement bibliographique. Actuellement, leur basculement dans la dernière version de la plateforme est en cours. De ce fait, les premiers tests ne peuvent être réalisés dans l'immédiat. Le délai ne peut pas encore être évalué, car il dépend des solutions envisagées et de leurs mises en place effectives.

Les aspects juridiques

La publication nécessite l'établissement d'une licence d'utilisation par le CEA. Le CEA est favorable à l'utilisation de licences de type Creative Commons (http://creativecommons.fr). Il les utilise pour d'autres créations mises à disposition sur Internet.

Les aspects techniques et organisationnels afin de pérenniser l'ontologie

Au regard des projets similaires pour d'autres communautés, la publication engage à pérenniser la ressource publiée. Elle engage, non pas juridiquement l'éditeur, mais l'image de celui-ci.

D'une part, documenter l'ontologie de plusieurs manières complémentaires est une garantie de sérieux et facilite son repérage ainsi que son identification. Documenter l'ontologie au sein même du fichier qui la constitue permet à ces informations d'être, elles aussi, exploitables par les machines : auteurs, contributeurs, éditeur, domaine d'application, sources, version, statut, historique. Ce procédé est nécessaire pour le référencement dans le catalogue LOV[3] par exemple. Il est préférable d'exprimer ces informations en plusieurs langues en combinant les propriétés de plusieurs schémas de métadonnées comme FOAF, VOAF et Dublin Core. De plus, l'édition devrait être accompagnée d'articles expliquant les objectifs remplis par cette ontologie et les lacunes qu'elle comble. Une page HTML devrait compléter la documentation autoportée par le fichier de l'ontologie. Ainsi une page pourrait lui être dédiée sur le site Toxcea avec la mention des références publiées à son propos. Pour simplifier sa maintenance, des outils comme Neologism, OntoSpec, LODE et Parrot génèrent la documentation HTML à partir du code RDF ou OWL. S'il est fait mention de maintenance, c'est aussi parce qu'une telle publication contraint d'une certaine façon son éditeur à pérenniser l'ontologie en veillant à sa mise à jour.

D'autre part, rassurer sur l'actualisation programmée ou régulière peut encourager les potentiels réutilisateurs. Pour cela, il est souhaitable de compléter le dispositif avec des outils de feedback. En contrepartie, il est nécessaire de s'astreindre à répondre aux signalements d'erreurs ou manques éventuels, de corriger ce qui doit l'être et publier de nouvelles versions en conséquence. Par ailleurs, il est intéressant que la documentation HTML signale également les perspectives que l'autorité éditrice envisage pour cette ontologie (maintenance et développement ultérieur).

[3] http://lov.okfn.org/dataset/lov

Certes l'évolution du Web vers un web encore plus social et interconnecté renforce les potentialités de mutualisation et de répartition de l'effort pour la construction d'une telle ressource. Mais elle n'en réduit pas les charges de maintenance, bien au contraire. Ainsi même si l'ontologie de la toxicologie nucléaire est envisagée à minima comme une terminologie, sa finalisation et sa publication nécessitent de coordonner plusieurs actions. Ainsi les aspects techniques, informatiques, juridiques, organisationnels et informationnels sont à piloter avant, pendant mais également après le développement. Pour l'après mise en œuvre, il s'agit de concevoir un ou des scénarios de continuité. L'étape suivante est une revue d'opportunité basée sur l'analyse du recueil de besoins. Elle explore les scénarios pour aider à la définition du projet. Enfin, si la décision de reprendre le projet est prise, une étude de faisabilité clôt la phase d'avant-projet en estimant la charge et le phasage. Les coûts, la répartition des tâches ainsi que l'attribution des responsabilités d'animation et de suivi devront y être étudiés.

Le déroulement général, les ressources nécessaires et les risques spécifiques

En l'absence d'objectifs plus précis, il est difficile d'établir le déroulé du schéma de réalisation global pour le projet d'ontologie de la toxicologie nucléaire au sein de la plateforme Toxcea. Il en est de même pour établir un calendrier prévisionnel. En ce qui concerne les ressources extérieures mobilisables (cf. livrable 3), leur étude sera à actualiser et à approfondir à l'aulne de la redéfinition du projet. L'étude de faisabilité aura aussi à envisager la nécessité de faire appel à la sous-traitance interne ou externe.

L'un des risques spécifiques pour ce projet est le ralentissement du programme Transversal et par là même la diminution de la consultation de la plateforme Toxcea. Toute entreprise faisant appel à la gestion des connaissances nécessitant de mobiliser plusieurs experts est coûteuse en temps et en énergie. C'est là aussi l'un des risques majeurs pour l'aboutissement du projet. Et c'est un risque auquel il a déjà été confronté depuis 2001. Pour éviter cet essoufflement de la part des différents acteurs, il est préconisé de se mettre en mode projet, comme l'envisage le Pôle GeCo - Gestion des Connaissances de l'INRA. Il s'agit alors de mettre en place une dynamique permettant de maintenir la motivation des différents participants. Pour éviter la concurrence entre les tâches quotidiennes et les tâches relatives au projet, le porteur du projet devrait aussi consacrer la majorité de son temps au projet. Les difficultés spécifiques à ce projet sont aussi liées à la distance professionnelle et géographique. Pour diminuer les temps de coordination et maintenir l'implication, il est préférable que le porteur de projet travaille avec une équipe resserrée. Il est par ailleurs également souhaitable que l'équipe colocalisée et que ses membres soient pleinement disponibles. Si l'équipe consacre beaucoup de temps au projet sur un temps court, autrement dit s'il s'agit d'une organisation de type task force, le projet a plus de chances d'aboutir.

Par ailleurs, les risques classiques d'un projet sont à envisager. Il s'agit des risques liés au budget, à la communication, à l'indisponibilité des membres pour des raisons diverses. Il s'agit encore des risques techniques, comme ceux liés à l'hébergement, à la sauvegarde des applications, mais aussi l'émergence de nouveaux formats, de nouvelles recommandations. Il peut éventuellement s'agir de se former à de nouveaux outils comme par exemple à ceux nécessaires à l'alignement de l'ontologie ToxNuc avec d'autres terminologies (OnAGUI, SMART, PROMPT). Les risques liés à la capitalisation et à la documentation du projet sont de plusieurs ordres. La documentation doit être accessible, simple d'utilisation, maintenable pour rester fiable et utilisable. Un soin devra être apporté au versionning de l'ontologie comme à celui de sa documentation de même qu'à sa sécurisation. La documentation est capitale pour évaluer les compétences nécessaires à la future équipe. Cette équipe encore plus resserrée est celle qui aura ensuite la charge maintenir et de faire évoluer le dispositif. Elle pourra alors s'appuyer sur cette documentation.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	04

Comme le montre le tableau suivant, la reprise du projet engendre de nouvelles charges. Et cela quel que soit le scénario choisi :

- amélioration de l'indexation et amélioration de la recherche de données bibliographiques ou de documents primaires
- conception d'un portail sémantique
- base de connaissance

Chacun d'entre eux nécessite des compétences en informatique et en ingénierie des connaissances mais pas seulement. La proximité avec des toxicologues est également nécessaire pour valider l'ontologie et les raisonnements produits. Un accompagnement du changement est donc nécessaire avec le soutien d'une volonté politique affirmée. En plus de cette volonté, il peut être nécessaire de créer une urgence pour reprendre et faire aboutir le projet. Cependant, le rapport coût/efficacité dans une période de resserrement budgétaire doit être mis en regard des gains. Ces derniers sont d'ordre symbolique comme l'image du CEA, les potentielles synergies entre communautés scientifiques et les diverses mutualisations possibles. Tous ces éléments font que ce type de projet n'est pas purement technique et informatique. En cas de non réalisation, des démarches pourraient malgré tout être entreprises. D'une part l'ensemble des réflexions et travaux menés autour du projet d'ontologie pourrait être capitalisés pour éclairer des projets similaires à l'avenir. D'autre part, il pourrait s'agir d'améliorer le signalement des ressources de la base documentaire de Toxcea. Ceci afin de garantir l'archivage et l'accès pérenne aux publications, ou encore faciliter les migrations futures. Des bénéfices pourraient d'ailleurs en être retirés pour améliorer la gestion de contenu d'autres programmes de recherche du CEA à l'avenir.

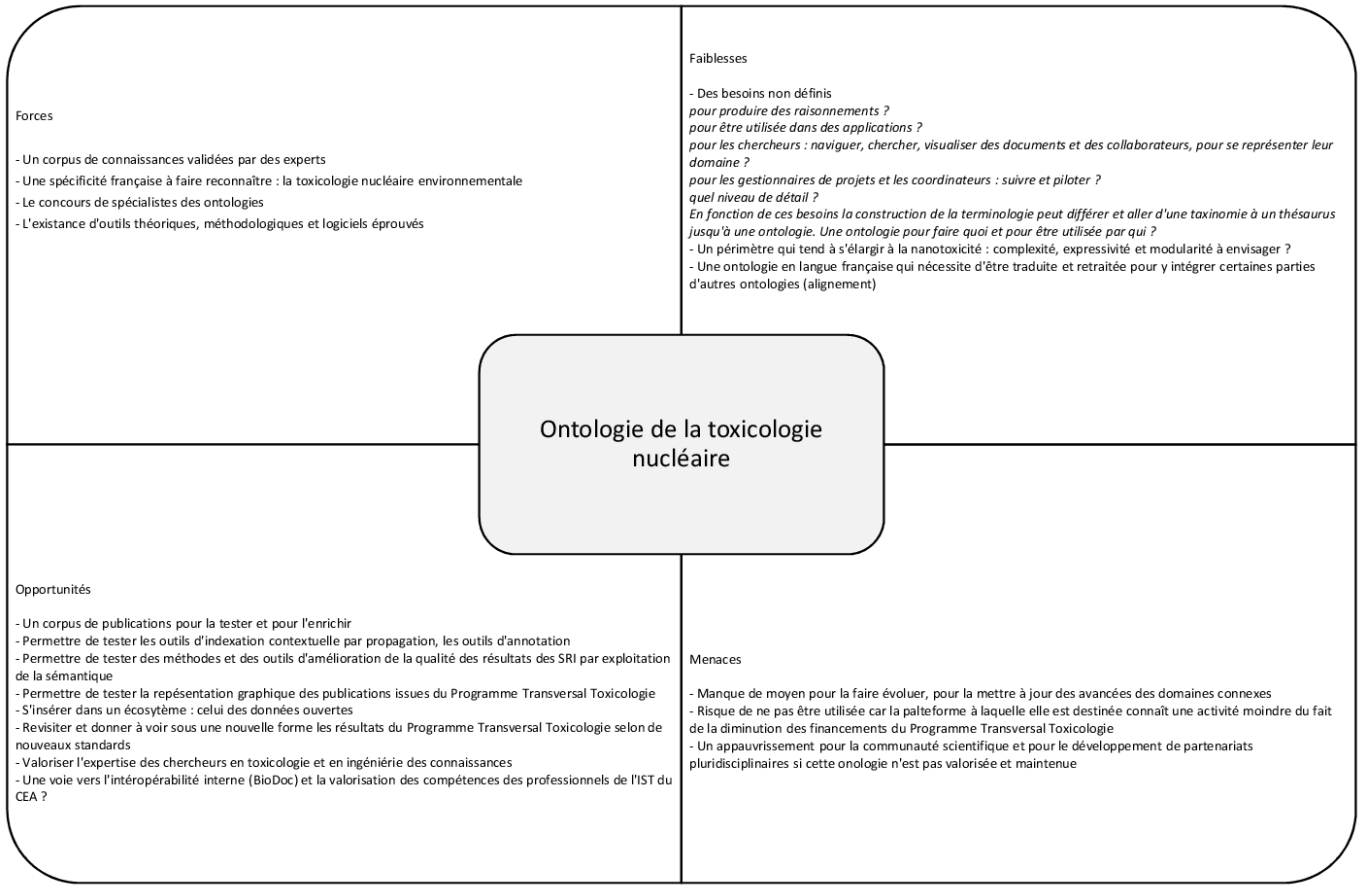
Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Quelles conséquences éventuelles en cas de réalisation et de non-réalisation du projet de reprise de l'ontologie ToxNuc ?

En cas de / Risques	non-précisions (attentes et besoins)	réalisation sans publication	réalisation et publication sans respect des standards	réalisation et publication sans mise à jour	non-apport de valeur à d'autres services du CEA (BioDoc, l'IST, HAL-CEA, Direction de la stratégie et des programmes)	non-réalisation
stagnation, mise en attente de la terminologie actuelle						
charges supplémentaires						
insuffisances, défauts de conception de l'ontologie						
manque de ressources pour la maintenance de Toxcea						
ontologie perçue uniquement comme une charge						
être menacé d'obsolescence						
non-interopérabilité pour l'échange de données						
moindre visibilité						
non-réutilisation par d'autres communautés						
pas de signalement d'erreurs						
pas de suggestion d'amélioration						
image négative pour le CEA						
opportunité de collaboration en moins						
engendrer la refonte de la plateforme Toxcea si les ressources sont suffisantes et les utilisateurs présents						
manquer de ressources et de compétences						

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	03

Le swot de l'ontologie ToxNuc



Annexe 3
Livrable 5 - Correction de l'ontologie formalisée en OWL

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	05

Commanditaires	Commande n°2
• Direction des sciences du vivant – DSV, Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) • Inserm UMR_S 1142 • Laboratoire de Génie Informatique et d'Ingénierie de Production – LGI2P, Centre de recherche de l'Ecole des Mines d'Alès – EMA (Nîmes)	**Correction de l'ontologie formalisée en OWL**

Rappel de la commande

Objectif général	Affiner l'ontologie
Contexte de la demande	
Environnement interne	Plateforme collaborative créée par le LGI2P et maintenue par la société ID-Alizée pour le collectif ToxNuc coordonné par le Commissariat à l'Energie Atomique et aux énergies alternatives (CEA)
Environnement externe	Expérimentations de nouvelles méthodes pour accompagner l'opérateur humain dans son investigation de vastes corpus de connaissances
Résultats attendus	
Tâches	A l'aide de l'éditeur d'ontologie Protégé : - enlever certaines formes plurielles - assurer la présence de labels français et anglais pour chaque classe de l'ontologie en traduisant certains labels d'une langue à l'autre - modifier et compléter l'ontologie selon les retours des experts du domaine
Calendrier	Du 2 juin 2014 au 22 août 2014
Personnes mobilisables	J. Charlet (AP-HP & INSERM UMR_S 1142) Marie Thérèse Ménager (CEA) G. Dray, S. Ranwez (LGI2P – EMA Nîmes)
Projet de rattachement	
Equipe de recherche	Knowledge and Image analysis for Decision - KID
Laboratoire	Laboratoire de Génie Informatique et d'Ingénierie de Production – LGI2P, Centre de recherche de l'Ecole des Mines d'Alès – EMA (Nîmes)
Projet	Automatisation des différentes phases d'analyse, d'indexation et d'exploitation des informations
Entre autres objectifs	Etude des apports des standards du Web sémantique et de la modélisation des connaissances à l'aide d'ontologies de domaine à ce type de projet
Mots-clés	Curation des données. Langage de représentation des connaissances. Ontologies. Protégé (logiciel). Qualité des données. Terminologie, Web Ontology Language (OWL)

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	05

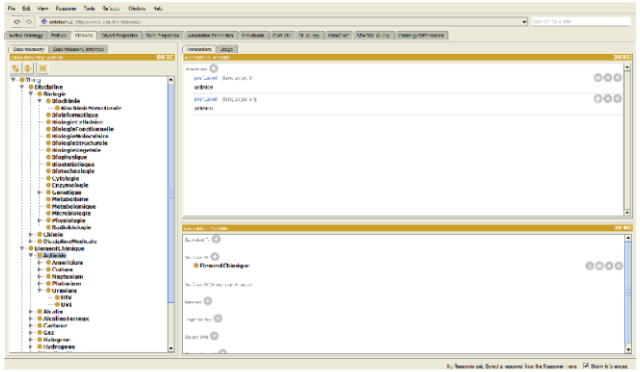
Plan

Auteur : Anne-Claire Le Picard	Date de création : 06-2014	Page **2** sur **9**
Nom du fichier : 2014_stage_intd_livrable5_affinage_fichier_owl_ontotoxnuc.pdf	Date de modification : 09-2014	N° de version : 01

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	05

Protégée, l'outil pour agir sur le fichier OWL

Le fichier a été nettoyé à l'aide de l'éditeur Protégé 4.3.0. Protégé est un éditeur Open Source pour la création d'ontologies. Il est développé par le Stanford Center for Biomedical Informatics Research de la Stanford University School of Medicine. Il existe une version en ligne de cet outil, mais c'est une version locale qui a été utilisée durant le stage INTD.



Vue arborescente de l'ontologie ToxNuc avec l'éditeur Protégé.

Les problèmes rencontrés

Les erreurs rencontrées

- un problème d'encodage pour les noms de classe
- des fautes de frappe
- des URI mal formées
- des labels absents
- des labels avec de mauvais libellés d'annotation
- des hiérarchies incohérentes

Là où subsistent des doutes

- le nommage des sous-classes de la classe principale « ModeleBiologique »
- la distinction entre les classes principales « Outil » et « TypeEtude »

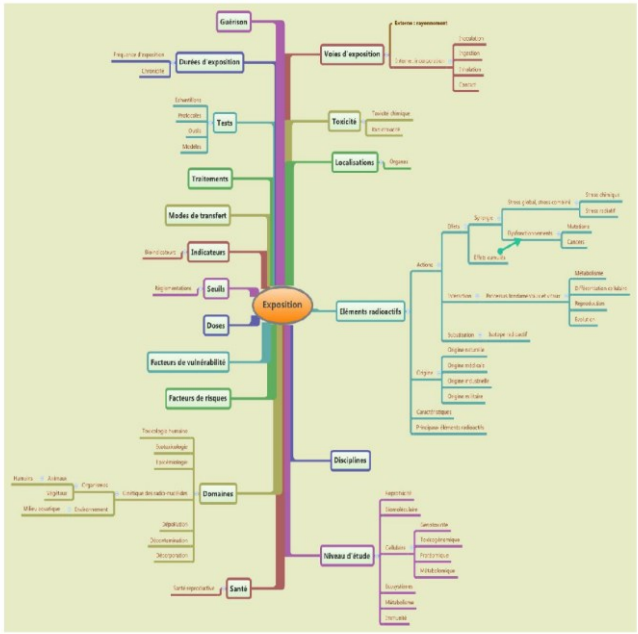
Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	05

Les actions qu'il aurait été intéressant de mener

- proposer et discuter des schémas de relations inspirés du processus de la toxicité
- tester quelques relations et inférences au sein de l'ontologie ToxNuc

Processus de la toxicité

Quoi ?	Eléments chimiques et radioactifs
Nature ?	Action selon l'état (solide, liquide, gazeux)
Comment ?	Effets - Réactions
Quand ?	Situation - Exposition
Où ?	Animaux : rats, humains — Environnement : milieux, végétaux



Les solutions d'attente

Concernant les polyhiérarchies, celles-ci ont été le plus possible évitées. Car, d'après un échange avec Jean Charlet, il est préférable de les éviter. En effet, en cas d'évolution vers une ontologie avec une expressivité plus forte, les polyhiérarchies peuvent engendrer des incohérences. Il en reste cependant quelques-unes. Ainsi le concept « Fluorochrome » a pour père à la fois le concept « Biomarqueur » et le concept « OutilDeLaBiologieCellulaire ». Certaines ont été éliminées en renommant des classes, comme dans le cas de Procaryote : « CelluleProcaryote » et « MicroorganismeProcaryote ». Des relations pourraient éventuellement être une solution supplémentaire pour les éviter.

Les méthodes et outils pour détecter des problèmes

Le repérage visuel dans Protégé

La détection des erreurs a principalement été réalisée de façon visuelle. Les problèmes d'encodage pour les noms de classes ont donc tous été corrigés à la suite de ce repérage. A cette occasion toutes les URI ont été revues et corrigées lorsque cela était nécessaire. De même, les lacunes concernant les annotations ont été détectées en faisant défiler les concepts.
Le défilage rapide a dans de nombreux cas empêché de constater les erreurs de libellés, comme « comment » ou « description », à la place de « label » ou « altLabel » au niveau des annotations.

Les moyens informatiques dans Protégé

Des requêtes SPARQL (Simple Protocole and RDF Query Language) peuvent être exécutées afin de détecter certains types d'erreurs, comme celles liées aux annotations évoquées ci-dessus. De même l'utilisation d'un plugin pour la visualisation sous la forme de graphe dans Protégé permet probablement de repérer des erreurs ou du moins les polyhiérarchies. Cependant, le manque de connaissances informatiques a conduit à explorer d'autres voies pour poursuivre ce repérage et le travail de correction. Ainsi, un outil dédié à la visualisation a permis de plus facilement détecter les erreurs du fichier OWL, toujours de manière visuelle et sans requête.

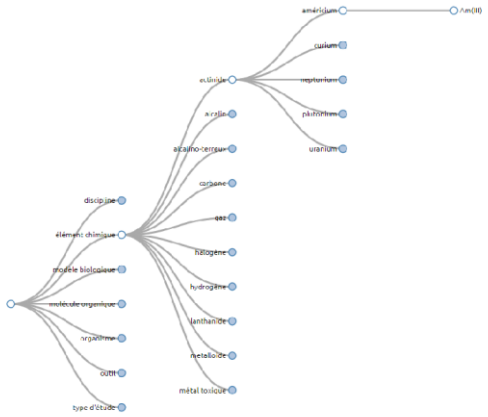
SKOS Play ! : un service de visualisation de thesaurus, taxonomies ou vocabulaires au format SKOS

Il s'agit d'un service en ligne gratuit[1]. Il est mis à disposition par le consultant Thomas Francart depuis son blog (http://blog.sparna.fr) à l'adresse suivante : http://labs.sparna.fr/skos-play. Cet outil génère différentes vues, dont l'une est dynamique, de thesaurus, taxonomies ou vocabulaires formalisés selon la recommandation SKOS.

> « SKOS, Simple Knowledge Organization System (Système simple d'organisation des connaissances) est une recommandation du W3C publiée le 18 août 2009 pour représenter des thésaurus, classifications ou d'autres types de vocabulaires contrôlés ou de langages documentaires. S'appuyant sur le modèle de données RDF, son principal objectif est de permettre la publication facile de vocabulaires structurés pour leur utilisation dans le cadre du Web sémantique. ». Wikipedia
> < http://fr.wikipedia.org/wiki/Simple_Knowledge_Organization_System >
>
> D'après la traduction réalisée par Thomas Francart de la recommandation W3C du 18 août 2009 (http://www.w3.org/TR/2009/REC-skos-reference-20090818), il s'agit d'« un modèle de données partagé pour échanger et relier des systèmes d'organisation de connaissances sur le Web ». < http://www.sparna.fr/skos/SKOS-traduction-francais.html >

[1] Il est aussi possible d'installer une version locale : «[…] you can install a local version of SKOS Play to work around the 5000 concepts limit on the online server. » < http://blog.sparna.fr/new-version-of-skos-play-for-taxonomy-visualization > (consulté le 26 août 2014)

Cet outil permet également de transformer des fichiers OWL en SKOS afin de pouvoir visualiser la hiérarchie de concepts contenus dans ce type de fichier. Ainsi le fichier de l'ontologie ToxNuc lui a été soumis. Il est possible de générer un arbre dynamique (cliquable) comme présenté par la figure ci-dessous, mais aussi une visualisation Carrée ("Icicle") et une visualisation Circulaire ("Sunburst").



L'outil génère également en HTML des listes et index plus classiques qui sont donc cliquables, ainsi que des fichiers d'impression PDF de ces listes et index. Ainsi plusieurs documents ont été générés à partir de la dernière version du fichier OWL. Ils sont joints aux livrables :

- Listes des concepts avec leurs traductions : liste tous les concepts, avec leurs traductions et tous leurs attributs (génériques, spécifiques mais pas de note ni de définition dans cette terminologie)
  liste_concepts_avec_attributs_traductions_fr_en.pdf
  liste_concepts_avec_attributs_traductions_en_fr.pdf
- Tableaux de correspondance des langues : table de traduction des libellés
  tableau_correspondance_langues_fr_en.pdf
  tableau_correspondance_langues_en_fr.pdf
- Index alphabétiques avec attributs : broader ou niveau le plus haut de la hiérarchie et terme générique (TT et TG), narrower ou terme spécifique (TS) (pas de related ou terme associé (TA) ni notation ou autres notes)
  index_alphabetique_avec_attributs_fr.pdf
  index_alphabetique_avec_attributs_en.pdf
- Arbre Hiérarchique :
  arbre_hierarchique_en.pdf
  arbre_hierarchique_en.pdf
- Index permuté :
  index_permute_fr.pdf
  index_permute_en.pdf
- Index KWIC ('KeyWord In Context')[2] :
  index_kwic_keyword_in_context_en.pdf

[2] L'index KWIC en français n'a pas pu être généré.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	05

Dans un premier temps, SKOS Play ! a été utilisé pour détecter les erreurs, toujours de manière visuelle et sans requête. Le travail de correction a consisté en une succession d'allers et retours dans Protégé pour effectuer les modifications. Dans un second temps, il a été envisagé pour dialoguer avec les interlocuteurs du projet situé à distance. En effet, cet outil ne nécessite pas d'être installé comme le logiciel Protégé et il a pour avantage d'avoir une interface conviviale. La navigation en ligne peut aussi être considérée comme plus facilement manipulable que des listes et index sous forme de fichiers PDF. En revanche ces listes sont intéressantes à l'issue du stage pour accompagner le fichier OWL corrigé et en présenter une version stabilisée.
Cet outil, conçu pour démontrer et illustrer le fonctionnement de certaines technologies sémantiques, a été détourné pour servir au contrôle de l'ontologie. Il a permis d'amorcer le dialogue pour susciter des retours sur les labels et les relations hiérarchiques de l'ontologie actuelle. Enfin il a permis de générer très facilement des listes et des index, qui sans cela, n'auraient probablement pu être présentés à l'issu du stage.
Par ailleurs, rappelons qu'à l'origine cette terminologie se présentait sous la forme d'un tableur car informatisée avec le format Excel. Ce test est donc aussi l'occasion de constater que le choix d'opérationnaliser la terminologie en un fichier OWL la rend facilement manipulable depuis le Web. Surtout ce test démontre qu'elle est conforme aux standards et qu'elle possède déjà certains éléments lui permettant d'être transférable dans l'univers du Web de données.

Partager des vues pour partager un langage commun, partager des vues pour faire connaître ce langage

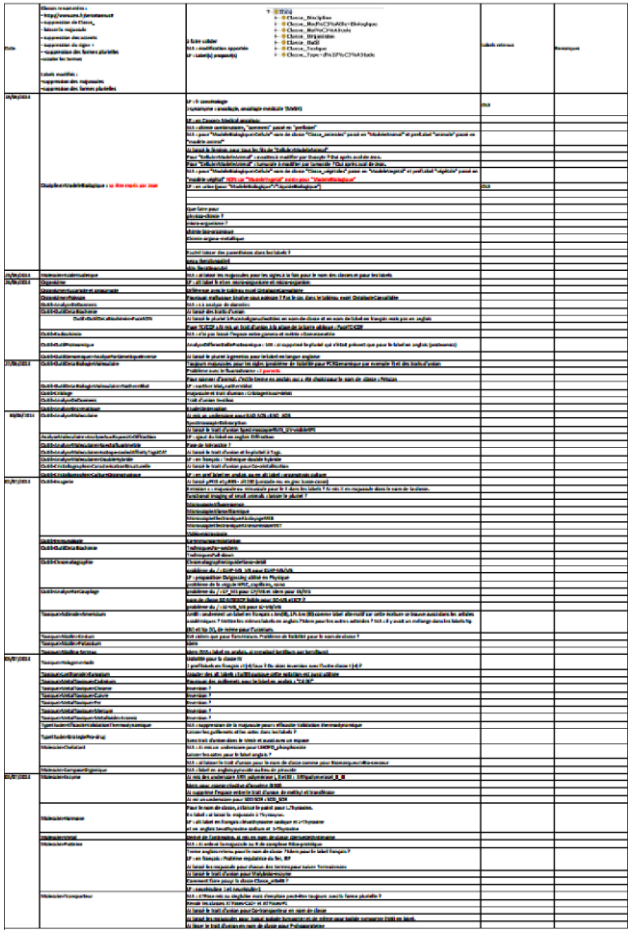
Ce type d'outil pourrait être envisagé comme un outil de management visuel et comme un outil de promotion supplémentaire :

- aide à la conception en cas de reprise du projet pour représenter les hiérarchies
- aide à l'instauration du dialogue avec les utilisateurs non informaticiens participant à sa conception ou à sa maintenance pour détecter les anomalies
- aide à la navigation hypertexte dans la terminologie pour indexer des contenus
- aide à la valorisation de l'ontologie
- aide pour susciter les feedback des utilisateurs externes en cas de mise à disposition

Les usagers sont de plus en plus habitués à visualiser des données. Les outils utilisés dans le contexte des activités professionnelles doivent également être attrayants. Il est préférable que ceux-ci correspondent aux habitudes des utilisateurs. Si ces interfaces font sens pour eux, il est permis de penser qu'elles susciteront une plus grande réactivité de la part des différents participants.

Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	05

Annexe : deux documents de suivi des modifications du fichier OWL



Nom du projet	N° du livrable
Conception d'une ontologie du domaine de la Toxicologie Nucléaire (ToxNuc)	05

*ModeleBiologique>ModeleVegetal>Organite :
Ajout de labels :
altLabel fr : organite cellulaire
altLabel en : cell organelle
altLabel en : organoid

*Classe Toxique renommée ElementChimique
ChimieMinerale aurait pu convenir si cette classe n'était pas déjà présente dans la classe Discipline
Labels à corriger >> C'est fait

*La classe Molecule a été renommée MoleculeOrganique
Labels à corriger >> C'est fait

*La classe Molecule>Metal a été renommée MoleculeOrganique>OrganoMetallique
Labels à corriger >> C'est fait

*MoleculeOrganique>ComposeOrganique>CaratenoideZeaxanthine
La caraténoïde zéaxanthine est une vitamine or il existe la classe Vitamine dans la classe
MoleculeOrganique. A revoir avec des spécialistes.

*MoleculeOrganique>OrganoMetallique
Travailler sur le concept Bicarbonate
Composelode : attention si volonté de renommer cette classe car une classe Iode est présente dans la classe Toxique renommée ElementChimique

*Faire un lien, voir ce qui pourrait être fait... entre les éléments chimiques de la classe
ElementsChimique>Metalloide et ceux de la classe
MoleculeOrganique>OrganoMetallique

*Ajout de la classe Iode129 dans la classe ElementChimique>Halogene>Iode
Labels à ajouter

*La classe TypeEtude pose problème

*La classe Outil n'a pas été observée

Par ailleurs, le concept Fluoramine a deux pères.

Voilà j'espère n'avoir rien oublié ni embrouillé
Merci et à bientôt,
Anne-Claire

Le 18 juillet 2014 17:34, Anne-Claire Le Picard <ac.lepicard@free.fr> a écrit :

Voici une liste de ce que nous avons vu avec Marie-Thérèse Ménager :

*Discipline>disciplineMedicale>Cancerologie
Labels modifiés et labels ajoutés validés

*ModeleBiologique>Cellule>CelluleAnimale :
Classe Peaukeratynocite devenue Keratynocite
altLabel : kératynocite (peau) et keratynocite (skin) supprimés

*MTM a validé la nouvelle arborescence de la classe Organisme réalisée avec Jean

*ModeleBiologique>Cellule>Vegetal :
Classe Vegetal renommée CelluleVegetale
Labels corrigés en conséquence

*ModeleBiologique>ModeleVegetal :
La sous-classe ModeleMicroOraganisme a été remontée d'un niveau ; ses sous-classes ont été supprimées
Sous ModeleBiologique>ModeleVegetal, il y a désormais
Cellule
LiquideBiologique
ModeleMicroOrganisme
ModeleVegetal
Organe
Y inclure la classe Organisme ? Et retravailler les sous-classes de celle-ci avec la classe
ModeleBiologique>ModeleVegetal ?
Ce à quoi il faudra réfléchir :
Redondance du concept Algue dans les sous-classes de Organisme
et de
ModeleBiologique>ModeleVegetal
Redondance du concept racine dans les sous-classes de
ModeleBiologique>Cellule>CelluleVegetale
et de
ModeleBiologique>ModeleVegetal

Auteur : Anne-Claire Le Picard	Date de création : 06-2014	Page **9** sur **9**
Nom du fichier : 2014_stage_intd_livrable5_affinage_fichier_owl_ontotoxnuc.pdf	Date de modification : 09-2014	N° de version : 01

Annexe 4
Entretien avec Éric Dagiral à propos de l'ontologie pour Orphanet

Entretien avec Eric Dagiral à propos du collectif du portail Orphanet 28 août 2014
Maître de conférences en sociologie à l'université Paris Descartes
Titulaire de la Chaire CNRS-Université "Technologie et lien social"
http://recherche.parisdescartes.fr/CERLIS/Equipe/Membres-statutaires/Dagiral-Eric

Anne-Claire Le Picard : introduction

- Exposition de la formation du Titre 1 dispensée au Cnam-INTD « Chef de projet en ingénierie documentaire » en lien avec une expérience professionnelle dans les bibliothèques depuis 1998.
- Présentation de la mission de stage portant sur une terminologie de la toxicologie nucléaire environnementale.
- Expression des souhaits quant à cet entretien par rapport à un travail sur les différents usages des ontologies, sur leur évaluation, la maintenance pérenne par les opérationnels d'un dispositif sociotechnique auquel on intègre une ontologie[1] : recueillir de l'information sur les phases en amont, sur la réflexion menée ou pas sur les enjeux, sur les conséquences pour la suite d'un projet. Soit quels gains, quelle rentabilité, et quels sont les coûts associés pour le collectif de production ?

Eric Dagiral : L'enquête[2] menée avec deux collègues sur la Plateforme Maladies Rares n'était pas dédié à Orphanet. Il s'agissait d'étudier les collaborations entre acteurs d'associations de patients, de la recherche et de la documentation en relation avec des dispositifs numériques. Nous avons alors co-construit le protocole d'ethnographie avec les différents acteurs.
Ce qui m'a fait m'intéresser à ces questions d'ontologie ne commençait pas par l'étude d'un objet ontologique. Nous nous intéressions à la constitution de connaissances qui se trouvent structurées dans une base de données relationnelle, mais aussi à l'informatique, soit des équipes techniques, des ingénieurs informaticiens. Et observer qu'en même temps il fallait élaborer, structurer, mais aussi écrire, rassembler des informations sur les maladies rares. Donc observer jamais des informaticiens tous seuls, mais des médecins avec différentes spécialités : cliniciens, biologistes... Observer finalement un ensemble d'acteurs hétérogènes qui doivent s'articuler.
Il est intéressant d'observer an amont les questions qui se posent :
Quelle est la solution qu'il faudrait mettre en œuvre ?
De quoi on parle ?
La question « De quoi a-t-on besoin ? » avant d'être amené à se dire que c'est d'une ontologie dont on aurait besoin. Et au final pourquoi ce machin, dont déjà le terme est encore assez peu connu, peu maîtrisé, et dont le lien avec l'informatique n'est pas a priori évident.
On pourrait dire que c'est peut-être dû une concomitance de facteurs :

[1] A propos de l'ontologie implémentée dans la base dont les données sont consultables librement via le portail multilingue de la plateforme **Orphanet** : http://www.orpha.net
OntoOrpha est une ontologie des maladies rares qui comprend des signes cliniques (phénotypes) et les gènes. La nomenclature complète Orphanet (étiquettes, synonymes) est disponible en six langues avec les références annotations externes (CIM, OMIM, HGNC, GENATLAS, Swissprot). Elle comprend aussi les données épidémiologiques relatives aux maladies rares en Europe (classe de prévalence, l'âge moyen d'apparition, l'âge moyen au décès) extraites de la littérature. Elle fournit en définitive un vocabulaire structuré pour les maladies rares, la capture des relations entre les maladies, les gènes et d'autres caractéristiques pertinentes, dans une langue compréhensible par les ordinateurs directement.
Par ailleurs des jeux de données sont également téléchargeables depuis le site **Orphadata** : http://www.orphadata.org/

[2] Etude ethnographique avec suivi durant 3 ans. « Les données recueillies in situ sont issues pour l'essentiel d'entretiens avec la direction et les membres des différentes équipes de travail, d'une série d'observations des réunions d'équipes et des situations concrètes de travail (activités rédactionnelles, conception et usages de la base de données), ainsi que de l'analyse d'un ensemble de documents internes (données statistiques, règles d'écriture des articles encyclopédiques et des résumés, documents et fichiers intermédiaires, traces d'activité informatisées...). » extrait de DAGIRAL É., PEERBAYE A. « Les mains dans les bases de données : connaître et faire reconnaître le travail invisible ». Revue d'anthropologie des connaissances [En ligne]. 2012. Vol. 6, 1, n°1, p. 229. Disponible sur : < http://dx.doi.org/10.3917/rac.015.0229 > (consulté le 27 août 2014)
« Cette recherche a été réalisée dans le cadre d'un Partenariat Institutions Citoyens pour la Recherche et l'Innovation (PICRI), financé et soutenu par la Région Île-de-France, ainsi que dans le cadre du projet BASICOM (Bases Informatiques et Coopération entre Mondes sociaux) financé par l'Agence nationale de la recherche. » extrait de DAGIRAL É., PEERBAYE A., « Voir pour savoir » Concevoir et partager des « vues » à travers une base de données biomédicales, Réseaux, 2013/2 n° 178-179, p. 163-196. DOI : 10.3917/res.178-179.0163

Entretien avec Eric Dagiral à propos du collectif du portail Orphanet 28 août 2014
Maître de conférences en sociologie à l'université Paris Descartes
Titulaire de la Chaire CNRS-Université "Technologie et lien social"
http://recherche.parisdescartes.fr/CERLIS/Equipe/Membres-statutaires/Dagiral-Eric

- Dans une certaine histoire la dénomination d'une technologie comme les ontologies était parvenue à une certaine maturité. Elle faisait donc du sens pour un certain nombre d'acteurs assez hétérogènes de ce collectif.
- La proximité entre ces différents acteurs
- La maturité de technologies Web associée au fait que la recherche d'information sur les maladies avait du sens en ligne
- Le fait qu'on parle de web sémantique
- Que pour produire les services d'Orphanet, il était question de catégorisation, de classification (du point de vue médical)
- La nécessité de faire évoluer cette base en ligne pour répondre au besoin de plus en plus grand de la rendre visible, mais aussi de la relier au travail de l'équipe qui œuvre pour l'enrichir et la mettre à jour, et de la relier d'autres initiatives en France, en Europe et dans le monde. La satisfaction d'un besoin qui confrontait à un problème d'interfaçage.
  Et l'articulation avec d'autres projets peut être considérée de plusieurs manières :
  - Une manière politique et scientifique : connaître les gens, se rapprocher d'eux pour collaborer avec eux, imaginer des projets communs
  - Une appréhension technique : la rencontre avec des acteurs qui font la promesse qu'un langage commun pourrait permettre à des formes techniques différentes, avec des histoires différentes (langage, objectifs) d'être mises en relation. Et donc un petit peu par la magie de cet opérateur qui est l'ontologie, articuler des mondes très différents dans un projet commun.

ACLP : Qui a sollicité explicitement de développer une ontologie ? La direction d'Orphanet ?
ED : Je n'ai que peu d'éléments. Je peux seulement dire que mon sentiment est que ce projet à émergé au sein de l'organisation d'Orphanet qui n'est pourtant pas une organisation très unifiée. Et malgré des cultures hétérogènes, des rapports de métiers différents, tous ces acteurs ont pu observer qu'il y avait des évolutions des systèmes d'information en santé (système d'aide au diagnostic, système d'aide à la décision médicale) des termes apparaissaient de plus en plus, que dans ces termes il avait l'idée d'ontologie. Mais chacune des communautés ayant sa définition de l'ontologie : pour les documentalistes, un ensemble de termes communs, de thésaurus et pour les informaticiens programmeurs, un ensemble de spécifications communes liées à des langages de programmation relatifs au web et aussi relatives à des institutions du Web comme le W3C. Et puis, dans les congrès le mot ontologie apparaissait de plus en plus, ce qui ne pouvait pas échapper à la direction d'Orphanet. Un autre type d'acteur a pu aussi contribuer à irriguer cette idée. C'est le conseiller associé au développement d'Orphanet, qui lui faisait de la prospective et qui venait faire de la formation aux ontologies aux équipes techniques, uniquement au départ, d'Orphanet pour voir comment on pouvait imaginer de structurer et assurer la pérennité de la base à l'avenir.

En résumé des préoccupations distinctes mais pour lesquelles semblent émerger la solution d'une ontologie pour toutes. Mais parle-t-on tous de la même chose en parlant d'ontologie ? Donc constatant ce fait au cours de dialogues, au cours parfois de quelques formations communes, il a fallu se lancer, car cela devenait un enjeu et un terme à l'ordre du jour. Mais je peux difficilement en dire plus. En fait, j'ai pu le constater en assistant à des réunions qui étaient plutôt techniques.

ACLP : A la lecture de l'article « les mains dans les bases de données [3] », j'ai cru comprendre que la justification de la réingénierie de la base était plus de répondre à la volonté de devenir plus visible sur le Web, et donc aussi plus reliée à d'autres bases présentes sur le Web. Et cela pour des questions aussi de légitimation de l'organisation auprès de ses tutelles, pour des questions liées à sa survie, et donc politiques. Il semble que les difficultés techniques des opérationnels, des documentalistes scientifiques, dont on peut lire les verbatims à ce propos dans votre article, n'aient pas été un déclencheur majeur de cette prise de décision.
ED : Oui une base sous optimale, jamais idéale. Oui c'est une lecture qui peut être faite de l'article, où la question de l'ontologie n'est pas centrale. Oui il y a peut-être de ça. La grande difficulté c'est de savoir comment l'ontologie, le projet d'une ontologie se pose dans l'organisation, à la fois dans une tension entre l'interne et les relations avec l'extérieur. Les relations avec l'extérieur peuvent sembler premières, et la question de l'ontologie comme dispositif sociotechnique nous a semblé la suivante : est-ce qu'on peut construire une ontologie sans reconsidérer complètement, depuis la base, l'organisation de la base et les formes de l'activité du travail. Dès lors au sein de l'organisation, on peut imaginer, et je pense l'avoir constaté, deux visions, et qui on l'imagine bien, peuvent devenir tout à fait conflictuelles à certains moments.
D'un côté, l'ontologie serait comme une couche logique supérieure qui permettrait de nouveaux interfaçages et qui pour autant serait relativement indépendante des façons de faire et des outils au quotidien des différents acteurs (beaucoup d'outils pour chacun d'entre eux). De l'autre côté, il y a l'idée, puisque le travail est si difficile à accomplir, que au contraire, cela serait l'opportunité pour pouvoir trouver ce qui serait le système idéal. Il y a l'idée que pourrait être trouvé le système idéal de la gestion de l'information au sein d'Orphanet dont le pivot serait l'ontologie. Et cela parce que les ontologues peuvent un peu faire la promesse qu'un système idéal d'organisation du travail, permettant la construction d'une information presque pure et parfaite, est possible. C'est une promesse extrêmement forte selon certains courants de la recherche en ingénierie des connaissances plus au moins évangélistes. Et se revendiquer de devoir être évangéliste dit beaucoup de la communauté technique, de l'ancrage dans le monde de l'informatique où l'idée d'évangélisation a une certaine histoire.

AC : Dans votre article « Les mains dans les bases de données… », vous parlez des parades pour avoir d'autres vues mises en places par l'équipe. Vous parlez de « médiations sociotechniques » pour toutes les pratiques et outils qui se déploient avec et autour de la base de données. Seuls certains de ces outils de gestion de l'activité sont documentés. Seules certaines pratiques ont fait, et tardivement l'objet de guides de procédures. Toutes les stratégies de mémorisation ne sont pas inscrites. La connaissance des documentalistes sur les connaissances contenues dans la base de données, et sur tous autres outils, n'est pas formalisée au moment où vous avez réalisé l'enquête. Vous mentionnez aussi la dimension collaborative. Par rapport à ce dispositif protéiforme, aux briques plus ou moins interfacées, à cette sédimentation d'outils, un seul outil comme un outil de GED pour documenter l'activité en train de se faire, pour centraliser tout ce qui participe à la production et à l'amélioration de la base de données, mais en dehors de celle-ci, n'a-t-il jamais été évoqué ?
ED : Si j'ai entendu ce mot.

[3] DAGIRAL É., PEERBAYE A. « Les mains dans les bases de données : connaître et faire reconnaître le travail invisible ». Revue d'anthropologie des connaissances [En ligne]. 2012. Vol. 6, 1, n°1, p. 229. Disponible sur : < http://dx.doi.org/10.3917/rac.015.0229 > (consulté le 27 août 2014)

Entretien avec Eric Dagiral à propos du collectif du portail Orphanet 28 août 2014
Maître de conférences en sociologie à l'université Paris Descartes
Titulaire de la Chaire CNRS-Université "Technologie et lien social"
http://recherche.parisdescartes.fr/CERLIS/Equipe/Membres-statutaires/Dagiral-Eric

AC : Je souhaiterais revenir sur un type d'acteur. Les documentalistes scientifiques qui ont une formation biomédicale étaient des documentalistes du fait de leurs activités. Certains parmi eux, avaient-ils des qualifications en gestion de l'information ?
ED : C'est difficile de répondre globalement car.il y a eu un taux de turnover importants de ces personnes. Il y a avait un mélange. Des personnes avaient une culture professionnelle de la documentation assez forte, pour d'autres c'était une acculturation et des usages très spécifiques tantôt liés au monde de la médecine ou de la biologie. Les gens avaient une culture de ce que sont les connaissances, de comment on les construit, de comment on y accède, de ce que sont les classifications et la catégorisation.

AC : Quels ont été les freins à la réingénierie de la base avec l'intégration d'une ontologie ?
ED : Il y a l'hétérogénéité des acteurs, mais peut-être plus encore la question de la hiérarchie des besoins et des attentes. Est-ce qu'on considère que malgré les difficultés, les documentalistes parviennent à travailler avec les outils ? Est-ce que c'est plutôt l'interface avec les autres dispositifs à un moment où l'un des enjeux d'Orphanet est l'interfaçage avec la Classification Internationale des Maladies ? Pour cette dernière attente, on a là typiquement un problème qui fait sens pour des ontologues.
Du point de vue des documentalistes, il pouvait aussi y avoir un souhait de garder la main sur des outils maison, plus ou moins ad hoc, et de ne se voir imposer quelque chose par d'autres métiers, avec lesquels il n'y a pas de partage de l'activité, ni aux niveaux technique et culturel, et de se sentir dessaisir. Donc comme dans toute organisation, la tension elle peut être entre les services vus comme plus informatiques et les services plus de documentation en l'occurrence. Et là la difficulté est que les services informatiques construisent les outils de travail. On imagine bien que le rapport de force, le rapport de pouvoir est assez dissymétrique.

AC : Et le rapport entre les informaticiens-développeurs de la base et les ontologues ?
ED : Oui bien sûr les ontologues rencontrent des difficultés en allant exposer devant les équipes techniques d'une organisation, qui comme dans le cas d'Orphanet, a déjà plus de quinze ans d'histoire, si on la prend au sens stricte car c'est plus en fait. En exposant qu'en fait, la seule solution pour régler les différents problèmes rencontrés, serait de complètement revenir sur la base de votre infrastructure pour envisager d'y parvenir, cela peut être difficile à entendre, à croire. On n'imagine pas une seule situation de travail, qu'elle soit informatisée pas, dans laquelle cela ne poserait pas des problèmes organisationnels et subjectifs, des problèmes du sens du travail pour les individus. Cela peut être perçu comme une remise en cause, un constat d'incapacité des équipes en place à régler le problème sans une intervention extérieure, peut-être parachutée. Dans le cadre de l'enquête ce n'est pas ce que l'on a spécifiquement étudié, mais c'est sûr que c'était un peu difficile ou pouvait être délicat.

AC : La présentation de la solution ontologique n'a pas été pensée, réfléchie, anticipée, de façon à ce que le collectif d'Orphanet y adhère de façon plus positive, alors qu'il y avait eu un intérêt pour le sujet, un apriori plutôt positif ? Au final, qui était sponsor du projet, hormis les ontologues forcément promoteur du projet ?
ED : Au final les ontologues. Leur posture, qui pourrait être perçue comme celle d'arriver avec une solution et une seule, n'est pas forcément souhaitable pour tout le monde. Il peut y avoir une crainte de se voir piquer son activité ou d'être redevable d'un nouveau type d'acteurs qu'on méconnaît le plus souvent. Mais au le début, les équipes techniques informatiques et notamment le consultant externe, ont eu la vision que c'était la solution, clairement. C'est difficile pour un informaticien développeur de ne pas s'intéresser aujourd'hui aux ontologies. Mais du point de vue de la sociologie des organisations, ce qui peut être un point de blocage, c'est quand il y a autant de difficultés et autant d'acteurs qui ont des perspectives différentes. Mais cela c'est le cas des réformes

informatiques dans des tas d'autres organisations. C'est la critique qui a été faite par la sociologie des organisations qui étudie l'informatisation à propos des modèles de changement organisationnels dits de type Big bang. Ces modèles où on était à un état 1 et on passe à un état 2 qui s'appuierait très peu sur le précédent. Et bien souvent, ce que la sociologie des organisations a montré, c'est que cela conduit dans quasiment ou au moins un cas sur deux à des échecs. Pour illustrer par un exemple, dans une grande société des télécommunications, on a investit énormément pour changer un système de paie et à un moment on a évalué que cela risquait de ne pas marcher et donc on a abandonné le projet.
Et même si pour les documentalistes qui attendaient et plaidaient pour un changement, une transformation, une adaptation de leurs outils de travail, peut-être que la manière de présenter la transformation était aussi inquiétante que rassurante, et difficile à saisir.
Orphanet c'est le produit d'une histoire. Les héritages du passé, comme les héritages dans les bases de données, pèsent toujours plus fortement qu'on ne le pense.
Pour les sociologues de l'informatique ou pour les chercheurs qui s'intéressent aux ontologies, les problèmes peuvent venir de l'affirmation que régler le problème du langage et ses ambiguïtés potentielles solutionnerait tous les problèmes de relations, de communication, c'est-à-dire venir du solutionnisme technologique. D'ailleurs, la directrice d'Orphanet, avait d'un côté un intérêt fort pour la technique. Elle était soucieuse d'effectuer une veille minimale sur le sujet. Ce qui lui permettait d'aller solliciter et d'un peu piquer ses équipes. Mais en même temps elle insistait sur une difficulté essentielle, qui n'était pas de dire que la solution était complètement technique, mais pour pointer le manque de culture technique partagée au sein des équipes. C'était aussi une manière de dire que du côté des documentalistes, il n'y avait pas une culture des bases de données relationnelles. Cette histoire de la base de données relationnelle qui traverse celle d'Orphanet, l'histoire de sa directrice et initiatrice, est-ce une culture des ontologies ? Quels sont les points communs entre une culture des bases de données relationnelles et une culture des ontologies ?
L'important pour une organisation comme Orphanet était aussi de pouvoir afficher que la piste ontologie était envisagée. En plus d'autres organisations se lançaient dans ce type de projet à l'époque. Mais, du coup, les freins, les difficultés pouvaient aussi venir du fait qu'il fallait conduire ces opérations en rapport avec l'extérieur. Cela n'était pas conduit de façon autonome. Il y avait des interactions avec l'équipe du Cismef à Rouen et d'autres institutions avec lesquelles il fallait mener des discussions.

AC : Savez-vous comment ce projet et ces discussions avec d'autres organisations ont été menés ?
ED : Non, cela ne faisait pas partie de nos observations. Il s'agissait vraiment d'une ethnographie d'Orphanet et pas de suivre ce projet. Donc cela pouvait apparaître lors de réunion mais je n'ai que très peu d'éléments à ce sujet.

AC : Concernant les aspects juridiques, avez-vous souvenir du moment où les questions du droit d'utilisation, et pour le modèle ontologique et pour la classification des maladies rares qu'ils avaient élaborée, ont pu faire leur apparition dans les interactions auxquelles vous avez pu assister ?
ED : Non je ne peux pas répondre du moment. Mais ce sont des aspects qu'ils abordaient.
Déjà l'idée que l'ensemble des données de la base appartenaient à l'Inserm n'était pas rien. Qu'il y eu des sessions de droit pour des activités d'exploitation tierces, comme l'exploitation par vidéodisques, ou par d'autres sites web exploitant des photos, notamment celui qui s'appelait POSSUMWeb, à la fin des années 80 et au début des années 90. Ça, ça a été des questions juridiques. Il fallait solliciter l'Inserm sur ces questions-là, car les compétences n'étaient complètement acquises en interne à Orphanet, bien qu'une personne soit référente sur ces questions. Ces aspects ont été exacerbés avec l'expansion, la diversification de l'offre qui touchait à des services. Comme notamment des extractions de la base non plus seulement pour des motifs de recherche mais aussi industriels. Et là aussi la question de la monétisation de l'information car on n'est pas

Entretien avec Eric Dagiral à propos du collectif du portail Orphanet 28 août 2014
Maître de conférences en sociologie à l'université Paris Descartes
Titulaire de la Chaire CNRS-Université "Technologie et lien social"
http://recherche.parisdescartes.fr/CERLIS/Equipe/Membres-statutaires/Dagiral-Eric

nécessairement tout le temps dans la connaissance scientifique avec différentes formes d'open access. Liés à ces aspects, il y avait aussi ceux de la dimension militante des acteurs associatifs et des nécessités de financement et le souci de pérennisation d'une structure qui l'est peu finalement. Et bien sûr les éléments éthiques n'étaient pas nuls. Se posait la question sur ce qu'on pouvait faire des données recueillies et des conséquences de leur interfaçage, et d'évaluer ce qui pourrait être des données personnelles qui ramenait à des considérations en lien avec la CNIL. Un des exemples était les difficultés pour les parties des bases pour la mise en relation de malades. Car pour cela il y a nécessairement des données personnelles. La mise en relation de malades faisait donc partie des projets d'Orphanet qu'il était le moins facile de rendre visible. On peut donc faire un constat de sciences sociales : la technique n'est pas purement technique. C'est une imbrication de social et de technique.