UFR SHA

Mention Information-Communication

Spécialité Documentation

Année universitaire 2017-2018

L'exploitabilité des systèmes d'organisation des connaissances dans le web des données pour améliorer la recherche d'information

Mémoire pour l'obtention du Master esDOC

Présenté par Carole Benoit

Le 19 septembre 2018

Sous la direction de

Monsieur David Guillemin

Université de Poitiers

UFR SHA

Mention Information-Communication

Spécialité Documentation

Année universitaire 2017-2018

L'exploitabilité des systèmes d'organisation des connaissances dans le web des données pour améliorer la recherche d'information

Mémoire pour l'obtention du Master esDOC

Présenté par Carole Benoit

Le 19 septembre 2018

Sous la direction de

Monsieur David Guillemin

Université de Poitiers

Remerciements

Mes remerciements vont en premier lieu à mon tuteur de mémoire, monsieur David Guillemin pour sa disponibilité et ses conseils avisés.

Je tiens également à remercier monsieur Miled Rousset de la Maison de l'Orient et de la Méditerranée Jean-Pouilloux, madame Isabelle Donze de la bibliothèque Eric-de-Dampierre et monsieur Thierry Bouchet du Centre national Rameau, d'avoir accepté de répondre à mes questions.

Ma gratitude va également à mon tuteur de stage, monsieur Etienne Marchand, et à madame Sandrine Clerc de l'Office International de l'Eau pour leur compréhension et pour m'avoir permis de prendre sur mon temps de stage afin de finaliser ce mémoire.

Enfin, un dernier remerciement, mais non des moindres, à l'ensemble de la promotion 2017-2018 des M2 du master esDOC, pour sa solidarité et la bonne humeur qui y règne.

Table des abréviations

Afnor : Association française de normalisation

BNE : *Biblioteca Nacional de España* (Bibliothèque nationale d'Espagne)

BNF : Bibliothèque nationale de France

CLIR : *Council on Library and Information Resources*

CNRS : Centre national de la recherche scientifique

Enssib : École nationale supérieure des sciences de l'information et des bibliothèques

FAO : *Food and Agriculture Organization of the United Nations* (Organisation des Nations unies pour l'alimentation et l'agriculture)

Gemet : *GEneral Multilingual Environmental Thesaurus* (thésaurus environnemental multilingue et général)

HTML : *HyperText Markup Language* (langage de marquage hypertexte)

HTTP: *HyperText Transfer Protocol* (protocole de transfert hypertexte)

IDC : *International Data Corporation*

Ifla : *International Federation of Library Associations and Institutions* (fédération internationale des associations et institutions de bibliothèques)

Inist : Institut de l'information scientifique et technique

Isidore : Interconnexion de services et interopérabilité des données pour la recherche et l'enseignement

Isko : *International Society for Knowledge Organization* (société internationale pour l'organisation des connaissances)

Iso : *International Organization for Standardization* (organisation internationale de normalisation)

LCSH : *Library of Congress Subject Headings* (liste de vedettes-matière de la bibliothèque du Congrès)

LRM : *Library Reference Model* (modèle de référence des bibliothèques)

OAI-PMH : *Open Archives Initiative Protocol for Metadata Harvesting* (protocole pour la collecte de métadonnées de l'Initiative pour les Archives ouvertes)

OWL : *Web Ontology Language* (langage d'ontologie web)

Pactols : Peuples et cultures, Anthroponymes, Chronologie, Toponymes, Lieux et Sujets

Rameau : Répertoire d'autorité-matière encyclopédique et alphabétique unifié

RDF : *Ressource Description Framework* (modèle de description de ressource)

RDFa : *Resource Description Framework* dans des Attributs

RDF-S : *RDF Schema*

RIF : *Rule Interchange Format* (règle de format d'échange)

RSS : *Really Simple Syndication* (syndication très simple)

Skos : *Simple knowledge organization system* (système simple d'organisation des connaissances)

Skos XL : *SKOS eXtension for Label* (extension de Skos pour les labels)

Soc : système d'organisation des connaissances (en anglais « KOS » *Knowledge Organization System*)

Sparql : *Sparql Protocol and RDF Query Language* (protocole Sparql et langage de requête RDF)

TED : *Technology, Entertainment and Design* (technologie, divertissement et design)

TGE : très grand équipement

TGIR : très grande infrastructure de recherche

Uri : *Uniform Resource Identifier* (identifiant uniforme de ressource)

Url : *Uniform Resource Locator* (localisateur uniforme de ressource)

XML : *Extensible Markup Language* (langage de balisage extensible)

W3C : *World Wide Web Consortium* (consortium du web)

Sommaire

Remerciements

Table des abréviations

Introduction

Partie 1 : état de l'art

I. La recherche d'information
II. Du web sémantique au web des données
III. Les systèmes d'organisation des connaissances dans le web des données

Partie 2 : méthodologie de l'étude de cas

I. Présentation de la plate-forme de recherche Isidore
II. L'analyse des enrichissements sémantiques
III. Les entretiens

Partie 3 : étude de cas : présentation des résultats de l'observation

I. Créer du lien entre les données : rebonds et alignements
II. Les doublons
III. La compréhension de la sémantique

Conclusion

Bibliographie

Bibliographie et webographie de l'état de l'art
Bibliographie et webographie de l'étude de cas

Table des illustrations

Table des annexes

Introduction

La démocratisation du web avec le développement des moteurs de recherche en requête libre sur le texte intégral des documents a un temps fait craindre l'obsolescence des langages documentaires. En effet, la formalisation des termes d'une recherche d'information en langage naturel est beaucoup plus intuitive pour un usager que le recours à un langage documentaire possédant une syntaxe et une logique propres, dont la maîtrise est souvent l'apanage des professionnels de l'information et de la documentation. Toutefois, si le langage naturel est plus immédiatement accessible à l'humain, il n'en va pas de même pour les machines, les systèmes de recherche d'information ayant besoin d'aide, de formalisme, pour être en mesure d'exploiter ce langage et notamment de passer outre ses ambiguïtés ; or, les vocabulaires contrôlés dans lesquels s'inscrivent les langages documentaires ont justement été conçus pour lutter contre ce problème.

En effet, dans un contexte de surabondance de l'information disponible sur le web ces Soc (systèmes d'organisation des connaissances) ont rencontré un regain d'intérêt pour lutter contre le chaos documentaire et faciliter la recherche d'information. Manuel Zacklad et Alain Giboin en proposent la définition suivante :

> « Les « systèmes d'organisation des connaissances » visent donc à définir des principes de description d'un domaine pour faciliter les opérations de classement et de recherche « d'items » plus ou moins abstraits : documents, personnes, lieux, produits, opinions ou activités. »[1]

Les systèmes d'organisation des connaissances réunissent donc différentes familles de modèles ayant vocation à formaliser et structurer des informations en vue de faciliter leur accès et leur recherche. Cette notion est très vaste et nous aurons l'occasion de constater au cours de ce mémoire que son périmètre ne fait pas l'unanimité dans la communauté scientifique, bien qu'un consensus semble exister sur le fait que les trois principales familles de langages documentaires : schémas de classification, listes de vedettes-matière et thésaurus, en fassent partie. L'exemple des thésaurus est d'ailleurs particulièrement éloquent lorsque nous considérons que, contrairement aux deux autres Soc mentionnés, leur

1 ZACKLAD, Manuel et GIBOIN, Alain. Introduction. Systèmes d'organisation des connaissances hétérogènes pour les applications documentaires, *Document numérique*. 2010/2, Vol. 13, p.8. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-document-numerique-2010-2-page-7.htm

apparition est concomitante à celle de l'informatique documentaire et qu'ils ont récemment dû faire face au changement de paradigme instauré par la norme Iso 25964 « Thésaurus et interopérabilité avec les autres vocabulaires », dont la première partie est parue en 2011 et la seconde en 2013, afin de pouvoir mieux s'insérer dans les logiques de recherche d'information sur le web ainsi que sur le web des données.

Le projet du « web des données » (« *linked data* ») tel que promu par Tim Berners-Lee est intrinsèque à celui du web sémantique et repose sur l'idée de création de liens logiques entre les différentes données présentes sur le web afin de pouvoir naviguer de l'une à l'autre et les exploiter conjointement. Cette initiative est portée par les technologies dites « du web sémantique », un ensemble de recommandations du W3C dont la vocation est de permettre aux machines d'accéder au sens (la sémantique) des termes du langage naturel afin de permettre la création de liens entre des données éparses et ainsi assurer une meilleure exploitation de celles-ci par les systèmes informatiques. Les systèmes d'organisation des connaissances ont été identifiés comme pouvant avoir un rôle à jouer dans la mise en œuvre de ce projet en servant de points de jonctions entre des ressources issues de silos de données épars, comme le prouve la création de Skos (*Simple Knowledge Organization System*), un format reposant sur RDF (*Resource Description Framework*), le modèle de données de base du web sémantique, et permettant la représentation, la publication et l'exploitation de Soc dans le web des données.

C'est donc dans ce contexte que ce mémoire pose la question de l'influence de l'ouverture des systèmes d'organisation des connaissances au web des données sur l'efficacité de la recherche d'information. Le sujet étant vaste, nous avons amorcé notre réflexion à partir de deux grandes hypothèses : la première est que l'enjeu du web des données pour les systèmes d'organisation des connaissances est avant tout celui de l'interopérabilité, et la seconde est que les systèmes d'organisation des connaissances compatibles avec le web des données ne sont pas forcément mieux exploitables pour les utilisateurs lorsqu'ils recherchent de l'information.

Afin d'arriver à un équilibre entre les aspects théoriques cadrant les notions de recherche d'information, de web des données et de systèmes d'organisation des connaissances, et un

exemple concret d'utilisation de Soc inscrits dans le web des données, le présent mémoire est divisé en trois parties. La première est un état de l'art prenant appui sur la littérature scientifique et professionnelle afin de définir et contextualiser les trois notions sus-citées, la deuxième présente la méthodologie mise en place afin de mener à bien l'étude d'un cas concret, en l'occurrence la plate-forme de recherche en sciences humaines « Isidore », présentant la particularité d'exploiter plusieurs systèmes d'organisation des connaissances afin d'enrichir des notices documentaires et de les interconnecter selon la logique du web des données, enfin, les résultats et les analyses de cette étude de cas sont présentés dans une troisième partie.

Partie 1 : état de l'art

Le présent état de l'art est divisé en trois parties, chacune se focalisant sur une notion introduite par la problématique alimentant ce mémoire.

Ainsi, un premier temps est consacrée à la recherche d'information afin de délimiter les contours de cette discipline avant de présenter le principe et le fonctionnement des systèmes dits « de recherche d'information », un second temps est dédié aux notions de « web sémantique » et de « web des données » afin de mettre en exergue les nuances entre ces deux idées ainsi que leur contexte de développement, enfin, la dernière partie porte sur les systèmes d'organisation des connaissances dans le web de données, le format Skos, la norme Iso 25964 et les changements de paradigme induits.

I. La recherche d'information

I.I La recherche d'information : entre sciences de l'information et informatique

A la fois activité humaine quotidienne, pratique professionnelle et objet d'étude pluridisciplinaire, la notion de « recherche d'information » est particulièrement polysémique[2]. En effet, l'Afnor (Association française de normalisation), telle que citée par Jean-Philippe Accart et Alexis Rivier, la définit de façon générique :

> « actions, méthodes et procédures ayant pour objet d'extraire d'un ensemble de documents les informations voulues. Toute opération (ou ensemble d'opérations) ayant pour objet la recherche, la collecte et l'exploitation en réponse à une question sur un sujet précis »[3]

En revanche, Yves Chiaramella et Philippe Mulhem en donnent une définition davantage centrée sur la discipline scientifique :

> « Dans sa définition la plus large, la RI [recherche d'information, NDLR] a pour thème central l'étude de modèles et systèmes d'interaction entre des utilisateurs humains et des corpus de documents numériques, en vue de la satisfaction de leurs besoins d'information »[4]

2 DINET, Jérôme. *La recherche d'information dans les environnements numériques*. Londres : ISTE Editions, 2014. p.8. Collection systèmes d'information, web et informatique ubiquitaire. ISBN 978-1-78405-018-4

3 ACCART, Jean-Philippe et RIVIER, Alexis. *Mémento de l'information numérique*. Paris : Éditions du Cercle de la Librairie, 2012. p.99. Collection Bibliothèques. ISBN 978-2-7654-1332-5

4 CHIARAMELLA, Yves et MULHEM, Philippe. La recherche d'information. De la documentation automatique à la recherche d'information en contexte, *Document numérique*. 2007/1, Vol.10, p.12. Également disponible en ligne à l'adresse :

Ces deux définitions s'accordent sur le fait que la recherche d'information naît du besoin humain d'accéder à des informations contenues dans des ressources de nature diverse et qu'il s'agit donc à ce titre d'une notion très large. Néanmoins, la seconde définition introduit l'idée que la recherche d'information en tant qu'objet d'étude scientifique n'existe que dans le contexte du numérique et en particulier de l'interaction homme-machine. Bien que cette définition soit trop restrictive, nous verrons que cette discipline ne se limite pas au domaine informatique, elle a le mérite de délimiter l'objet d'étude ; en outre dans la mesure où le présent mémoire porte sur la recherche d'information dans un environnement numérique, c'est donc sur cette seconde définition que nous nous appuierons en priorité.

L'apparition de la recherche d'information en tant qu'objet d'étude scientifique est concomitante à celle des premiers ordinateurs. En effet, dès les années 1940 de nombreux chercheurs voyaient dans les super calculateurs des outils pour « aider puissamment à appréhender l'univers en rapide croissance des connaissances humaines ».[5] A ce titre Vannevar Bush apparaît dès 1945 comme l'un des précurseurs de ce questionnement avec le concept de « memex », un appareil électronique conceptuel capable d'effectuer des recherches automatiques dans une vaste bibliothèque de documents servant d'extension à la mémoire humaine.[6]

En dépit de son contexte de naissance, la recherche d'information est généralement rattachée aux sciences de l'information ; toutefois d'autres disciplines questionnent et analysent cet objet. Parmi elles, Jérôme Dinet évoque en plus de l'informatique, l'économie et la robotique.[7] Selon l'approche économique, la recherche d'information est un outil d'aide à la prise de décision. Plusieurs études menées par des organismes économiques estiment que bien que de plus en plus de temps de travail soit consacré dans les entreprises à la recherche d'information, 50 à 90 % des documents produits par les organismes sont déjà existants ailleurs. Ces études soulignent également que l'activité de recherche d'information représente un coût financier, humain et matériel important, et qu'il est donc important de bien la maîtriser afin d'être en mesure de tirer de la plus-value de l'analyse des informations trouvées, plutôt que de s'adonner à l'activité improductive de reproduction d'informations déjà existantes. Notons tout de même que les études parlent d'informations « déjà

https://www.cairn.info/revue-document-numerique-2007-1-page-11.htm

5 *Ibid*.

6 *Ibid*.

7 DINET, Jérôme. *op.cit*. pp.14-19.

existantes » mais sans préciser leur niveau de mise à disposition : les informations reproduites étaient-elles trouvables sur le web ? dans le système d'information de l'entreprise ? étaient-elles librement accessibles ou en accès restreint ? le chercheur d'information aurait-il pu avoir accès à l'information d'une manière ou d'une autre ? Ces questions mériteraient d'être également prises en considération. En effet, dans certains contextes il peut être plus rapide, si ce n'est nécessaire, de reproduire une information déjà existante car elle n'est pas facilement voire pas du tout accessible autrement au chercheur d'information. Les résultats de ces études sont donc à prendre avec précaution. En outre, Jérôme Dinet cite parmi les entreprises ayant mené les études sus-citées Google et IDC (*International Data Corporation*), deux acteurs des technologies de l'information ; il n'est donc pas étonnant que les résultats des études montrent un réel besoin de perfectionnement de la recherche d'information dans les entreprises.

Concernant les apports de la recherche d'information dans la robotique, ceux-ci se manifestent surtout par l'amélioration des déplacements des robots puisqu'ils ont besoin de rechercher et de prélever des informations dans leur environnement afin de pouvoir se déplacer de façon autonome. Dans ce contexte les travaux en recherche d'information vont également puiser dans la psychologie et l'éthologie afin de permettre l'interprétation et l'acquisition de comportements humains, et permettre ainsi aux robots de se déplacer mais également d'adapter leurs actes à la présence ou non d'êtres humains à proximité.

Ces deux exemples confirment que la recherche d'information est un objet d'étude faisant sens dans de nombreux domaines scientifiques.

I.II Des recherches d'information

L'emploi du terme « recherche d'information » est quelque peu abusif. En effet, comme nous l'avons vu précédemment les différentes définitions de cette notion sont souvent vagues car sous l'expression générique « recherche d'information » se cachent différents objets plus spécifiques. En effet l'expression « recherche d'information » est notamment la traduction de l'anglais « *information retrieval* »[8], or là où la littérature française emploi presque exclusivement le terme « recherche d'information », les anglo-saxon disposent d'un vocabulaire plus nuancé. Ainsi, « *information retrieval* » renvoie davantage à la recherche

8 CHIARAMELLA, Yves et MULHEM, Philippe. *op.cit.* p.12

d'information en tant qu'objet d'étude scientifique multidisciplinaire[9] ou encore à la recherche d'un document, ou fragment de document, numérisé dont l'existence est avérée mais l'emplacement inconnu[10], tandis que « *information seeking* » renvoie davantage à la « recherche ouverte d'information » c'est-à-dire au processus d'identification d'informations existantes sur un sujet donné[11]. Or, contrairement à une recherche d'information que l'on pourrait qualifier de « classique », renvoyant donc à l'*information retrieval*, l'utilisateur entreprenant une recherche ouverte d'information n'a aucune certitude quant à l'existence de ressources pouvant répondre à son besoin d'information.[12] Par conséquent la recherche ouverte d'information est un processus au cours duquel le système informatique utilisé pour la recherche va accompagner l'utilisateur dans la formalisation de son besoin d'information par l'identification des informations existantes et de celles qui ne le sont pas.[13]

Néanmoins la définition que nous donnons ici de la « recherche ouverte d'information » n'est pas toujours équivalente à celle de « *information seeking* ». En effet, si pour certains auteurs l'*information seeking* est bien un processus de recherche d'information, pour d'autres il s'agit d'un processus d'acquisition d'information, d'autres encore la considèrent comme un processus global, de recherche et d'acquisition d'information, tandis que pour d'autres il s'agit d'un processus global permettant de créer des connaissances à partir d'informations données.[14] Si les différences entre ces définitions peuvent paraître minimes elles induisent tout de même des nuances impactant l'objet d'étude.

En outre il peut également être intéressant d'expliquer pourquoi est-ce que nous parlons ici de « recherche d'information » et non pas de « recherche documentaire ». En effet, la

9 DINET, Jérôme. *op. cit.* p.9

10 ZACKLAD, Manuel. Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (ROI). In : ARSENAULT, Clément et DALKIR, Kimiz. (dir.). CAIS/ACSI 2007, *Actes du 35e Congrès annuel de l'Association Canadienne des Sciences de l'Information. Partage de l'information dans un monde fragmenté : Franchir les frontières*. Montréal, 2007. p.2. Également disponible en ligne à l'adresse : https://archivesic.ccsd.cnrs.fr/sic_00202440/document

11 DESFRICHES DORIA, Orélie et ZACKLAD, Manuel. Améliorer la recherche d'information à l'aide de thésaurus « ad hoc ». Expérimentations et réflexions méthodologiques, *Document numérique*. 2010/2, Vol.13, p.17. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-document-numerique-2010-2-page-13.htm

12 ZACKLAD, Manuel. 2007. *op.cit.* p.2

13 *Ibid*.

14 IKOJA-ODONGO, Robert et MOSTERT, Janneke. Information seeking behaviour: A conceptual framework, *South African Journal of Libraries and Information Science*. 2006/3, Vol.72, p.148. Également disponible en ligne à l'adresse : http://sajlis.journals.ac.za/pub/article/view/1112

définition même de la notion de « recherche documentaire » ne fait pas l'unanimité dans la mesure où, par exemple, Jean-Philippe Accart et Alexis Rivier la définissent comme étant propre à la profession de documentaliste[15], tandis que Jacques Maniez adopte une définition plus globale en partant de l'objet d'une recherche documentaire, en l'occurrence « un lot sélectionné de sources d'information spécifique »[16] mais ne précise pas si les sources d'informations retenues ont été recherchées et/ou triées par un professionnel ou directement par le chercheur d'information. Enfin, le dictionnaire de l'Enssib (École nationale supérieure des sciences de l'information et des bibliothèques) souligne les évolutions qu'a connu la notion, précisant qu'il s'agissait initialement d'un acte mené par des professionnels mais considère qu'il s'agit désormais d'un synonyme de « recherche d'information », les évolutions technologiques ayant permis aux individus de s'émanciper des documentalistes pour mener à bien leurs recherches documentaires, le dictionnaire de l'Enssib attribue donc la même entrée aux deux termes.[17] Or, à une époque où les évolutions de l'informatique et la multiplication des corpus informationnels directement accessibles par l'utilisateur font que l'intermédiation du professionnel de l'information est effectivement de plus en plus rare,[18] il pourrait être plus judicieux de considérer la recherche documentaire avec la définition de Michel Beigbeder,[19] c'est-à-dire la recherche de document, physique ou numérique, tandis que la recherche d'information ne serait quant à elle pas liée à la recherche d'un support précis mais d'un contenu sur un sujet donné, se rapprochant ainsi de la définition que nous avons évoqué précédemment de la « recherche ouverte d'information ». Par conséquent, désormais lorsque nous parlerons de « recherche d'information » il faudra comprendre cette notion avec le sens de « recherche ouverte d'information ».

15 ACCART, Jean-Philippe et RIVIER, Alexis. *op.cit.* p.99

16 MANIEZ, Jacques. *Actualité des langages documentaires : fondements théoriques de la recherche d'information.* Paris : ADBS Edition, 2002. p.118. Collection Sciences de l'Information Série Études et techniques. ISBN 2-84365-060-7

17 DEROCHE, Frédéric. Recherche d'information (recherche documentaire). *Enssib* [en ligne]. Mise à jour le 19 août 2015. [Consulté le 22 avril 2018]. Disponible à l'adresse : http://www.enssib.fr/le-dictionnaire/recherche-dinformation-recherche-documentaire

18 CHIARAMELLA, Yves et MULHEM, Philippe. *op.cit.* p.13

19 BEIGBEDER, Michel. Les temps du document et la recherche d'information. *Document numérique*, 2004/4. Vol.8. p.55. Également disponible en ligne à l'adresse : https://www.cairn.info/article.php?ID_ARTICLE=DN_084_0055

I.III Les systèmes de recherche d'information

La recherche d'information relève d'un processus complexe pouvant être représenté schématiquement sous la forme d'un « U ».[20] Cette figure montre que le processus de recherche d'information se joue à deux niveaux : celui de l'utilisateur demandeur d'information et celui du système informatique fournisseur d'information. En effet, afin de combler son besoin informationnel, l'individu doit tout d'abord le comprendre afin de pouvoir formuler une requête (phase d'indexation), l'exécuter, analyser ses résultats, extraire les informations pertinentes puis éventuellement formuler une nouvelle requête[21] tandis que le système informatique interrogé doit être en mesure de fournir une sélection de documents adaptée à la requête de l'utilisateur. Le processus de recherche d'information consiste donc en la mise en relation d'un demandeur d'information (l'usager) avec un fournisseur d'information (entreprise, bibliothèque, etc.) par le biais d'un système informatique permettant de rechercher et consulter tout ou partie de documents.[22]

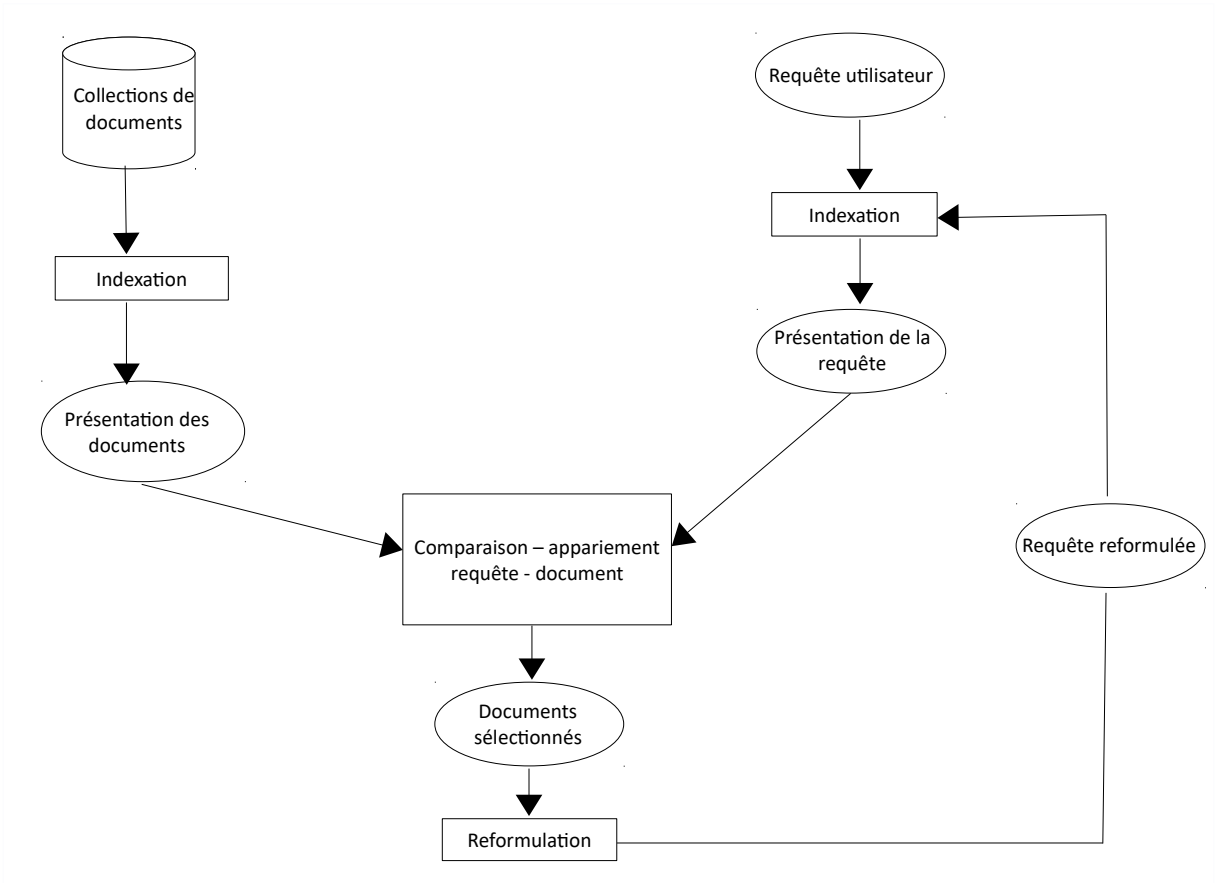


Illustration 1: processus en « U » de recherche d'information Source : BOUGHANEM, Mohand. *op. cit.*

20 BOUGHANEM, Mohand. Chapitre 1. Introduction à la recherche d'information. In : BOUGHANEM, Mohand et SAVOY, Jacques. *Recherche d'information : état des lieux et perspectives*. Paris : Lavoisier, 2008. p.22. Collection Recherche d'information et web. ISBN 978-2-7462-2005-8

21 IKOJA-ODONGO, Robert et MOSTERT, Janneke. *op.cit.* p.148

22 DESFRICHES DORIA, Orélie et ZACKLAD, Manuel. *op.cit.* p.16

Cependant afin d'être en mesure de mener à bien ce processus de recherche d'information Boubée et Tricot remarquent que le chercheur d'information doit s'appuyer sur trois types de connaissances : sur le contenu recherché, puisqu'il est impossible de faire une recherche sur un objet dont on ne sait absolument rien, sur le domaine de l'information, c'est-à-dire être capable de mobiliser les ressources les plus adaptées à son besoin, et enfin sur le fonctionnement des documents électroniques et des systèmes d'informations afin d'être en mesure de les exploiter.[23] Pourtant plusieurs études consacrées aux usages des moteurs de recherche tendent à montrer que les stratégies de recherche mobilisées sont rudimentaires puisque les utilisateurs privilégient des requêtes courtes formulées en langage naturel et limitent leur lecture des résultats aux premiers affichés.[24] La recherche d'information en tant que discipline scientifique (*information retrieval*) ayant pour vocation première de permettre la création de ses interfaces et systèmes informatiques appelés « systèmes de recherche d'information », cela conduit Boubée et Tricot à associer les travaux relatifs à l'*information retrieval* à ceux sur l'organisation des connaissances,[25] notion que nous détaillerons davantage dans le troisième chapitre de cet état de l'art.

Les systèmes de recherche d'information permettent un accès rapide et facilité à l'information pertinente en donnant accès à un corpus documentaire constitué à partir des correspondances entre la requête formulée par l'utilisateur et les documents présents dans l'index du système de recherche.[26] Comme son nom l'indique, cet index, parfois aussi appelé « dictionnaire », est le résultat d'une phase d'indexation c'est-à-dire de description du contenu des documents présents dans le système d'information. Il est donc constitué à partir d'une liste de termes significatifs, pouvant être issus d'un vocabulaire contrôlé, bien que cela ne soit pas toujours le cas, afin de simplifier et normaliser l'indexation[27] c'est-à-dire limiter la

23 BOUBEE, Nicole et TRICOT, André. *Qu'est-ce que rechercher de l'information ?* Villeurbane : Presses de l'ENSSIB, 2010. 286 p. Également disponible en ligne à l'adresse : http://books.openedition.org/pressesenssib/799

24 MENON, Bruno. Journée d'étude ADBS. Optimiser l'accès à l'information, une opportunité pour les langages documentaires ?, *Documentaliste-Sciences de l'Information*. 2007(a)/6, Vol.44, p.385. Également disponible en ligne à d'adresse : https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-6-page-385.htm

25 BOUBEE, Nicole et TRICOT, André. *op.cit.*

26 ZARGAYOUNA, Haïfa, ROUSSEY, Catherine et CHEVALLET, Jean-Pierre. Recherche d'information sémantique : état des lieux, *Traitement automatique des langues*. 2015/3, Vol.56. p.50. Également disponible en ligne à l'adresse : http://www.atala.org/sites/default/files/2.Zargayouna-56-3.pdf

27 MANIEZ, Jacques. *op.cit.* p.169

part de subjectivité en imposant l'utilisation d'une liste de mots prédéterminés. A chacun de ces termes sont associés des poids afin de définir leur degré de représentativité ; ce sont donc les termes avec le plus haut degré de représentativité qui sont identifiés par le système de recherche d'information comme étant les plus pertinents et ce sont donc les documents qui leur sont associés qui apparaissent en premier dans les résultats de recherche lorsque le système identifie une similarité entre la requête et un ou plusieurs termes de l'index.[28]

Mohand Boughanem et Jérôme Dinet s'accordent pour dire que la notion de « pertinence » est au cœur de tout système de recherche d'information dans la mesure où c'est c'est le degré de pertinence qui permet d'établir le lien de correspondance entre une requête et un document. Par conséquent cette notion permet également d'évaluer l'efficacité d'un système de recherche d'information grâce au calcul du taux de rappel et du taux de précision.[29] En effet, le taux de rappel correspond au nombre de documents pertinents retournés par rapport au nombre total de documents pertinents disponibles, tandis que le taux de précision renvoie au nombre de documents pertinents retournés par rapport au nombre total de documents retournés. Ainsi, plus un système de recherche d'information est efficace plus ses taux de rappel et de précision sont élevés. Par extension cela signifie également que le bruit, c'est-à-dire la présence dans les résultats de recherche de documents non pertinents, et le silence documentaire, c'est-à-dire l'absence de documents pertinents dans la page de résultat, sont faibles.[30]

La pertinence est établie par le système de recherche d'information à l'aide d'un cadre théorique mathématique appelé « modèle de recherche d'information ». Il existe trois grands courants de modèles de recherche d'information : les modèles basés sur la théorie des ensembles, les modèles algébriques et les modèles probabilistes. Les modèles basés sur la théorie des ensembles reposent sur la combinaison des termes d'une requête à l'aide des opérateurs logiques « *AND* » (ajout : les deux termes doivent être présents), « *OR* » (choix : l'un ou l'autre des termes doit être présent), « *NOT* » (exclusion : le terme ne doit pas être présent), l'illustration la plus connue des modèles basés sur la théorie des ensembles est le modèle booléen. Les modèles algébriques, comme par exemple le modèle vectoriel, permettent quant à eux de définir la pertinence d'un document pour une requête par des

28 BOUGHANEM, Mohand. *op.cit.* p.22.

29 *Ibid*. p.36

30 DELESTRE, Nicolas et MALANDAIN, Nicolas. *Du web des documents au web sémantique*. Bois-Guillaume : Éditions KLOG, 2017. pp.63-64. ISBN 979-10-92272-18-5.

mesures de distance dans un espace vectoriel, tandis que les modèles probabilistes permettent de calculer une probabilité de pertinence entre un document et une requête donnés.[31] En outre, afin d'assurer la meilleure pertinence possible entre une requête et les résultats proposés, certains systèmes d'information prennent en considération la signification des mots utilisés dans les requêtes et les ressources documentaires, ce sont les « systèmes de recherche d'information sémantique ». Ils fonctionnent grâce à des « ressources sémantiques » généralement générées sous la forme de thésaurus ou d'ontologie afin de formaliser informatiquement les différentes significations identifiées. Ainsi un système de recherche d'information sémantique est capable d'associer à un mot une ressource sémantique grâce aux techniques de traitement automatiques des langues ; cette étape, préalable à l'indexation, se nomme « annotation sémantique ». [32]

Cependant, si le système de recherche d'information calcule le degré de pertinence, c'est l'usager seul qui est en mesure de réellement juger de la pertinence ou non-pertinence d'une ressource[33]. En effet, au regard du système informatique les documents et les informations qu'ils contiennent sont simplement des données et c'est tout le contexte qui les entoure, comme leur contexte de production ou leur objectif d'utilisation, qui permet d'établir leur pertinence pour une requête donnée. Or, à ce jour, l'être humain demeure plus performant que la machine lorsqu'il s'agit d'interpréter ces données afin d'en établir la pertinence.[34] Par conséquent il s'agit bien d'une notion subjective dépendant du besoin d'information de l'usager, et donc de sa bonne expression dans une requête, des ressources informationnelles disponibles (document intégral, notice documentaire, métadonnées) et du contexte de l'activité de recherche d'information c'est-à-dire le cadre dans lequel elle est réalisée mais aussi les caractéristiques physiques et psychologiques de l'utilisateur : connaissance du sujet, familiarité avec le système de recherche, handicap éventuel, etc.[35]

Comme nous l'avons vu précédemment, la recherche d'information au sens d'*information retrieval* puise dans de nombreuses disciplines parmi lesquelles les sciences de l'information qui portent un intérêt particulier à l'utilisateur et à son double rôle de demandeur et

31 BOUGHANEM, Mohand. *op.cit.* pp.28-29
32 ZARGAYOUNA, Haïfa, ROUSSEY, Catherine et CHEVALLET, Jean-Pierre. *op.cit.* p.50
33 BOUGHANEM, Mohand. *op.cit.* p.22
34 DELESTRE, Nicolas et MALANDAIN, Nicolas. *op.cit.* pp.59-60
35 DINET, Jérôme. *op.cit.* pp.21-22

bénéficiaire du processus d'interaction avec les documents.[36] Cet intérêt se matérialise en particulier à travers l'axe de recherche dans cette discipline dit « document structuré » qui consiste à étudier différents moyens de fournir aux chercheurs d'informations non pas des documents dans leur intégralité mais plutôt uniquement les passages, ou unités documentaires, répondant à leur requête.[37] Or, ces travaux font surtout sens dans le contexte du web, et notamment des moteurs de recherche, afin d'éviter que les utilisateurs ne se perdent dans la surabondance informationnelle disponible. Néanmoins, cette multiplicité de l'information et de ses points d'accès ne sont pas les seuls freins à une recherche d'information efficace. En effet, en prenant l'exemple du contexte universitaire, Michael Keller explique que la multiplicité et le manque de précision des outils de recherche, mais aussi des schémas de métadonnées, et leur mauvaise indexation par les moteurs de recherche du web nuisent à une découverte et à un accès facile à l'information. Il propose donc le web sémantique comme solution au chaos informationnel[38], tout comme Jean-Philippe Accart et Alexis Rivier qui voient dans le web sémantique le web des professionnels des sciences de l'information.[39]

II. Du web sémantique au web des données

II.I Le projet initial : le web sémantique

Lorsque Tim Berners-Lee publie la feuille de route pour le déploiement du web sémantique en 1998 :

> « le Web sémantique est alors présenté comme une extension du Web des documents qui constituerait une base de données globale à l'échelle du réseau pour permettre aux machines de mieux appréhender les données et aux personnes de mieux coopérer ».[40]

En d'autres termes, le web sémantique apparaît comme une surcouche du web actuel permettant aux ordinateurs d'accéder au sens sémiotique des données, de la même manière qu'un être humain. Ce faisant les machines peuvent traiter et exploiter ces données de façon

36 CHIARAMELLA, Yves et MULHEM, Philippe. *op.cit.* pp.12-13

37 *Ibid.* p.25

38 KELLER, Michael. Linked Data: a way out of the information chaos and toward the semantic web. *EDUCAUSE* [en ligne]. 21 juillet 2011. [Consulté le 14 novembre 2017]. Disponible à l'adresse : https://er.educause.edu/articles/2011/7/linked-data-a-way-out-of-the-information-chaos-and-toward-the-semantic-web

39 ACCART, Jean-Philippe et RIVIER, Alexis. *op.cit.* p.121

40 POUPEAU, Gautier. Petite histoire du Web sémantique. *Les petites cases* [en ligne]. 15 août 2011. [Consulté le 21 novembre 2017]. Disponible à l'adresse : http://www.lespetitescases.net/petite-histoire-du-web-semantique

automatique. Cette définition est appuyée par Roger T. Pedauque qui présente le web sémantique comme une : « infrastructure enrichit et [qui, NDLR] s'appuie sur les standards actuels du Web et est susceptible d'en intégrer bien d'autres »[41]. En effet, si le web sémantique repose sur trois piliers du web traditionnel : les identifiants uniques Uri (ou URL), le protocole HTTP, le langage HTML, ainsi que sur le langage XML, il possède également des technologies propres[42]. En fait, le web sémantique est une infrastructure composée de six couches elles-mêmes divisées en un ou plusieurs blocs représentant ce que le W3C appelle « la pile du web sémantique ».

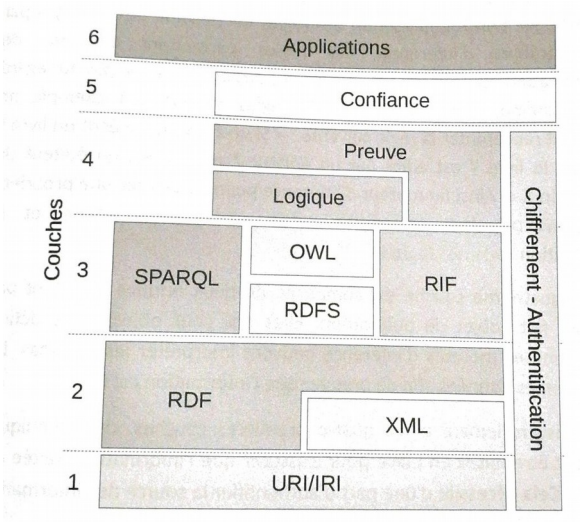


Illustration 2:pile du web sémantique

Source : DELESTRE, Nicolas et MALANDAIN, Nicolas. *op.cit*. p.169

A ce jour seules les trois premières couches de la pile sont effectives. Les couches 4 à 6 étant encore principalement prospectives ; la couche 4 permettrait de prouver la fiabilité des inférences établies par les moteurs d'inférence tandis que la couche 5 permettrait d'attribuer un indicateur de confiance aux informations traitées par les applications de la couche 6[43]. Par conséquent lorsque nous parlons des « technologies du web sémantique » il est généralement question du modèle RDF, un modèle de données propre exploitant les Uri sous

41 PEDAUQUE, T. Roger. *La redocumentarisation du monde*. Toulouse : Cépaduès, 2007. ISBN 978-2-85428-728-8. Chapitre 6 : Sur des aspects primordiaux du web sémantique, p. 100.

42 GANDON, Fabien, FARON-ZUCKER, Catherine et CORBY, Olivier. *Le web sémantique : comment lier les données et les schémas sur le web ?* Paris : Dunod, 2012. 206 p. ISBN 978-2-10-058140-5.

43 *Ibid*. pp.169-171

la forme de triplets exprimés avec le langage XML, ainsi que des langages SPARQL, RDF-S, OWL et RIF. Le premier permet l'interrogation des triplets RDF, les deux autres sont des langages de description et le dernier un langage de règles permettant la compatibilité avec des moteurs d'inférence[44]. Soulignons au passage la présence des Uri et du langage XML dans les deux premières couches, ce qui confirme l'idée que le web sémantique repose en grande partie sur des technologies web existantes et n'a pas vocation à supplanter le web des documents mais plutôt à exister en parallèle.

En outre il est intéressant de noter que même si seule la moitié des couches de la pile du web sémantique est pour l'instant utilisable, toutes les briques technologiques sont effectives, celles manquantes étant relatives à la question de la fiabilité des contenus des données. Nous pouvons d'ailleurs souligner que ces technologies sont surtout présentes dans des gisements de données collaboratifs dont la fiabilité repose sur les connaissances et le bon vouloir de milliers d'individus, comme GeoNames[45] ou Wikidata[46]. En effet, reprenant l'exemple de DBpedia[47], un autre gisement de données basé sur Wikipédia, Alexandre Monnin insiste sur le fait que « l'objectivité sur laquelle repose nos objets est bien une objectivité de second degré »[48] ; en d'autres termes les objets décrits sur le web à travers des données ne correspondent pas à la réalité du monde physique car ils dépendent des interprétations des êtres humains qui créent et documentent ces données. Néanmoins les technologies du web sémantique se sont également développées dans des institutions comme la BNF (Bibliothèque nationale de France) avec le portail data.bnf.fr[49]. Or, d'après Gautier Poupeau, la preuve et la confiance dans les données sont produites par des informations de contexte[50], par conséquent nous pouvons supposer que les données produites et partagées par de grandes institutions nationales ou internationales sont de qualité. Notons également la distinction entre la notion de « vérité », remise en question

44 *Ibid*. p.170

45 http://www.geonames.org/

46 https://www.wikidata.org/wiki/Wikidata:Main_Page

47 https://wiki.dbpedia.org/

48 MONNIN, Alexandre. Du cycle de vie des données au cycle de vie des objets. In : CALDERAN, Lisette, *et. al*. *Le document numérique à l'heure du web de données. Séminaire Inria Carnac, 1er-5 octobre 2012*. Paris : ADBS Éditions, 2012. p.225. Sciences et techniques de l'information. ISBN 978-2-84365-142-7.

49 http://data.bnf.fr/

50 POUPEAU, Gautier. Histoire(s) de notices. In : CALDERAN, Lisette, *et. al*. *Le document numérique à l'heure du web de données. Séminaire Inria Carnac, 1er-5 octobre 2012*. Paris : ADBS Éditions, 2012. p.37. Sciences et techniques de l'information. ISBN 978-2-84365-142-7.

pour les raisons sus-citées par Alexandre Monnin, et celle de « fiabilité » qui fait partie intégrante des objectifs du web sémantique. Ainsi, pour être considérées comme fiables les données n'ont pas besoin de se rattacher à des vérités objectives du monde physique, le monde virtuel du web n'étant pas une représentation du monde physique mais plutôt une construction parallèle.[51]

II.II La notion de « sémantique » : entre centralité et ambiguïté

L'enrichissement permis par le web sémantique tel que décrit par Roger T. Pedauque est principalement dû au fait que ces technologies permettent d'associer à n'importe quel contenu web, existant ou à venir, des informations sous la forme de métadonnées permettant l'exploitation automatique des ressources en question par des machines[52]. Cette exploitation étant rendue possible du fait que ces technologies servent à établir des liens logiques et des relations d'inférences entre différentes données dans un langage compréhensible par les ordinateurs. Une vision similaire de la plus-value du web sémantique est partagée par Gregory Grefenstette lorsqu'il souligne la distinction entre la notion de « sémantique » dans le domaine de la linguistique et dans le domaine informatique, insistant sur le fait qu'en informatique la sémantique permet aux machines d'accéder au sens intellectuel des contenus, autrement uniquement accessible aux humains :

> « En linguistique on entend couramment par sémantique l'étude de la signification, en faisant une distinction entre signification et sens. Dans l'informatique, particulièrement dans l'idée du Web sémantique, la sémantique désigne l'assignation des chaînes des caractères dans une structure formelle qui explicite les relations entre choses, rendant opérationnelle l'idée de sens. »[53]

Notons tout de même que l'emploi du terme « sémantique », en raison de son rattachement traditionnel à la linguistique, est à l'origine d'un certain nombre d'incompréhensions et donc de freins au développement de ce web dit « sémantique »[54]. En effet, la conception d'un usage informatique de cette notion, basé sur la structuration des données, n'est pas forcément évidente à appréhender, particulièrement en l'absence de cas d'application concrets, les travaux sur le web sémantique étant initialement purement théoriques. Ainsi,

51 MONNIN, Alexandre. *op.cit.*

52 PEDAUQUE, T. Roger. *op.cit*. pp.102-103

53 GREFENSTETTE, Gregory, *et al*. Chapitre 3 : L'utilisation de la sémantique dans les applications basées sur la recherche d'information. In : GRIVEL, Luc. *La recherche d'information en contexte : outils et usages applicatifs*. Paris : Lavoisier, Hermès Sciences, 2011. p.97. Traité des sciences et techniques de l'information. ISBN 978-2-7462-2581-7.

54 POUPEAU, Gautier. 2011. *op.cit*.

entravés par une dénomination n'étant peut-être pas la plus adaptée et par les contraintes techniques de l'époque, le web sémantique ne trouve qu'un faible écho auprès des professionnels de l'informatique ou de l'information et demeure quasiment inconnu du grand public.

II.III Le web des données, un web sémantique

Le tournant arrive en 2007 lorsque deux chercheurs travaillant sur le sujet, James Hendler et Chris Welty, admettent la trop grande complexité des notions qu'introduisait le web sémantique ; Tim Berners-Lee arrive au même constat et décide de recentrer le projet sur le « web des données »[55]. Le lancement la même année du gisement de données structurées DBpedia, puis le discours de Tim Berners-Lee sur la libération des données (« *Raw data now* ») lors de la conférence TED de 2009 augmentent la visibilité du projet et permettent de le faire connaître d'un public plus large[56].

Toutefois, aussi bien dans la littérature scientifique que dans la littérature et les discours professionnels, les notions de « web sémantique » et de « web des données » sont souvent présentées comme des synonymes interchangeables. En effet, la différence entre les deux est subtile comme le souligne la page du W3C consacrée aux données liées (*linked data*) et expliquant que : « Le web sémantique est un web de données »[57]. Cependant la définition donnée sur cette page introduit tout de même une distinction notable : le web des données est rendu possible par les technologies du web sémantique qui ont vocation à permettre une standardisation des formats de données et à assurer leur réutilisation. Par conséquent nous devrions plutôt parler du « web des données liées » par opposition à un « web des données » qui ne serait qu'une base de données géante dans laquelle les données seraient uniquement recensées, et au « web social » qui, comme le souligne Jean-Michel Salaün, met l'accent sur le partage de documents, dont la stabilité et la pérennité n'est pas toujours

55 *Ibid*.

56 *Ibid*.

57 W3C. Linked Data. *World Wide Web Consortium (W3C),* [en ligne]. 2015. [Consulté le 9 mars 2018]. Disponible à l'adresse : https://www.w3.org/standards/semanticweb/data
« *The Semantic Web is a Web of Data — of dates and titles and part numbers and chemical properties and any other data one might conceive of. The collection of Semantic Web technologies (RDF, OWL, SKOS, SPARQL, etc.) provides an environment where application can query that data, draw inferences using vocabularies, etc. However, to make the Web of Data a reality, it is important to have the huge amount of data on the Web available in a standard format, reachable and manageable by Semantic Web tools. Furthermore, not only does the Semantic Web need access to data, but relationships among data should be made available, too, to create a Web of Data (as opposed to a sheer collection of datasets). This collection of interrelated datasets on the Web can also be referred to as Linked Data.* »

assurée, plutôt que sur le partage des données permettant de générer les-dits documents[58]. Enfin le « web des données liées » est également à distinguer du « web des documents » dont la lecture et la compréhension sont destinées aux êtres humains[59].

Toutefois, dans la mesure où le web a été créé pour répondre au besoin d'êtres humains de stocker et partager de l'information grâce à des technologies informatiques, faire une distinction nette entre d'un côté un web conçu uniquement pour les humains (le web des documents) et de l'autre un web conçu uniquement pour les machines (le web des données) n'aurait aucun sens puisque ces deux « cibles » sont indissociables. Cette relation de co-dépendance entre l'humain et la machine dans le contexte du web est expliquée ainsi par Stéphane Crozat :

> « En fait, toute interaction homme-machine procède d'une documentarisation et d'une donnéisation puisque l'ordinateur fabrique des signes avec des données pour se faire comprendre de l'homme et que l'homme, en retour, fabrique des données avec des signes pour se faire comprendre de la machine. »[60]

Ici la définition de la « documentarisation » reprend celle de Roger T. Pedauque puisqu'il s'agit d'associer à un objet des données représentées sous la forme de signes intelligibles pour les humains tandis que la « donnéisation » apparaît comme une traduction du terme anglais « *datafication* » renvoyant au fait de transformer en données informatiques des productions originellement sémiotiques[61]. Par conséquent, l'interaction homme-machine est rendue possible par le fait que l'ordinateur convertisse des données en signes linguistiques et qu'en retour l'être humain transmette des données à l'ordinateur. Il est alors nécessaire que l'humain connaisse différents langages informatiques afin que le dialogue avec la machine soit possible, ou bien qu'il utilise des interfaces, développées au préalable par des informaticiens connaissant les langages informatiques, servant à traduire le langage humain en langage interprétable par un ordinateur. En outre, l'explication de Stéphane Crozat part du principe que « le document est ce qui est accessible et interprétable par l'homme et la donnée est ce qui est accessible et calculable par la machine »[62]. Or il s'agit là de définitions

58 SALAÜN, Jean-Michel. Du document à la donnée et retour : la fourmilière ou les Lumières. In : CALDERAN, Lisette, *et. al*. *Le document numérique à l'heure du web de données. Séminaire Inria Carnac, 1er-5 octobre 2012*. Paris : ADBS Éditions, 2012. p.13. Sciences et techniques de l'information. ISBN 978-2-84365-142-7.

59 DELESTRE, Nicolas et MALANDAIN, Nicolas. *op.cit*. p.115

60 CROZAT, Stéphane. Écrire avec une machine à calculer, écrire pour une machine à calculer, *I2D – Information, données & documents*. 2016/2, Vol.53, pp.62-64. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-i2d-information-donnees-et-documents-2016-2-page-62.htm

61 *Ibid*.

62 *Ibid*.

quelque peu simplifiées dans la mesure où un être humain peut également accéder à et travailler sur des données, à moins que nous considèrions qu'en raison de sa capacité à analyser et comprendre le contenu des données, l'être humain accède plutôt à de l'information, la donnée brute, c'est-à-dire non interprétée, restant alors l'apanage des machines.

En somme, juxtaposées les unes aux autres ces définitions des différents web et de la relation homme-machine confirment l'idée précédemment évoquée selon laquelle le web des données liées, que nous appelons ici par convention « web des données », et parfois encore appelé par abus de langage « web sémantique », est une surcouche à destination des machines du web des documents, qui est lui à destination des humains, développée grâce aux technologies du web sémantique.

I.IV Les entraves à l'essor du web des données

L'expansion du web des données est limitée par des contraintes techniques et de temps. En effet, sa mise en place nécessite un travail conséquent de mise en forme et de mise en relation des données nécessitant des compétences techniques poussées[63] ; or les organisations et les individus n'ont pas forcément à disposition l'expertise requise ou le temps nécessaire à la formation d'un expert, pas plus que le temps de concevoir et mettre en place ce type de projet. De plus la généralisation du web de données nécessiterait également de retravailler toutes les données déjà disponibles sur le web avec les standards du web sémantique et notamment d'attribuer un Uri à chaque entité sur chaque page web[64]. Un travail titanesque qui ne sera probablement jamais réalisé, d'autant plus que certains acteurs, notamment les acteurs privés présents sur des marchés concurrentiels, n'ont que peu d'intérêt à rendre leurs données publiques et réutilisables par d'autres. Toutefois, si beaucoup de sémantique est difficile à mettre en place, l'exploitation d'un peu de

63 POUCHOT, Stéphanie et EPRON, Benoît. Chapitre 8 : Classer numérique. In : SALAÜN, Jean-Michel et HABERT, Benoît (dir). *Architecture de l'information : méthodes, outils, enjeux*. Louvain-la-Neuve : de boeck, 2015. p.179. Information & Stratégie. ISBN 978-2-8041-9140-5.

64 RIGOUSTE, Loïs, et al. Chapitre 4 : Analyse sémantique et moteurs de recherche, apport des entités nommées. In : GRIVEL, Luc. *La recherche d'information en contexte : outils et usages applicatif*. Paris : Lavoisier, Hermès Sciences, 2011. pp. 119-120. Traité des sciences et techniques de l'information. ISBN 978-2-7462-2581-7.

sémantique suffit à créer de la valeur ajoutée[65], c'est-à-dire que dans la mesure où les technologies sémantiques permettent de désambiguïser des objets, notamment grâce à l'utilisation d'identifiants, cela peut donner un avantage compétitif et un gain d'efficacité à celui qui les exploitent. Ainsi, si le web des données ne présente pas d'intérêt réel dans tous les contextes, l'utilisation des technologies du web sémantique fait sens dans de nombreux environnements, y compris concurrentiels. Par conséquent, bien que « web sémantique » et « web des données » soient souvent utilisés de manière interchangeable, et que les technologies du web sémantique aient été créées pour permettre la mise en place du web des données liées, le web sémantique peut exister sans le web des données, dans des systèmes fermés contenant par exemple des ontologies maisons n'ayant pas vocation à être diffusées sur le web.

En outre, concernant l'utilisation de technologies sémantiques dans le cadre d'une recherche d'information celle-ci n'est pas toujours pertinente notamment car les outils devant permettre d'assurer la fiabilité des données et de leurs relations ne sont pas encore au point, par conséquent ces dernières peuvent être d'une qualité incertaine et donc juger inexploitables[66]. Cependant l'intégration des technologies du web sémantique présente un intérêt certain pour l'amélioration des systèmes de recherche d'information dans la mesure où les objectifs des deux projets se rejoignent, comme par exemple l'utilisation d'Uri pour désambiguïser les concepts auxquels sont associés les termes linguistiques.[67]

II.V La recherche d'information dans le web de données

Bien que la feuille de route du web sémantique ait été publiée il y a vingt ans par Tim Berners-Lee, force est de constater que le web des données n'en est encore qu'à ses débuts. En effet, il a été imaginé pour faciliter la recherche d'information en permettant aux machines de répondre à des requêtes utilisateurs complexes. Ainsi plutôt que de multiplier les requêtes simples et faire lui-même les liens logiques entre les différentes réponses,

65 HENDLER, James. The dark side of the semantic Web. *IEEE Intelligent Systems*, Janvier/Février 2007, Vol.22, n°1, pp.2-4. Également disponible en ligne à l'adresse : https://www.computer.org/csdl/mags/ex/2007/01/x1002.html

66 AUSSENAC-GILLES, Nathalie.,Chapitre 8. Le web sémantique, quel renouvellement pour la recherche d'information ? In : BOUGHANEM, Mohand et SAVOY, Jacques. *Recherche d'information : état des lieux et perspectives*. Paris : Lavoisier, 2008. p.259. Collection Recherche d'information et web. ISBN 978-2-7462-2005-8

67 *Ibid*. p.237

l'utilisateur doit pouvoir formuler une requête élaborée et c'est la machine qui agrège alors automatiquement les différentes informations et serait, par exemple, capable de donner l'ensemble des activités d'une personne sur une année donnée[68]. Si dans les faits les moteurs de recherche ne sont pas encore capables de répondre à des requêtes complexes, d'importants progrès ont tout de même été réalisés comme le prouve notamment l'évolution du leader des moteurs de recherche. En effet, depuis 2012 l'algorithme de Google prend en compte les langages reposant sur des triplets sémantiques et intègre leurs données sous forme de *Knowledge Graph*[69], un encart situé en première page de résultats dans lequel sont agrégées des informations structurées issues de différentes sources web.

En parallèle de l'évolution lente des moteurs de recherche traditionnels apparaissent des moteurs de recherche dits « sémantiques » « c'est-à-dire potentiellement capables d'interpréter et de désambiguïser le sens des données du web (pages textuelles, vidéos, images, etc.) »[70]. Lorsqu'un utilisateur effectue une requête sur un moteur de recherche celui-ci se réfère à un index de concepts recensant tous les concepts de tous les documents fournis au moteur de recherche et ayant fait l'objet d'une indexation, afin de retourner à l'utilisateur la liste des documents comprenant le terme de sa requête[71]. Suivant ce schéma, les modules sémantiques peuvent intervenir à trois niveaux : entre l'utilisateur et l'interface, l'interface et l'index, l'index et les documents[72]. Gregory Greffenstette *et. al.* soulignent l'existence de trois fonctions sémantiques : de type, c'est-à-dire la possibilité d'ajouter son type à une entité, d'égalité, ce qui permet de spécifier que deux choses distinctes renvoient au même sens, de relation, c'est-à-dire spécifier la nature de la relation entre deux objets[73]. Or les données structurées possèdent déjà une, si ce n'est toutes, ces fonctions, par conséquent un moteur de recherche sémantique pourrait les interroger et faire bénéficier l'utilisateur de ces informations supplémentaires afin de le guider dans sa requête grâce à de l'auto-suggestion par exemple. De plus, bien qu'ayant un temps de réponse plus long que

68 DELESTRE, Nicolas et MALANDAIN, Nicolas. *op.cit*. pp.37-38

69 KEMBELLEC, Gérald. Que voit réellement Google de la sémantique des pages web ? , *I2D – Information, données & documents*. 2016/2, Vol.53, p.65. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-i2d-information-donnees-et-documents-2016-2-page-65.htm

70 RIGOUSTE, Loïs, *et al. op.cit*. p.121

71 GREFENSTETTE, Gregory, *et al. op.cit*. p.100

72 *Ibid*.

73 *Ibid*. p.98

leurs homologues généralistes, les moteurs de recherche sémantiques parfois aussi appelés « moteurs de recherche intelligents » présentent l'intérêt de permettre la désambiguïsation du langage humain et donc l'obtention de réponses plus pertinentes. Selon leur configuration ils font généralement appel à une ou plusieurs des techniques suivantes : analyse linguistique, analyse syntaxique, correction orthographique, calcul des cooccurrences, reconnaissance et extraction automatique des entités nommées[74], alignements entre les termes d'une requête et ceux d'un thésaurus[75] ou d'un autre système d'organisation des connaissances.

Toutefois les technologies du web sémantique et le web de données n'ont pas seulement vocation à améliorer les moteurs de recherches. Parmi les différentes applications possibles figurent également une meilleure adaptabilité des outils et des interfaces aux différents types d'utilisateurs, ou encore le perfectionnement de la fouille de textes et de données (*text mining*, *data mining*) ainsi que des systèmes d'organisation des connaissances.[76]

III. Les systèmes d'organisation des connaissances dans le web des données

III.I Les systèmes d'organisation des connaissances

L'organisation des connaissances est une discipline scientifique s'intéressant initiallement à : « l'ordonnancement des sujets traités dans les documents, grâce à des plans de classement, des thésaurus, des listes de mots-clés et des systèmes d'organisation des connaissances »[77]. Fondée en 1989 par un groupe de chercheurs allemands réunis sous la bannière de l'Isko (*International Society for Knowledge Organization*), cette discipline est structurée autour de quatre axes : la théorie de l'organisation des connaissances, les systèmes d'organisation des connaissances, la représentation de l'organisation des connaissances, et les applications d'organisation des connaissances.[78] Pour les besoins de ce mémoire nous nous intéresserons plus particulièrement au second axe dans sa globalité, au troisième à travers le langage Skos

74 *Ibid*. pp.122-125

75 ACCART, Jean-Philippe et RIVIER, Alexis. *op.cit*. p.119

76 AUSSENAC-GILLES, Nathalie. *op.cit*. p.237

77 GNOLI, Claudio. Chapitre 2 : Des métadonnées représentant quoi ?. In : EL HADI Widad Mustafa. *L'organisation des connaissances : dynamisme et stabilité*. Paris : Lavoisier, 2012. p.51. Traité des sciences et techniques de l'information. ISBN 978-2-7462-3227-3

78 *Ibid*. p.52

et au quatrième pour ce qui concerne le rôle des systèmes d'organisation des connaissances dans le contexte de la recherche d'information dans le web des données.

Les systèmes d'organisation des connaissances (souvent abrégés « Soc » ou « KOS » pour « *Knowledge Organization Systems* ») servent à organiser l'information dans des domaines plus ou moins larges en formalisant l'expression des constituantes du domaine en question de façon à ce qu'elles puissent être efficacement comprises et réutilisées par son audience cible.[79] Une autre caractéristique des systèmes d'organisation des connaissances semble être la présence d'une organisation de ses composantes de différentes natures : terminologique, sémantique[80] ou encore hiérarchique, associative, etc.[81] Toutefois, contrairement aux points précédents, celui-ci semble ne pas faire l'unanimité étant donné que par exemple Manuel Zacklad considère que les Soc ne possèdent pas forcément de règles d'association explicites.[82] Néanmoins cette position peut se comprendre par le fait que cet auteur adopte une conception large des systèmes d'organisation des connaissances qu'il divise en six catégories : les classifications bibliothéconomiques, les langages documentaires et les thésaurus, les ontologies formelles et le web sémantique, les approches multidimensionnelles, les annuaires de ressources internet et les folksonomies, et enfin, les index automatiques des moteurs de recherche[83]. Il est d'ailleurs intéressant de noter que la typologie de Manuel Zacklad établit une distinction entre « les classifications épistémiques universelles de la bibliothéconomie et les approches à facettes universelles » et « les langages documentaires et les thésaurus » alors même que les classifications sont généralement considérées comme un type de langage documentaire au même titre que les

79 MASTORA, Anna. *et. al*. SKOS Concepts and Natural Language Concepts: an Analysis of Latent Relationships in KOSs, *Journal of Information Science*. 2017/4, Vol.43, p.4. Également disponible en ligne à l'adresse : https://www.researchgate.net/publication/303322229_SKOS_Concepts_and_Natural_Language_Concepts_an_Analysis_of_Latent_Relationships_in_KOSs
« KOSs are meant to organize information, either for a particular domain or with a broader coverage, serving, in any case and in principle, as a convention for expressing an area of interest in a way that its constituents are effectively communicated towards and among the targeted audience. »

80 ISAAC, Antoine. Entre thésaurus et ontologies : une affaire d'interopérabilité et d'alignement, *Documentaliste – Sciences de l'Information*. 2011/4, Vol.48, p.49. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2011-4-page-42.htm

81 MASTORA, Anna. *et. al*. *op. cit.* p.4

82 ZACKLAD, Manuel. Évaluation des systèmes d'organisation des connaissances, *Les Cahiers du numérique*. 2010/3, Vol.6, p.135. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-les-cahiers-du-numerique-2010-3-page-133.htm

83 *Ibid*. pp.136-145.

listes de vedettes-matière et les thésaurus.[84] Cette distinction serait due au fait que les classifications portent sur les sujets des documents tandis que les thésaurus représentent des concepts[85] ; leur objet étant différent cela donne lieu à deux catégories de systèmes d'organisation des connaissances différentes. Concernant les autres catégories, les ontologies dites « formelles » renvoient à des modèles de représentation de concepts permettant leur manipulation par des machines[86] ; cependant cette définition vague et l'essor des ontologies permis par celui du web sémantique font que ce terme est souvent utilisé de façon abusive pour faire référence à toute classification partageable sur le web grâce aux formats XML et RDF[87] auxquels nous pouvons également ajouter OWL, trois standards du web sémantique promus par le W3C. Les approches multidimensionnelles regroupent les ontologies dites « sémiotiques », qui sont des schémas de classification regroupant des expressions contextualisées dans un cadre transactionnel donné[88], et les approches à facettes locales. Ces deux outils s'inscrivent dans la logique du web « socio sémantique » visant à aider l'usager dans sa recherche d'information via des activités communicationnelles de coopération à plus ou moins grande échelle, par opposition au web sémantique « formel » proposé par Tim Berners-Lee, c'est-à-dire orienté vers l'exploitation automatique du web par les machines.[89] Cette vision axée sur la collaboration des usagers se retrouve également dans la cinquième catégorie de Soc qui réunit les annuaires de ressources internet collaboratifs et les folksonomies. Enfin, la sixième et dernière catégorie de Soc représentée dans la typologie de Manuel Zacklad concerne les index automatiques des moteurs de recherche dont le fonctionnement n'est pas sans rappeler celui des systèmes de recherche d'information abordés en première partie de ce mémoire.

Rappelons tout de même qu'il n'existe pas de liste standardisée des différents types de systèmes d'organisation des connaissances existants et qu'en raison de leur définition très

84 FEYLER, François. Vocabulaires contrôlés. *Savoirs CDI*, [en ligne]. Avril 2009. [Consulté le 05 juin 2018]. Disponible à l'adresse : https://www.reseau-canope.fr/savoirscdi/societe-de-linformation/tic-et-documentation/veille-technologique/formats-normes-et-standards/vocabulaires-controles.html

85 ZACKLAD, Manuel. *op. cit*. pp.136-137

86 MENON, Bruno. Les langages documentaires. Un panorama, quelques remarques critiques et un essai de bilan, *Documentaliste-Sciences de l'Information*. 2007(b)/1, Vol.44, p.26. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-18.htm

87 ZACKLAD, Manuel. *op. cit*. p.140

88 ZACKLAD, Manuel. *Introduction aux ontologies sémiotiques dans le Web Socio Sémantique*. 2005. p.4 et 8. Disponible en ligne à l'adresse : https://archivesic.ccsd.cnrs.fr/file/index/docid/62630/filename/sic_00001479.pdf

89 *Ibid*. pp.2-3 et ZACKLAD, Manuel. 2010. *op. cit*. p.141

large de nombreux outils peuvent être considérés comme des Soc. Par conséquent les typologies des systèmes d'organisation des connaissances semblent dépendre de leur contexte de création et d'utilisation. Ainsi la typologie proposée par Manuel Zacklad est fortement marquée par le contexte du web socio-sémantique faisant parti de ses objets de recherche, tandis que par exemple celle de la CLIR (*Council on Library and Information Resources*) est axée bibliothèque. Elle comprend ainsi trois grandes catégories elles-mêmes divisées en sous-catégories : les listes de termes réunissant les fichiers d'autorités, les glossaires, les dictionnaires et les index géographiques, les classifications et catégories qui comprennent les listes de vedettes-matière, les schémas de classification et de catégorisation ainsi que les taxonomies, et enfin les listes relationnelles, c'est-à-dire les thésaurus, les réseaux sémantiques et les ontologies.[90] Toutefois si nous comparons cette typologie à celle de Manuel Zacklad nous pouvons remarquer que bien que le nombre et les intitulés des catégories diffèrent une partie de leur contenu est semblable ; nous pouvons donc noter la présence des classifications, des thésaurus et des ontologies dans les deux typologies. Si les classifications et les thésaurus semblent faire unanimement partis des Soc, le cas des ontologies formelles, souvent appelées simplement « ontologies », fait davantage débat, ainsi certains auteurs, comme Antoine Isaac, considèrent qu'il ne s'agit pas de systèmes d'organisation des connaissances.[91] Par conséquent la prochaine sous-partie réalisera un focus sur les langages documentaires en tant que systèmes d'organisation des connaissances.

III.II Les langages documentaires

Initialement développés pour faire face à la croissance du volume d'information disponible,[92] les langages documentaires sont des vocabulaires contrôlés créés arbitrairement afin de lutter contre les ambiguïtés de la langue naturelle et d'harmoniser l'expression des « sujets »[93] de ressources documentaires de tous types. Ils permettent ainsi aux professionnels de l'information d'indexer ces ressources de façon homogène et peuvent également être exploités lors de la formulation de requêtes par les utilisateurs de systèmes

90 HODGE, Gail. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Council on Library and Information Resources. 2000. ISBN 1-887334-76-9. 1. Knowledge Organization Systems: An Overview. Également disponible en ligne à l'adresse : https://www.clir.org/pubs/reports/pub91/1knowledge/

91 ISAAC, Antoine. *op. cit*. p.49.

92 MENON, Bruno. 2007(b). *op. cit*. p.18

93 MANIEZ, Jacques. *op.cit*. p.172

documentaires.[94] Comme nous l'avons vu précédemment, ces langages sont traditionnellement divisés en trois catégories : les classifications, les listes de vedettes-matière et les thésaurus ; toutefois il peut arriver qu'on leur associe également les taxonomies et les ontologies. Or, tout comme la notion d'« ontologie », celle de « taxonomie » (ou « taxinomie ») est vague dans la mesure où elle renvoie aujourd'hui aussi bien à son usage initial de système de classification des organismes vivants qu'à tout langage documentaire possédant une organisation hiérarchique[95], ainsi, si nous nous en tenons à cette définition un thésaurus peut être une taxonomie. Par conséquent, afin d'éviter toute ambiguïté, nous nous en tiendrons à la définition de Bruno Menon :

> « Pour nous, une taxonomie est un cadre d'organisation pour des ressources numériques de toutes natures (et pas seulement documentaires – en cela les taxonomies ne sont pas toujours des langages documentaires) destiné à en permettre une présentation ordonnée et y donnant accès par navigation hypertextuelle. »[96]

Chacune des trois familles de langages documentaires est apparue dans un contexte particulier afin de répondre à des besoins spécifiques. Ainsi, l'accroissement de la production éditoriale au XIX^{ème} siècle a conduit à l'apparition des classifications bibliographiques, comme la classification décimale de Dewey et la classification décimale universelle. Ces systèmes à vocation encyclopédique ont également été pensés afin de faciliter le rangement physique des documents[97]. Apparues à la même époque[98], les listes de vedettes-matière répondent à un besoin pourtant différent : celui d'indexer des unités documentaires distinctes pouvant être contenues dans une même ressource physique[99], par exemple des articles scientifiques publiés dans une revue. Ce langage documentaire, dont parmi les exemples les plus connus figurent la LCSH (*Library of Congress Subject Headings*), le répertoire de vedettes-matière de l'université de Laval ou encore Rameau (Répertoire d'autorité-matière encyclopédique et alphabétique unifié), se construit autour du sujet principal (la « vedette ») d'un document. Les listes de vedettes-matière étant des langages pré-coordonnés il est possible d'associer les vedettes selon des règles et une syntaxe définies, voire même, dans le cas du langage Rameau, de leur ajouter certain nombre de subdivisions afin de permettre une indexation

94 FEYLER, François. *op. cit.*
95 MENON, Bruno. 2007(b). *op. cit*. p.23
96 *Ibid*.
97 ZACKLAD, Manuel. 2010. *op.cit.* pp.136-137
98 MENON, Bruno. 2007(b). *op.cit*. p.19
99 HODGE, Gail. *op.cit*.

plus fine des documents.[100] Enfin, la dernière famille de langage documentaire, celle des thésaurus, est apparue à la fin des années 1950 et a connu un essor concomitant à celui de l'informatique documentaire dans les années 1980.[101] Langages post-coordonnés, les thésaurus sont composés d'un ensemble de termes contrôlés représentant des concepts ainsi que de relations hiérarchiques, associatives et d'équivalence permettant de lier ces différents concepts en un ensemble thématique ou de sous-thématiques (micro-thésaurus) cohérentes.[102] Parfois comparés aux listes de vedettes-matière, les thésaurus sont toutefois plus simples à comprendre et à manier, que ce soit par les professionnels ou par les usagers.[103]

En dépit de leurs différences de structure et de contexte d'usage les langages documentaires ont tous la même fonction : harmoniser l'expression des sujets de documents donnés. Or, cette fonction d'harmonisation est assurée par le recours à des vocabulaires contrôlés qui, s'ils ont l'avantage de dissiper l'ambiguïté des langues naturelles, en limitent également la richesse et provoquent donc systématiquement une perte d'information.[104] En outre, reposant sur des choix arbitraires de la part des concepteurs des vocabulaires, le choix des termes retenus peut ne pas faire l'unanimité aussi bien pour les professionnels chargés d'indexer les documents que pour les auteurs des documents, qui ne trouvent pas forcément de termes adéquats parmi ceux retenus dans le vocabulaire, et les utilisateurs finaux du système d'organisation des connaissances qui ne sont toujours familiers du fonctionnement des langages documentaires.[105] Bien que d'après Jacques Maniez les thésaurus aient été inventés pour passer outre la rigidité des langages documentaires[106], force est de constater qu'ils n'en restent pas moins des langages artificiels construits par des êtres humains à partir de leurs propres représentations et de celles de leur institution sur des thématiques données ; un thésaurus ne peut donc pas non plus avoir de vocation universelle.

Enfin, en tant que langages normalisés pouvant être traduits dans plusieurs langues, pouvant faire l'objet d'alignement et représentant des sujets ou des concepts plutôt que des termes,

100 Centre national RAMEAU. *Guide d'indexation RAMEAU*, [en ligne]. 7ème édition. Bibliothèque nationale de France, 2017, p.4, 10 et 26. [Consulté le 09 juin 2018]. Disponible à l'adresse : http://rameau.bnf.fr/docs_reference/pdf/Guide_RAMEAU_2017.pdf#page=7

101 FEYLER, François. *op.cit.*

102 ZACKLAD, Manuel. 2010. *op.cit*. pp.137-138

103 MASTORA, Anna, *et. al. op.cit.* p.4 et MENON, Bruno. *op.cit.* p.21

104 ZARGAYOUNA, Haïfa, ROUSSEY, Catherine et CHEVALLET, Jean-Pierre. *op.cit*. p.60.

105 DESFRICHES DORIA, Orélie et ZACKLAD, Manuel. *op.cit*. p.22.

106 MANIEZ, Jacques. *op.cit*. pp.323-324

les langages documentaires permettent de passer outre les barrières de la langue et du temps.[107] Ces avantages font qu'ils sont également utilisés dans des systèmes de recherche d'information afin d'en améliorer les performances en constituant par exemple l'index de mots-clefs exploitables par le système lors d'une recherche[108], ou encore en permettant à l'utilisateur de naviguer à travers des catégories établies à partir des systèmes d'organisation des connaissances, comme par exemple les micro-thésaurus.[109] En effet, le recours à des vocabulaires contrôlés permet d'éviter la synonymie, génératrice de silence documentaire, ainsi que la polysémie et l'homonymie, responsables du bruit documentaire ; ainsi le recours aux langages artificiels permet d'améliorer les taux de rappel et de pertinence des systèmes de recherche d'information.[110] Cet avis est également partagé par Bruno Menon qui insiste sur le fait qu'en l'absence de langage documentaire pour contrôler la polysémie et la synonymie dans un système de recherche d'information, cette tâche, cognitivement coûteuse, revient à l'usager et est à réitérer pour chaque nouvelle recherche[111], ce qui n'est pas sans incidence sur l'expérience utilisateur. Ainsi les langages documentaires, utilisés en arrière-plan des systèmes de recherche d'information, permettent de donner l'illusion de moteurs de recherche « intelligents » capables d'effectuer des liens logiques en exploitant la structure et les relations établies dans et entre les vocabulaires contrôlés.[112]

Les systèmes d'organisation des connaissances sont également utilisés pour établir un cadre de référence spécifiant le modèle à adopter afin de structurer des documents ou des

107 DEXTRE CLARKE, Stella G. *Thesauri, interoperability and the role of ISO 25964*. Diapositive n°8. [en ligne]. [Consulté le 02 juin 2018]. Disponible à l'adresse : http://www.docslides.com/thesauri-interoperability-and-the-role-of-iso-25964

108 SUOMINEN, Osma, *et. al. Publishing SKOS vocabularies with Skosmos*. Manuscript submitted for review. 2015, p.2. Disponible en ligne à l'adresse : http://skosmos.org/publishing-skos-vocabularies-with-skosmos.pdf

109 MERCEDES MARTÍNEZ-GONZÁLEZ, M. et ALVITE-DÍEZ, María-Luisa. On the evaluation of thesaurus tools compatible with the Semantic Web, *Journal of Information Science*. 2014/6, Vol.40. p.1. Également disponible en ligne à l'adresse : https://www.researchgate.net/publication/271731043_On_the_evaluation_of_thesaurus_tools_compatible_with_the_Semantic_Web

110 FEYLER, François. *op.cit*.

111 MENON, Bruno. 2007(b). *op.cit*. pp.21-22

112 HUDON, Michèle. ISO 25964 : pour le développement, la gestion et l'interopérabilité des langages documentaires, *Documentation et bibliothèques*. 2012(a)/3, Vol.58, p.131. Également disponible en ligne à l'adresse : https://www.erudit.org/fr/revues/documentation/2012-v58-n3-documentation01721/1028903ar.pdf

données dans un outil logiciel compatible avec le Soc en question.[113] Ainsi, par exemple, depuis la publication de la norme sur les thésaurus Iso 25964 sont apparus de nombreux outils de gestion de thésaurus (Ginco, VocBench, OpenTheso, TemaTres, etc.) compatibles avec le format Skos.

III.III Skos : un format pour intégrer les systèmes d'organisation des connaissances au web des données

Initialement imaginé pour permettre la structuration des thésaurus[114], le *Simple Knowledge Organization System* (Skos) est un format de données à destination de n'importe quel système d'organisation des connaissances. Basé sur le modèle RDF, il est donc compatible avec les technologies du web sémantique et fait l'objet d'une recommandation du W3C depuis le 18 août 2009.[115]

Antoine Isaac en propose une définition assez complète :

> « modèle, qui se veut à la fois simple et compatible avec une majorité d'approches existantes (thésaurus, classifications, etc.) permet la représentation de concepts et vocabulaires (ConceptScheme), de données terminologiques attachées à un concept (libellé préféré ou alternatif), potentiellement multilingue, de liens sémantiques entre concepts (relations générique ou associative) et de notes (d'application, définitions). »[116]

L'intérêt du format Skos apparaît comme une évidence dès lors que tout comme Laurence Maroye nous considérons que :

> « Le web sémantique repose sur l'exploitation des ontologies conceptuelles, qui représentent le réel, et les systèmes d'organisation des connaissances (SOC), comme les thésaurus, qui permettent d'indexer des documents qui traitent du réel. »[117]

En effet, Skos repose principalement sur l'utilisation d'Uri afin d'identifier les concepts présents dans les langages documentaires. L'utilisation de ces identifications uniques permet d'indexer les ressources avec davantage de précision qu'en utilisant les termes issus de vocabulaires contrôlés.[118] Cependant les Uri sont utilisés pour faciliter la compréhension et le

113 MAHE, Sylvain *et. al*. Gestion des connaissances et systèmes d'organisation de connaissances : premier modèle et retours d'expérience industriels, *Document numérique*. 2010/2, Vol.13, p.66. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-document-numerique-2010-2-page-57.htm

114 AUSSENAC-GILLES, Nathalie. *op.cit*. p.261

115 W3C. *SKOS Simple Knowledge Organization System,* [en ligne]. Mise à jour le 13 décembre 2012. [Consulté le 10 juin 2018]. Disponible à l'adresse : https://www.w3.org/2004/02/skos/

116 ISAAC, Antoine. *op.cit*. p.49

117 MAROYE, Laurence. ISO 25964 : de la distinction formelle concept/terme préconisée par la norme pour la création et la gestion des thésaurus, *I2D – Information, données & documents*. 2015/1, Vol.52, p.75. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-1-page-72.htm

118 SUOMINEN, Osma, *et. al*. *op. cit*. p.2

traitement de concepts par des ordinateurs et non pas des êtres humains, par conséquent le format Skos permet également d'associer à ces identifiants un certains nombre de termes issus du langage humain afin de les rendre intelligibles par ces derniers. Permettant de caractériser les relations entre les concepts, Skos est également complété par une extension intitulée « Skos XL » (S*KOS eXtension for Labels*) utilisée afin de caractériser les relations entre les différents libellés représentant un concept.[119]

Toutefois, précisons que l'utilisation d'un format basé sur RDF n'est pas suffisant pour assurer la présence et l'utilisation des systèmes d'organisation des connaissances dans le web des données, dans la mesure où ce format doit pouvoir être exploitable par des outils logiciels adaptés à la publication du contenu du Soc en question. C'est cette réflexion qui a conduit la Bibliothèque nationale de Finlande a développer « Skosmos »[120], une plate-forme *open source* de diffusion de vocabulaires contrôlés au format Skos aujourd'hui utilisé notamment pour la diffusion du thésaurus de la FAO (Organisation des Nations unies pour l'alimentation et l'agriculture) « Agrovoc », du thésaurus de l'Unesco et des ressources terminologiques de l'Inist-CNRS (Institut de l'information scientifique et technique-Centre national de la recherche scientifique) disponibles sur la plate forme « Loterre ». Toutefois Skosmos n'est pas le seul outil permettant la diffusion de vocabulaires contrôlés au format Skos, nous pouvons, par exemple, également citer « Ginco-Diff »[121] développé par le Ministère de la Culture et de la Communication pour permettre la publication sur le web de thésaurus conçus avec le logiciel dédié « Ginco ».

Bien que les exemple cités ci-dessus renvoient principalement à des thésaurus, rappelons que Skos peut être utilisé avec une large variété de systèmes d'organisation des connaissances, bien que certains, comme la classification décimale de Dewey ou les listes de vedettes-matière, soient dans les faits peu adapter à ce format en raison de leur construction pré-coordonnée basée sur la juxtaposition de concepts[122], ce qui rend notamment plus difficile la mise en place d'alignement[123] (parfois aussi appelés « *mapping* »), c'est-à-dire l'établissement de liens d'équivalence, hiérarchiques ou associatifs entre des concepts issus

119 MASTORA, Anna, *et. al. op. cit*. p.3
120 http://skosmos.org/
121 https://github.com/culturecommunication/ginco
122 MASTORA, Anna, *et. al. op. cit*. p.17
123 DEXTRE CLARKE, Stella G. *op. cit*. diapositive 22

de vocabulaires différents.[124] En effet l'enjeu central du format Skos est celui de l'interopérabilité entre différents vocabulaires[125] dans la logique d'interconnexion du web des données liées. La notion d'alignement va de paire avec celle d'interopérabilité qui renvoie à la capacité d'un système informatique d'échanger des informations avec un autre système.[126] Dans le web des données l'interopérabilité est basée sur deux modèles : « la roue et l'essieu » (« *hub and spoke* ») dans lequel des référentiels de type thésaurus ou listes d'autorité servent de point d'entrée commun à plusieurs jeux de données ou la « navigation intuitive » (« *follow your nose* »)[127] qui consiste à parcourir les liens d'un jeu de données, ou de façon plus générale d'un document, se rapprochant ainsi de la logique de la sérendipité. Notons que le fonctionnement du modèle de la roue et de l'essieu est assuré par l'emploi d'Uri dans chacun des jeux de données interconnectés, ce qui confirme l'intérêt de l'utilisation du format Skos afin de doter les systèmes d'organisation d'Uri et les rendre ainsi intégrables au web des données. Pourtant, si l'interopérabilité présente des avantages certains en termes d'exposition des données elle peut également être un frein au développement de systèmes d'organisation des connaissances au contenu très spécifiques et donc moins facilement alignable avec d'autres.[128]

III.IV Iso 25964, une norme pour rendre les thésaurus compatibles avec le web des données

En perte de vitesse depuis les années 1990, les thésaurus connaissent un regain d'intérêt avec l'apparition du web des données[129]. En effet, le fonctionnement des thésaurus reposant sur la mise en relation de relations de concepts n'est pas sans rappeler celui du triplet RDF « sujet – prédicat – objet », ce qui peut éventuellement également expliquer pourquoi est-ce que les principaux exemples d'application du format Skos sont les thésaurus. En outre, alors que des systèmes d'organisation des connaissances comme les classifications servent à organiser des sujets, aussi bien intellectuellement que dans l'espace physique, la vocation

124 DEXTRE CLARKE, Stella G. *op. cit*. diapositive 19
125 ISAAC, Antoine. *op.cit.* p.49
126 DEXTRE CLARKE, Stella G. *op. cit*. diapositive 10
127 BERMES, Emmanuelle et POUPEAU, Gautier. Les technologies du web appliquées aux données structurées. In : CALDERAN, Lisette *et. al*. *Le document numérique à l'heure du web de données. Séminaire Inria Carnac, 1er-5 octobre 2012*. Paris : ADBS Éditions, 2012. p.75. Sciences et techniques de l'information. ISBN 978-2-84365-142-7.
128 MAROYE, Laurence. *op. cit*. p.75
129 DEXTRE CLARKE, Stella G. *op. cit*. diapositive 3

première des thésaurus a toujours été de décrire des concepts[130] , qui, comme nous l'avons expliqué précédemment, sont représentés par des URI exploitables par le format Skos. Par conséquent il est tout à fait possible d'imaginer une exploitation des thésaurus en arrière-plan des systèmes de recherche d'information afin de pouvoir effectuer des recherches directement basées sur les concepts et non plus sur les termes les représentant.[131]

Ce ré-engouement pour les thésaurus a été rendu possible par la standardisation du format Skos en 2009 et la création de la norme Iso 25964 dont la première partie « Thésaurus et recherche d'information » a été publiée en 2011 et la seconde « Interopérabilité avec les autres vocabulaires » en 2013.[132] Avant la création de cette norme, les thésaurus étaient encadrés par deux normes distinctes : Iso 5964 pour les thésaurus multilingues et Iso 2788 pour les thésaurus monolingues. Publiées respectivement en 1985 et 1986, ces deux normes, en dépit d'une évaluation tous les cinq ans, n'avaient encore jamais été révisées. C'est sous l'impulsion d'un groupe de spécialistes britanniques travaillant sur une nouvelle norme nationale que le travail autour de la norme Iso 25964 a pu être entamé.[133] En effet, dans les années 2000, face aux importantes évolutions technologiques des dernières années, plusieurs pays, en particulier le Royaume-Uni, prennent l'initiative de réviser la norme à l'échelle nationale. La norme britannique BS 8723 « *Structures vocabularies for information retrieva*l » est présentée en 2007 au TC46 (Comité technique 46 « Information et documentation ») de l'Iso et servira de base à la réflexion autour de la nouvelle norme.[134]

La norme Iso 25964 formalise la distinction entre « concept » et « terme », le terme servant à représenter l'« unité de pensée » qu'est le concept.[135] Cette norme invite donc à un changement de paradigme dans la mesure où elle demande à ce que les thésaurus ne soient

130 ZACKLAD, Manuel. 2010. *op. cit*. p.138

131 MERCEDES MARTÍNEZ-GONZÁLEZ, M. et ALVITE-DÍEZ, María-Luisa. *op.cit*. p.1

132 DALBIN, Sylvie, *et. al*. *Livre banc ISO 25964-1 – Thésaurus pour la recherche documentaire* [en ligne]. AFNOR, 2013. p.8. [Consulté le 12 juin 2018]. Disponible à l'adresse : http://dossierdoc.typepad.com/files/iso25964-1-livre-blanc-janvier-2013-vfinale.doc

133 HUDON, Michèle. 2012(a). *op. cit*. p.130

134 HUDON, Michèle. Chapitre 13 : ISO 25964 : vers une nouvelle norme pour l'organisation et l'accès à l'information et aux connaissances. In EL HADI Widad Mustafa. *L'organisation des connaissances : dynamisme et stabilité*. Paris : Lavoisier, 2012(b). pp.209-210. Traité des sciences et techniques de l'information. ISBN 978-2-7462-3227-3

135 MAROYE, Laurence. *op.cit*. p.76

plus construits autour des termes mais plutôt autour des concepts, ce qui montre bien que la norme a été pensée pour être compatible avec le standard Skos[136]. Enfin, tout comme la norme britannique dont elle s'inspire, Iso 25964 accorde une place importante à l'interopérabilité entre les thésaurus et les autres systèmes d'organisation des connaissances, en l'occurrence : les classifications, les *file plans* (classifications utilisées en *records management*), les taxonomies, les listes de vedettes-matière, les fichiers d'autorité, les anneaux de synonymes, les terminologies et les ontologies. Toutefois la norme est exclusivement destinée aux thésaurus ainsi son traitement des autres Soc se limite à une brève description et à des recommandations afin d'effectuer des alignements entre ces Soc et les thésaurus.[137]

Il est d'ailleurs intéressant de noter que si la norme Iso 25964 prévoit l'alignement entre thésaurus et ontologies cette pratique est pourtant déconseillée, et il est plus souvent recommandé de faire évoluer un thésaurus en ontologie.[138] Si la largeur de la notion d'« ontologie » que nous avons précédemment abordée fait que les travaux de modélisation des thésaurus sont parfois abusivement appelés ainsi[139], il est effectivement possible de faire évoluer les thésaurus en ontologies afin de pouvoir ajouter davantage de sémantique dans les attributs des concepts et leurs relations et ainsi vers évoluer le thésaurus vers une base de connaissance.[140]

Pour conclure, le présent état de l'art nous a permis de définir et contextualiser différentes notions au centre de notre problématique : la recherche d'information au sens d'*information seeking*, le web des données rendu possible par les technologies du web sémantique, et enfin les systèmes d'organisation des connaissances avec plus particulièrement l'exemple du thésaurus, un Soc ayant dû faire face à un changement de paradigme afin de pouvoir s'insérer dans le web des données et dont les utilisations ne se limitent plus à l'indexation de document.

136 MERCEDES MARTÍNEZ-GONZÁLEZ, M. et ALVITE-DÍEZ, María-Luisa. *op.cit*. p.2
137 DEXTRE CLARKE, Stella G. *op. cit*. Diapositives 23 et 24
138 *Ibid*. diapositive 26
139 MANIEZ, Jacques. *op.cit*. p.331
140 CYROT, Catherine et PREUSS, Christian. Réingénierie de thésaurus : une étude de cas, *Documentaliste-Sciences de l'Information*. 2009/3, Vol.46, p.5. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2009-3-page-4.htm

Si cet état de l'art prouve que les systèmes d'organisation des connaissances sont bel et bien exploités pour améliorer la recherche d'information et que leur compatibilité avec les technologies du web sémantique, en particulier le format Skos, ouvre les Soc à de nouvelles possibilités d'exploitation, il ne nous permet pas pour autant de déterminer clairement leur influence sur les systèmes de recherche d'information. C'est pourquoi cet état de l'art est à compléter par une étude de cas qui nous permettra d'observer et mieux comprendre l'utilisation pouvant être faite de systèmes d'organisation des connaissances présents dans le web des données, à travers un exemple concret.

Partie 2 : méthodologie de l'étude de cas

La seconde partie du présent mémoire introduit la méthodologie permettant de confronter les hypothèses formulées précédemment à travers l'étude d'un cas concret reposant sur plusieurs des notions abordées dans l'état de l'art. Pour ce faire nous avons choisi d'observer et analyser les enrichissements sémantiques permis par l'exploitation de systèmes d'organisation des connaissances dans la plate-forme de recherche en sciences humaines et sociales « Isidore ». Afin de donner davantage de consistance à ces observations, de les approfondir et d'apporter des éléments d'explications, dépassant le stade des conjectures, aux conclusions qui en sont tirées, nous avons désiré les compléter avec des entretiens menés avec les responsables des principaux systèmes d'organisations des connaissances exploités par Isidore.

I. Présentation de la plate-forme de recherche Isidore

« Isidore », pour « Interconnexion de Services et Interopérabilité des Données pour la Recherche et l'Enseignement », est une plate-forme de recherche, d'agrégation et de diffusion de données, publications et informations en sciences humaines et sociales librement accessible sur le web depuis 2011. Elle a été développée par la très grande infrastructure de recherche (TGIR) Huma-Num (à l'époque TGE Adonis) en partenariat avec le Centre pour la Communication Scientifique Directe ainsi qu'avec l'aide d'un consortium de sociétés privées spécialisées dans les technologies du web sémantique : Antidot, Sword et Mondeca.[141]

Le développement d'Isidore poursuit principalement un double objectif : offrir un point d'accès unique et pérenne à l'ensemble des données de la recherche en sciences humaines et sociales, constituer un ensemble de services à haute-valeur ajoutée à destination des chercheurs par le développement d'outils complémentaires permettant d'exploiter la base de

141 POUYLLAU, Stéphane. *Isidore : signaler, enrichir et valoriser les documents, données, informations numériques des sciences humaines*, [en ligne]. Paris : TGIR Huma-Num, 2013. Diapositives 4 et 16. [Consulté le 02 août 2018]. Disponible à l'adresse : https://www.rnbm.org/supports_anf/cirm2013/plateforme-ISIDORE.pdf

données d'Isidore de différentes façons.[142] La plate-forme met donc à profit les standards du web afin de normaliser et enrichir les notices documentaires qu'elle moissonne puis les exposer dans le web des données. Pour ce faire Isidore peut procéder de trois façons : par la récupération des métadonnées des notices documentaires mises à disposition par les producteurs de données partenaires dans des entrepôts dédiés au protocole OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*), par l'utilisation du protocole Sitemaps, permettant de signaler des pages web à un moteur de recherche, associé à la syntaxe RDFa pour l'expression des métadonnées, ou encore par des flux de syndication de contenus de types RSS ou Atom.[143] Une fois les données récupérées, Isidore effectue un certain nombre de traitements automatiques afin de les homogénéiser et de les enrichir à l'aide de référentiels-métiers. Enfin, l'ensemble des notices structurées et enrichies est convertie au format RDF et déposé dans un *triplestore* interrogeable via une interface Sparql *end point*.[144]

Isidore est également développé dans le contexte de la « science ouverte ». A la fois vague et vaste, cette notion, généralement abordée sous l'expression anglo-saxonne « *open science* », s'inscrit dans un paradigme considérant la science comme étant un bien commun devant être librement accessible et utilisable par tous. La science ouverte s'appuie sur plusieurs autres notions et mouvements en faveur du partage et de la réutilisation des données : le libre-accès aux publications, l'*open data*, c'est-à-dire la mise à disposition libre et gratuite de données sur le web, ou encore l'utilisation de licences ainsi que de formats ouverts permettant d'assurer l'interopérabilité entre différentes sources de données.[145] Son développement ayant été permis par l'évolution des technologies numériques, elle englobe à la fois des pratiques d'édition et de partage de la recherche scientifique, tout comme de nouvelles méthodes de travail[146], par exemple participatives, comme dans le cas des sciences dites « citoyennes » associant les citoyens à des projets de recherche. Isidore, en particulier, s'inscrit dans la démarche du libre-accès, revendiquant être « le plus gros projet d'open data

142 CAPELLI, Laurent, *et. al. Les guides de bonnes pratiques : comment contribuer à Isidore avec ses données numériques ?*, [en ligne]. Paris : TGIR Huma-Num, 2014. p.7. [Consulté le 04 août 2018]. Disponible à l'adresse : https://www.huma-num.fr/sites/default/files/guide-isidore.pdf

143 Ibid. pp.10-13

144 *Ibid*. p.16

145 Direction de l'information scientifique et technique – CNRS. *Livre blanc : une science ouverte dans une République numérique*, [en ligne]. Marseille : OpenEdition press, 2016. pp.80-81. Collection Laboratoire d'idées. Disponible à l'adresse : https://books.openedition.org/oep/1548

146 *Ibid*. pp.10-11

scientifique en France ».[147] A ce titre la plate-forme donne accès à plus de « 5500000 ressources en sciences humaines et sociales (SHS) provenant de 6000 sources différentes ».[148] Ces chiffres sont tout de même à relativiser car, bien qu'Isidore encourage cette démarche, la version en texte intégral des ressources indexées sur la plate forme n'est pas toujours librement consultable sur le web.

Enfin, Isidore s'inscrit dans l'offre de services proposée par la très grande infrastructure de recherche pour les humanités numériques. A ce titre, bien que la plate-forme soit librement accessible sur le web, elle s'adresse en priorité aux chercheurs et étudiants en sciences humaines et sociales, de par le caractère spécialisé de son contenu. Depuis janvier 2015 le projet affiche clairement des ambitions internationales puisque l'interface de consultation et les enrichissements sont désormais également disponibles en anglais et en espagnol afin d'aider les chercheurs non-francophones à évaluer l'intérêt des notices qu'ils consultent.[149] Notons également que si cette mise à jour de la plate-forme rend possible le moissonnage de documents en anglais et espagnol en provenance d'organismes anglophones ou hispanophones, comme l'archive ouverte de la *School of Oriental and African Studies* de l'université de Londres[150] ou la Bibliothèque nationale de l'Uruguay[151], des ressources en langues autres que le français étaient déjà disponibles sur la plate-forme, probablement car elles sont mises à disposition par des organismes francophones.

I.I Les enrichissements sémantiques sur Isidore

Isidore effectue des « enrichissements sémantiques » pour chacune des ressources qu'il indexe à l'aide de référentiels-métiers au format Skos. Ces référentiels, parfois également

147 TGIR Huma-Num. Signaler ses données avec Isidore. *Huma-num l'infrastructure des humanités numériques*, [en ligne]. Mise à jour le 09 avril 2018. [Consulté le 04 août 2018]. Disponible à l'adresse : https://www.huma-num.fr/services-et-outils/signaler

148 CO.SHS. Entrevue avec Huma-Num sur ISIDORE à la demande. *Produire – Découvrir – Explorer – CO.SHS,* [en ligne]. 10 mai 2018. [Consulté le 03 août 2018]. Disponible à l'adresse : https://co-shs.ca/fr/nouvelles/entrevue-avec-huma-num-sur-isidore-a-la-demande/

149 TGIR Huma-Num. Isidore speaks English, sino también español et toujours en français. *Le blog d'huma-num et de ses consortiums*, [en ligne]. Mise à jour le 19 mai 2015. [Consulté le 03 août 2018]. Disponible à l'adresse : https://humanum.hypotheses.org/921

150 http://eprints.soas.ac.uk/
Fiche de la collection sur Isidore : https://www.rechercheisidore.fr/annuaire/source/?collection=10670/3.mum5jl

151 http://www.bibna.gub.uy/
Fiche de la collection sur Isidore : https://www.rechercheisidore.fr/annuaire/source/?collection=10670/3.hrsv7p

appelés « référentiels scientifiques », sont des systèmes d'organisation des connaissances de type « vocabulaire contrôlé », ainsi qu'une base de données géographiques. Comme nous l'avons vu dans la partie précédente, les vocabulaires contrôlés reposent sur une logique de « concepts » et non pas de « termes », par conséquent Isidore peut les utiliser suivant le modèle de la roue et de l'essieu : les référentiels agissant comme un point nodal mettant en relation des ressources issues de différents silos de données par l'identification de concepts communs.

Actuellement les référentiels exploités par Isidore sont au nombre de neuf et sont utilisés sur trois sections distinctes de la notice documentaire enrichie : « Classification », « Enrichissements ? », et « Espaces géographiques ». La classification, c'est-à-dire l'attribution d'un ou plusieurs thèmes à la ressource indexée, se fait donc à partir des disciplines Hal et des catégories d'OpenEdition, tandis que l'identification des espaces géographiques mentionnés dans la ressource est assurée par GeoNames. Les six référentiels restant sont donc utilisés pour la section « Enrichissements ? », toutefois précisons que si les thésaurus Pactols, GeoEthno et Gemet sont des vocabulaires multilingues, gérant *a minima* des termes en français, anglais et espagnol, les trois autres sont monolingues : Rameau est en français, le LCSH est en anglais et le vocabulaire de la BNE (Bibliothèque nationale d'Espagne) est en espagnol. Par conséquent, Isidore étant disponible en français, anglais ou espagnol, le quatrième vocabulaire utilisé dans le champ « Enrichissements ? » dépend de la langue de consultation. Notre étude ne portant pas sur le multilinguisme et notre compréhension de la sémantique des termes dans une langue autre que le français étant plus limitée, nous n'avons pas jugé utile d'examiner de façon systématique les enrichissements sémantiques en provenance de la LCSH et de la BNE.

Bien que les référentiels Hal et OpenEdition s'apparentent à des classifications et plans de classement, et fassent donc partis de la famille des systèmes d'organisation des connaissances, nous avons choisi de les exclure de notre étude compte-tenu du nombre restreints de concepts qu'ils contiennent (28 pour Hal, 124 pour OpenEdition), et du fait que leur structure est très simple de par l'absence de relations autres que hiérarchiques ainsi que de toute autre information permettant de baliser la sémantique de ces concepts (notes d'application, définitions, termes alternatifs, etc.). Dans le même ordre d'idée, une base de données géographiques comme GeoNames obéit à une logique différente, quoique proche

sur certains points, des systèmes d'organisation des connaissances, et a donc été également exclue du cœur de l'analyse. Par conséquent, nous avons fait le choix de limiter notre étude à l'observation des quatre systèmes d'organisation des connaissances francophones exploités dans la section « Enrichissements ? » : Rameau, Gemet, Pactols et GeoEthno.

I.II Présentation de la section « Enrichissements ? »

Les notices documentaires publiées sur Isidore sont divisées en trois zones[152] : la zone de gauche, intitulée « Fiche de la ressource », correspond à la notice moissonnée une fois l'écriture de certains champs automatiquement normalisée lors du traitement documentaire effectué par Isidore, la zone de droite accueille les trois catégories d'enrichissements susmentionnées : « Classification », « Enrichissements ? », et « Espaces géographiques », ainsi que la section « Rebondir ?» donnant accès à un système de facettes établi à partir de ce que nous pourrions appeler les « métadonnées enrichies » de la ressource, c'est-à-dire une combinaison de certaines métadonnées de la notice (« Collection », « Source », « Organisation », « Type », « Langue », « Auteur ») et des enrichissements des sections « Classification » et « Enrichissements ? », ici rebaptisées « Sujets », afin de permettre au lecteur d'affiner sa recherche documentaire ; enfin la zone du bas de la notice, intitulée « Pour aller plus loin… », correspond à des suggestions de ressources.

La section « Enrichissements ? » se présente sous la forme d'un nuage de mots-clefs qui correspondent aux termes préférentiels des concepts identifiés par Isidore dans les métadonnées et le texte intégral du document. Les systèmes d'organisation des connaissances exploités figurent en bas du nuage de mots-clefs. Un clic sur leur nom provoque le surlignage dans une couleur donnée de tous les concepts issus du référentiel sélectionné. Sur l'illustration ci-dessous, les concepts issus de Rameau sont surlignés en vert, ceux issus de Gemet en jaune, ceux de Pactols en bleu, et enfin, celui de GeoEthno en violet ; notons que ces couleurs sont les mêmes sur toutes les notices ce qui facilite l'identification de chaque référentiel. En outre, sur certaines notices quelques concepts sont écrits en plus gros (par exemple sur l'illustration 3 : « Poitiers », « histoire », « Histoire » et « histoire ») et d'autres en plus petits. Ce procédé de mise en forme est assez commun sur les nuages de mots-clefs et permet de bien distinguer les termes les plus utilisés de ceux qui le sont le moins par la majoration de la taille des caractères ; par conséquent si ce procédé est bel et

152 cf. annexe 1

bien utilisé sur Isidore, cela signifierait que les concepts les plus récurrents lors de la fouille de texte sont ceux apparaissant en plus gros dans le nuage d'enrichissements. Concernant l'ordre dans lequel sont rangés les enrichissements, s'il est certain qu'il ne s'agit ni d'un classement alphabétique ni d'un classement par référentiel, ni d'un classement en lien avec les variations typographiques sus-mentionnées, nous pouvons en déduire que cette disposition serait représentative de l'ordre d'apparition des différents concepts dans le document, avec une priorité donnée aux concepts identifiés dans le texte intégral sur ceux présents dans les métadonnées. Cependant, n'ayant pas pu nous entretenir avec un responsable d'Isidore, le fruit de nos observations demeure au stade de conjectures.

Illustration 3: enrichissements d'une notice sur Isidore

Source : Isidore. OpenEdition, La scène punk à Poitiers (1976-2016). *ISIDORE – accès aux données et services numériques des SHS*, [en ligne]. [Consulté le 15 août 2018]. Disponible à l'adresse : https://rechercheisidore.fr/search/resource/?uri=10670/1.aphqrr

Lorsque nous cliquons sur l'un des concepts de la section, Isidore n'opère pas de transfert automatique vers la page de résultats correspondant à la recherche avec le concept sélectionné comme mot-clef, mais déploie un nouvel onglet détaillant le concept. Cet onglet comprend au minimum le nom du référentiel d'où est issu le concept sélectionné ainsi que la forme préférentielle retenue pour une ou plusieurs langues. Selon le concept, ces informations peuvent également être complétées par une ou plusieurs formes alternatives, une définition, une ou plusieurs notes, ainsi qu'un ou plusieurs « concepts liés », c'est-à-dire

des alignements vers des concepts issus d'autres vocabulaires contrôlés représentés sous la forme d'un Uri ; notons que hormis pour les alignements, la mention de la langue figure pour chaque champ. Les concepts liés et les termes préférentiels sont les seules mentions cliquables de l'onglet : la première renvoie à la page du concept avec lequel l'alignement a été effectué, tandis que la seconde renvoie à une nouvelle page de résultats de recherche sur Isidore à partir du concept sélectionné.

I.III Les quatre langages documentaires exploités dans les enrichissements sémantiques francophones d'Isidore

Présentation de Rameau

Rameau est le langage d'indexation utilisé par la Bibliothèque nationale de France, développé depuis 1980, et mis à jour par un service dédié : le Centre national Rameau. Il s'agit d'une liste de vedettes-matière, c'est-à-dire un langage pré-coordonné reposant sur une logique de « vedettes » représentant des concepts. Tandis que la « tête de vedette » équivaut au concept principal de la ressource à indexer, il est possible de lui associer jusqu'à quatre subdivisions : de sujet, géographique, chronologique, de forme, afin d'en améliorer la pertinence. A la fin du mois de décembre 2017 Rameau était composé de 186821 vedettes.[153]

Sa syntaxe complexe étant peu adaptée au contexte du web, et notamment du web des données, un vaste chantier de restructuration du langage a été entrepris dès 2016, pour la partie réflexive, et se prolonge sur la période 2017-2022 pour la partie mise en œuvre.[154] La réforme de Rameau s'appuie sur les préconisations du rapport du groupe de travail dédié qui s'articulent autour de trois grands axes : transformer les entités de genre et de forme, de lieux et de temps en référentiels distincts afin de les rendre utilisables à l'indexation et dans le cadre du Fichier National des Entités, simplifier les règles syntaxiques afin de rendre l'utilisation de Rameau plus rapide et aisée, et enfin mener la réforme en cohérence avec le modèle LRM (*Library Reference Model*) mis en place par l'Ifla (*International Federation of Library Associations and Institutions* – Fédération internationale des associations et

153 Centre national RAMEAU. Statistiques d'accroissement du référentiel RAMEAU. *BNF – RAMEAU*, [en ligne]. Décembre 2017. [Consulté le 14 août 2018]. Disponible à l'adresse : http://rameau.bnf.fr/informations/chiffres.htm

154 Centre national RAMEAU. Réformer RAMEAU. *BNF – RAMEAU*, [en ligne]. Mise à jour le 03 mai 2018. [Consulté le 14 août 2018]. Disponible à l'adresse : http://rameau.bnf.fr/chantier_syntaxe/intro.html

institutions de bibliothèques).[155] En France, le passage au modèle LRM s'inscrit dans le contexte de la « Transition bibliographique », un programme national lancé en 2014 afin de permettre aux catalogues des bibliothèques d'obtenir de la visibilité sur le web par l'ouverture et l'interconnexion des silos de données, suivant ainsi la logique du web des données.[156] La réforme Rameau s'inscrit donc en partie dans ce contexte notamment dans la mesure où la création de référentiels dédiés au lieu et au temps n'est pas sans rappeler les deux champs du même nom présent dans le modèle LRM. Par soucis de cohérence, le groupe de travail « Concepts, Lieux, Temps » qui pilote la réforme Rameau a donc été inclus au groupe « Normalisation » du programme Transition bibliographique.[157]

Rameau est un langage documentaire créé et exploité à la Bibliothèque nationale de France au format Intermarc, toutefois une conversion au format Skos a été effectuée en 2008 dans le cadre du projet européen Telplus[158] et est, depuis février 2012, en libre accès sur data.bnf. Isidore ne prenant en charge que les référentiels au format Skos, c'est donc cette version qui est utilisée par la plate-forme de recherche, ce qui, comme nous le verrons dans la partie suivante, n'est peut-être pas sans incidence sur la bonne exploitation de ce langage par la plate-forme de recherche.

Présentation de Gemet

Gemet (*GEneral Multilingual Environmental Thesaurus*) est un thésaurus multilingue sur le thème de l'environnement développé depuis 1996 par l'Agence européenne pour l'environnement. Se revendiquant « généraliste », il était composé en 2001 de 5298 concepts disponibles en treize langues.[159] Désormais presque intégralement traduit en 37 langues, la

155 MENARD, Florence. *Rapport du groupe de travail national sur la syntaxe de Rameau : préconisations et pistes d'évolution*, [en ligne]. Version 1.0. Paris : Centre national RAMEAU – Bibliothèque nationale de France, 05 mai 2017. 41 p. Disponible à l'adresse : http://rameau.bnf.fr/chantier_syntaxe/pdf/rapport_final_syntaxe_rameau.pdf

156 CANTIE, Philippe. Transition bibliographique : en avant marche ! *Bulletin des bibliothèques de France*, [en ligne]. 23 septembre 2015. [Consulté le 14 août 2018]. Disponible à l'adresse : http://bbf.enssib.fr/tour-d-horizon/transition-bibliographique-en-avant-marche_65461

157 MENARD, Florence et ROUSSEAUX, Olivier. Quand Rameau se greffe au programme national. *Arabesques*. 2017, Vol. 87. p.12. Egalement disponible en ligne à l'adresse : http://www.abes.fr/Publications-Evenements/Arabesques/Arabesques-n-87

158 BOUCHET, Thierry. Le vocabulaire Rameau en SKOS. *Journée nationale Rameau*, [en ligne]. 30 mai 2008. 13 diapositives. [Consulté le 14 août 2018]. Disponible à l'adresse : http://rameau.bnf.fr/informations/pdf/journee2008/rameau_skos.pdf

159 Eionet. About Gemet. *European Environment Information and Observation network – Eionet*, [en ligne]. Mise à jour le 29 août 2017. [Consulté le 14 août 2018]. Disponible à l'adresse :

version du thésaurus utilisée par Isidore est toutefois restreinte au français, à l'anglais et à l'espagnol, les trois langues disponibles sur la plate-forme, probablement pour éviter d'alourdir le détail des enrichissements en proposant pour chaque concept une liste d'au minimum 37 termes. Plus de 4000 concepts se sont vus attribuer une définition afin de faciliter leur compréhension et leur bonne réutilisation. Toutefois, contrairement aux termes, les définitions ne sont disponibles qu'en quatre langues : l'anglais, le bulgare, le russe et le slovène.[160] Si jamais la définition n'a pas été traduite dans la langue de l'interface de consultation du thésaurus, c'est par défaut la définition en anglais qui est utilisée. Par conséquent, Isidore exploitant la version française de Gemet, les définitions sont toutes en anglais.

Présentation de Pactols

Pactols est un thésaurus sur l'archéologie composé de sept micro-thésaurus auxquels il doit son nom : Peuples, Anthroponymes, Chronologie, Toponymes, Œuvres, Lieux, et Sujets. Il est développé par le réseau Frantiq (Fédération et ressources sur l'Antiquité) avec le soutien de la Maison de l'Orient et de la Méditerranée pour la partie logicielle et de l'Inist qui effectue les alignements géographiques du micro-thésaurus « Lieux » avec GeoNames et Wikidata.[161] Un groupe de travail dédié composé d'archéologues et de documentalistes est en charge de l'alimentation et de la mise à jour continue du thésaurus qui comporte à ce jour plus de 30000 concepts[162] traduits en sept langues. Nous pouvons à ce propos souligner le fait que les termes exprimés dans les sept langues, parmi lesquelles figurent le français, l'anglais et l'espagnol, sont présents dans la section « Enrichissements ? » d'Isidore, là où Gemet avait été restreint aux trois langues sus-mentionnées.

Présentation de GeoEthno

GeoEthno est un thésaurus géographique adapté au domaine de l'anthropologie, développé depuis 1985 par la bibliothèque Eric-de-Dampierre, une bibliothèque de recherche rattachée à l'université Paris Ouest Nanterre la Défense.[163] Le thésaurus revendique plus de 18000

http://www.eionet.europa.eu/gemet/fr/about/

160 *Ibid*.

161 FRANTIQ. Le thésaurus. *FRANTIQ : Fédération et ressources sur l'Antiquité*, [en ligne]. [Consulté le 16 août 2018]. Disponible à l'adresse : https://www.frantiq.fr/fr/thesaurus

162 *Ibid*.

163 Bibliothèque Eric-de-Dampierre. Thésaurus géographique GeoEthno – Bref historique du thésaurus. *Le site de la bibliothèque Eric de Dampierre*, [en ligne]. Mis à jour le 30 mars 2012. [Consulté le 16 août 2018]. Disponible à l'adresse : http://www.mae.u-paris10.fr/bibethno/spip.php?article20

« descripteurs »[164] disponibles en une ou plusieurs langues, généralement cinq ; tout comme pour Pactols, toutes les langues sont visibles sur Isidore.

Depuis 2016 le thésaurus comprend également une base dédiée aux ethnonymes, toutefois seule celle concernant les toponymes est exploitée sur Isidore.

II. L'analyse des enrichissements sémantiques

Comme expliqué précédemment, notre étude de cas porte principalement sur l'observation et l'analyse des enrichissements sémantiques de la section francophone « Enrichissements ? », autrement dit de l'utilisation qu'Isidore fait des langages documentaires Rameau, Gemet, Pactols et GeoEthno. Pour ce faire nous avons décidé de procéder à l'analyse de cinq corpus de documents indexés sur Isidore déterminés par les cinq « natures » de ressources disponibles sur le site : données bibliographiques, données de recherche, événements, publications, autres. Cependant entre le moment où nous avons entamé cette étude cas et celui où nous l'avons terminé, la nature de ressources « autres » a disparu de l'interface d'Isidore et a donc finalement été exclue de l'étude. En outre, initialement chaque corpus devait être composé de dix ressources afin de pouvoir faire des comparaisons au sein d'une même nature de ressource et ainsi avoir une idée d'ensemble des modalités et de la nature des enrichissements. Cependant, face à la redondance des observations et à la portée limitée de l'analyse qu'elles amenaient, celle-ci pouvant difficilement dépasser le stade des conjectures, nous avons fait le choix de réduire notre corpus à six documents pour chacune des quatre natures de ressources sus-mentionnées, ce qui fait un total de 24 documents enrichis observés.

Les ressources présentes sur Isidore sont publiées par plus d'une centaine d'organisations. Afin de pouvoir procéder à une analyse qui soit également comparative, nous avons décidé de privilégier des ressources issues de producteurs communs aux différentes natures de ressources. Cependant, sur Isidore, aucun producteur n'est commun aux quatre natures de ressources. Néanmoins nous avons pu identifier deux producteurs communs à trois sources : le « Centres de recherche et laboratoires en SHS » qui comprend les natures de ressources : données bibliographiques, données de recherche et événements, et « Criminocorpus.

164 *Ibid*.

Histoire de la justice, des crimes et des peines » qui possède des collections relatives aux natures de ressources : données bibliographiques, données de recherche et publications ; nous avons donc fait le choix de privilégier ces deux organismes producteurs dans notre corpus quitte à examiner plusieurs ressources provenant d'une même nature pour chacun de ces deux producteurs, en particulier lorsque la nature de ressource en question était peu adaptée à une comparaison trans-nature de sources. En effet, à titre d'exemple, les ressources de nature « événements » sont publiées par des organismes très spécifiques (seulement six) et qui, à l'exception du « Centres de recherche et laboratoires en SHS », sont tous propres à cette nature de sources, toutefois nous avons limité l'analyse des « événements » issus de ce producteur à trois ressources afin de pouvoir tout de même observer le traitement des enrichissements pour une nature de source donnée sans forcément tenir compte de l'établissement producteur.

Une grille d'analyse a été réalisée afin de faciliter l'étude du corpus. Elle se présente sous la forme d'un tableau composé de 25 lignes (ligne de titre puis une par ressource composant le corpus) et de dix-huit colonnes. Les quatre premières colonnes renvoient à des informations d'ordre général sur la ressource observé, les onze suivantes concernent explicitement la section « Enrichissements ? », et enfin, les trois dernières permettent d'ouvrir légèrement le périmètre de notre étude à des objets proches ou sur lesquels les enrichissements sémantiques pourraient, potentiellement, avoir un impact.

Détail des colonnes :

- nature de ressource ;
- producteur et collection ;
- nom de la ressource ;
- URL de la notice sur Isidore ;
- mots-clefs de la notice. Ce critère correspond à la section « Mots-clés » présente en haut à gauche de chaque notice. Nous faisons l'hypothèse que ces mots-clefs correspondent à ceux saisis lors de la création de la notice chez le producteur. Il

pourrait être intéressant d'observer s'il y a des doublons ou au contraire d'importants écarts entre ces mots-clefs et ceux de la section « Enrichissements ? » ;

- nombre d'enrichissements, c'est-à-dire le nombre de concepts présents dans la section « Enrichissements ? » pour chaque notice ;
- concepts Gemet, c'est-à-dire le nombre et les noms des concepts issus du thésaurus Gemet présents dans la section « Enrichissements ? » de la ressource analysée ;
- concepts communs entre les mots-clefs et Gemet ;
- concepts Pactols, c'est-à-dire le nombre et les noms des concepts issus du thésaurus Pactols présents dans la section « Enrichissements ? » de la ressource analysée ;
- concepts communs entre les mots-clefs et Pactols ;
- concepts Rameau, c'est-à-dire le nombre et les noms des concepts issus de la liste de vedettes-matières Rameau présents dans la section « Enrichissements ? » de la ressource analysée ;
- concepts communs entre les mots-clefs et Rameau ;
- concepts GeoEthno, c'est-à-dire le nombre et les noms des concepts issus du thésaurus GeoEthno présents dans la section « Enrichissements ? » de la ressource analysée ;
- concepts communs entre les mots-clefs et GeoEthno ;
- doublons dans les enrichissements. Ce critère consiste à énumérer tous les concepts d'intitulé identiques figurant dans la section « Enrichissements ? » d'une même notice, et à les analyser notamment en ce qui concerne leur portée sémantique, notamment lorsque celle-ci est balisée par une note d'application ou une définition, ou encore le nombre de ressources sur lequel chacun des doublons permet de rebondir ;
- entités GeoNames, c'est-à-dire leur nom et leur nombre. Ce critère sert surtout à établir une comparaison entre les concepts toponymiques de GeoEthno et les entités géographiques de GeoNames afin de repérer des doublons, des différences et comparer le nombre de ressources sur lesquels chacun permet de rebondir ;

- composition de la section « Pour aller plus loin... – Ressources ayant le(s) même(s) auteur(s) ». Ce critère cherche à établir si les enrichissements sémantiques, qu'il s'agisse des classifications, des enrichissements ou des espaces géographiques, ont un impact sur le choix des ressources suggérées ;
- composition de la section « Pour aller plus loin... – Suggestions de lecture ». Ce critère cherche à établir si les enrichissements sémantiques, qu'il s'agisse des classifications, des enrichissements ou des espaces géographiques, ont un impact sur le choix des ressources suggérées.

La première version de cette grille d'analyse comprenait également deux colonnes dédiées aux rebonds permis à partir des « sujets » de la section « Rebondir ? » et de la section « Enrichissements ? », l'objectif étant de pouvoir comparer les ressources proposées à partir de ces deux types de rebonds. Cependant nous avons très rapidement réalisé qu'il n'y avait aucune différence et que la plus-value de la section « Rebondir ? » venait de la possibilité de combiner des critères, sujets ou autres, ce qui n'est pas possible depuis la section « Enrichissements ? ». Ce critère n'étant par conséquent plus jugé pertinent, nous l'avons exclu de notre grille d'analyse définitive ; toutefois dans la mesure où cette section exploite les enrichissements sémantiques, nous aurons tout de même l'occasion d'y revenir dans la troisième partie de ce mémoire.

Sur Isidore, les ressources des contributeurs sont divisées en une ou plusieurs « collections » et à chacune de ces collections correspond une seule nature de ressource, toutefois un même producteur peut avoir plusieurs collections se référant à la même nature de document. Le choix des organismes de producteurs s'est fait selon le critère sus-mentionnés de recherche de producteurs ayant des collections renvoyant à différentes natures de ressources, cependant lorsqu'une même nature de ressource était présente, pour un même organisme, dans plusieurs collections, c'est le hasard qui a guidé notre choix. En effet, l'étude ne portant pas tant sur le contenu des notices que sur les enrichissements faits par Isidore, peu d'attention a été accordée aux spécificités, notamment thématiques, des ressources analysées. Nous avons tout de même pu, ponctuellement, utiliser comme critère d'analyse la

position du document dans les résultats de recherche pour tâcher d'enrichir notre analyse des suggestions de ressources, toutefois, nous verrons dans la troisième partie de ce mémoire, que ces tentatives ce sont toujours révélées infructueuses. Précisons également qu'un certain nombre de notices présentes sur Isidore ne contient aucun enrichissement sémantique, c'est donc logiquement que nous les avons exclues de l'étude.

Au regard de tous ces critères notre corpus était donc composé de :

- deux ressources produites par le Centre de documentation Regards : une en provenance de la collection de données bibliographiques « Banque de données documentaires REGARDS », l'autre en provenance de la collection de données de recherche « Collection numérisée de la cartothèque-photothèque REGARDS » ;
- six ressources du Centres de recherche et laboratoires en SHS : trois en provenance de la collection d'événements « Archéologie, Terre, Histoire, Sociétés », une issue de la collection de données de recherche « Chronique des fouilles en ligne », une de la collection de données bibliographiques « EUREL : Données sociologiques et juridiques sur la religion en Europe », et enfin la dernière est issus d'une autre collection de données bibliographiques « Système d'information en philosophie des sciences » ;
- trois ressources en provenance de Criminocorpus : une issue de la collection de données bibliographiques « Criminocorpus : bibliographie d'histoire de la justice française », deux de la collection de publications « Criminocorpus : musée d'histoire de la justice », et enfin une de la collection de données de la recherche « Criminocorpus : sources pour l'histoire de la justice, des crimes et des peines » ;
- deux ressources produites par l'Irstea (Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture) : une en provenance de la collection de données de la recherche « Gestion des territoires - CemOA », l'autre issue de la collection de publications « Sciences Eaux & Territoires - Irstea » ;
- deux ressources issues des Archives ouvertes : une de la collection de données de recherche « MédiHAL, l'archive ouverte de photographiques et d'images scientifiques », l'autre de la collection de publications « MoDyCo, Modèles, Dynamiques, Corpus - UMR 7114 » ;

- une ressource de l'ENS (École nationale supérieure) de Lyon issue de la collection de données bibliographiques « Atelier numérique de l'histoire » ;
- une ressource de la Maison des sciences de l'Homme de Dijon issue de la collection de données bibliographiques « CArGOS, catalogue de données géographiques en SHS » ;
- une ressource de la Maison des sciences de l'Homme de Clermont-Ferrand issue de la collection de publications « K@iros » ;
- une ressource en provenance de la collection éponyme d'événements de Calenda ;
- une ressource de type « événement » de l'unique collection des Actualités des Écoles françaises et Instituts français à l'étranger ;
- une ressource de type « événement » de l'unique collection de Fabula : la recherche en littérature ;
- deux ressources en provenance de GéoProdig, portail d'information géographique : une issue de la collection de publications « Atlas de l'UMR PRODIG », l'autre de la collection de données de recherche « Images des chercheurs de l'UMR PRODIG ».

III. Les entretiens

Afin d'enrichir notre analyse, nous avons choisi de compléter notre observation des enrichissements d'Isidore par des entretiens avec les professionnels en charge des référentiels francophones exploités par la plate-forme. Néanmoins, pour des raisons pratiques, nous n'avons pas pris contact avec les responsables de Gemet, l'Agence européenne pour l'environnement étant située à Copenhague au Danemark, il nous a semblé peu probable d'avoir un interlocuteur francophone et nous n'avons pas voulu prendre le risque de mener un entretien téléphonique ou par visioconférence en anglais par crainte de ne pas être en mesure de rebondir efficacement sur les déclarations de notre interlocuteur. En outre, nous avons fait le choix de nous adresser aux individus en charge de considérations techniques car ce n'est pas tant le cheminement intellectuel menant à l'alimentation des langages documentaires ou de la plate-forme Isidore, que leurs perspectives d'exploitation et d'évolution sur le plan informatique, qui nous intéressait. Nous avons donc sollicité un entretien téléphonique auprès de : Stéphane Pouyllau, co-directeur du projet Isidore, en

charge des aspects techniques, Thierry Bouchet, personnel du Centre national Rameau ayant été en charge de la conversion du langage au format Skos, Isabelle Donze, bibliothécaire à la bibliothèque Eric-de-Dampierre et responsable du thésaurus GeoEthno, et enfin Miled Rousset, responsable informatique du logiciel Opentheso à partir duquel Pactols est géré et responsable de la plate-forme « Têtes de réseaux documentaires » de la Maison de l'Orient et de la Méditerranée Jean Pouilloux.

Sur ces quatre sollicitations, seule celle de Stéphane Pouyllau est restée sans réponse, par conséquent la plupart de nos réflexions sur le fonctionnement d'Isidore sont restées au stade d'hypothèses, bien que cela ne les rendent pas pour autant caduques compte-tenu de nos observations et des quelques croisements ayant pu être faits avec les entretiens des trois autres professionnels.

Chacun des entretiens a été abordé comme un complément aux observations faites sur la section « Enrichissements ? » d'Isidore mais aussi comme une opportunité de prolonger notre réflexion sur les perspectives d'évolution de deux types de systèmes d'organisation des connaissances particuliers : les thésaurus et les listes de vedettes-matière. Les trois langages documentaires ayant fait l'objet d'un entretien ont chacun un contexte propres, les trois échanges ont donc fait l'objet de guides d'entretiens différents centrés sur des questions spécifiques à chaque contexte. Afin de faciliter la liberté de parole du répondant et les interactions, nous avons opté pour des entretiens semi-directifs structurés autour d'une grille comprenant les principaux thèmes et sous-thèmes à aborder dans l'entretien ainsi que quelques questions sous la forme de relances sur certains points précis. Par conséquent l'entretien de Thierry Bouchet du Centre national Rameau était axé sur la conversion de Rameau en Skos : la raison d'être de ce projet, l'utilisation qui en est faite à la BNF et son lien éventuel avec la réforme Rameau, l'entretien de Miled Rousset était centré sur le gestionnaire de thésaurus Opentheso : les spécificités de ce logiciel et le lien existant entre un système d'organisation des connaissances et son outil de gestion, enfin l'entretien d'Isabelle Donze abordait les perspectives d'évolution de GeoEthno et sa raison d'être dans Isidore. Bien que nous voulions éviter de poser des questions dont les réponses auraient été redondantes avec les informations que nous avons pu trouver sur chacun des systèmes

d'organisation des connaissances et du logiciel sus-mentionnés, il nous a tout de même semblé nécessaire de débuter chaque entretien par une question de contexte.

Tous les entretiens ont été menés par téléphone entre le 16 juillet et le 07 août 2018, ils ont été enregistrés numériquement puis ont fait l'objet d'une retranscription intégrale.[165] Bien que nous estimions leur durée à une vingtaine de minutes, celui d'Isabelle Donze pour GeoEthno a duré près du double ; si cela peut s'expliquer par un différent de personnalité, certaines personnes étant plus loquaces que d'autres, nous pensons surtout que cela est dû au fait qu'Isabelle Donze est la seule professionnelle que nous ayons interrogée ayant une vue d'ensemble du système d'organisation des connaissances qu'elle gère, aussi bien sur le plan technique qu'au niveau de son contenu et de sa participation à Isidore.

En outre, dans la mesure où nous n'avons pas pu nous entretenir avec un responsable d'Isidore afin d'en savoir plus sur les conditions en termes de contenu éditorial, d'ajout d'un référentiel à la section « Enrichissements ? » d'Isidore, il aurait pu être intéressant de compléter l'entretien de Miled Rousset par celui de l'un des gestionnaires de Pactols afin de savoir pourquoi est-ce qu'ils ont décidé d'intégrer le thésaurus à la plate-forme de recherche en sciences humaines et sociales. Concernant la version de Rameau en Skos, il s'agit d'un cas un peu différent, le vocabulaire étant librement téléchargeable depuis le site web data.bnf, Thierry Bouchet nous a expliqué ne rien savoir sur son utilisation par Isidore.

165 cf. annexes 2, 3 et 4

Partie 3 : étude de cas : présentation des résultats de l'observation

Cette troisième et ultime partie présente les résultats de nos observations concernant les enrichissements sémantiques automatiques de notices documentaires présentes sur la plate-forme de recherche Isidore par l'exploitation des quatre référentiels francophones Gemet, Rameau, Pactols et GeoEthno. Afin d'élargir notre propos et d'étayer notre analyse, celle-ci est complétée par les propos recueillis auprès de professionnels, comme cela a été expliqué dans la partie précédente.

Avant de commencer, rappelons qu'Isidore intègre en permanence de nouvelles ressources, il est donc possible que les chiffres cités dans les prochains paragraphes concernant le nombre de résultats obtenus par rebond à partir des termes de la section « Enrichissements ? » ne soient plus d'actualité. Toutefois, compte-tenu du fait qu'Isidore recense à ce jour presque six millions de ressources, les écarts de quelques dizaines de résultats que nous avons pu observer nous ont semblé dérisoires et sans incidence sur nos analyses.

I. Créer du lien entre les données : rebonds et alignements

I.I Les alignements

Comme nous l'avons expliqué dans notre seconde partie, la section « Enrichissements ? » d'Isidore affiche les alignements entre le référentiel exploité et d'autres vocabulaires. Cependant nous avons pu remarquer qu'en ce qui concerne Rameau, si les alignements avec le « Thésaurus W » du ministère de la Culture et « Agrovoc » apparaissent bien, cela n'est pas le cas de ceux qui ont pu être faits avec le LCSH et certains autres vocabulaires. Tout d'abord rappelons que les alignements n'obéissent pas à une logique de réciprocité : par exemple le fait qu'un concept issu de Rameau soit aligné avec un du Thésaurus W, ne signifie pas que l'alignement sera également fait sur le thésaurus du ministère de la Culture vers la liste de vedettes-matière de la Bibliothèque nationale de France. Or, dans le cas du LCSH, dans la mesure où ce système de vedettes-matière est également présent sur Isidore, il est intéressant de noter que les alignements avec des concepts issus de Rameau apparaissent

lorsque nous sélectionnons des concepts issus du vocabulaire de la bibliothèque du Congrès. Si nous prenons l'exemple du concept Rameau « Quatorzième siècle » utilisé sur Isidore nous pouvons voir que l'onglet de détail contient la forme préférentielle et les différentes formes alternatives du concept ainsi qu'une note et un alignement vers un concept du Thésaurus W, or, lorsque nous examinons le détail du concept depuis data.bnf, nous pouvons remarquer que celui-ci est aligné avec d'autres vocabulaire que celui du ministère de la Culture, notamment le LCSH. Cela pourrait s'expliquer par le fait que lorsque l'on regarde la page de code du concept Rameau au format RDF-XML, le concept du Thésaurus W figure en « *Skos:exactMatch* » et celui du LCSH, et des autres vocabulaires, en « *Skos:closeMatch* »[166], nous pourrions donc imaginer qu'Isidore ne référence que les alignements exacts. Cependant cette hypothèse est contredite par le fait que si nous prenons l'exemple du concept LCSH « *Feathers* » Isidore affiche, entre autres, un alignement vers le concept Rameau « Plumes », or lorsque nous examinons le code RDF du concept sur le site du LCSH, le concept Rameau sus-mentionné figure en tant que « *Skos:closeMatch* ».[167] Une autre hypothèse que nous pouvons faire concernant les alignements est qu'Isidore ne conserve les alignements que lorsqu'ils sont effectués avec un vocabulaire francophone : le Thésaurus W est en français, et Agrovoc, tout comme Eurovoc parfois utilisé par Gemet, possèdent également une version française contrairement au LCSH. Toutefois cette hypothèse est à considérer avec prudence car nous n'avons pu observer qu'une infime partie des alignements présents sur Isidore. La dernière hypothèse que nous pouvons formuler est que la version de Rameau utilisée par Isidore n'est pas la même que celle présente sur data.bnf, il pourrait s'agir d'une version antérieure comprenant des alignements différents. Thierry Bouchet nous a expliqué que les enrichissements avec le LCSH, entre autres, ont été effectués dans le cadre du projet Telplus ayant mené à la conversion de Rameau en Skos, or, Rameau au format Skos n'est disponible sur data.bnf que depuis 2012, soit un an après le lancement officiel d'Isidore, il était auparavant accessible sur un site dédié déployé dans le cadre du projet Stitch porté par TelPlus : « *RAMEAU subject headings as SKOS linked data* ».[168] Cependant, étant donné que les alignements ont été effectués dans le cadre de ce projet, il est fortement possible que la version de Rameau pouvant alors être récupérée comprenait déjà des alignements avec le

166 cf. annexe 5

167 cf. annexe 6

168 TelPlus. RAMEAU subject headings as SKOS linked data. *STITCH Project*, [en ligne]. [Consulté le 28 août 2018]. Disponible à l'adresse : https://www.cs.vu.nl/STITCH/rameau/index-fr.html

LCSH, l'hypothèse de l'incidence de la version utilisée par Isidore pour expliquer la présence ou l'absence d'un certain nombre d'alignement semble dès lors peu plausible. De plus, dans la mesure où les systèmes d'organisation des connaissances sont des objets en constante évolution afin de rester à jour et en adéquation avec des besoins qui se transforment en permanence, il serait étonnant que la version de Rameau utilisée en 2018 sur Isidore date d'il y a plus de six ans ; rappelons, à titre de comparatif, qu'Isabelle Donze a expliqué transmettre une version actualisée de GeoEthno tous les ans ou tous les deux ans aux gestionnaires d'Isidore.

En outre, nous pouvons nous interroger sur l'exploitation qui est faite de ces alignements sur Isidore et sur l'intérêt d'une telle pratique pour relier deux systèmes d'organisation des connaissances. Comme nous avons pu l'expliquer précédemment, l'exploitation qu'Isidore fait des concepts alignés se limite à l'affichage du lien cliquable vers le concept cible, aucune de nos observations n'a permis de mettre en exergue d'autres façons d'exploiter ces alignements. Néanmoins certains point ont attiré notre attention concernant le multilinguisme et la potentielle exploitation pouvant être faite d'alignements afin d'assurer la représentation de concepts issus d'un vocabulaire monolingue sur un document dans une autre langue. En effet, si nous prenons l'exemple de la notice documentaire intitulée sur Isidore : « *D 7.2 Design and Sustainability Plan for an Open Humanities Data Platform* », nous pouvons remarquer que ce document est entièrement en anglais aussi bien en ce qui concerne le texte intégral que le contenu des métadonnées, mots-clefs et le résumé, par conséquent il est logique que le seul concept issu du vocabulaire monolingue français « Rameau » présent dans la notice soit « Vente jumelée » puisqu'il s'agit de l'un des très rares concepts de ce vocabulaire à posséder un terme alternatif en anglais, en l'occurrence « *Package* ». Une rapide recherche sur le texte intégral et la notice originale sur Hal SHS grâce au raccourci clavier CTRL+F nous a permis de constater que le terme « *package* » est bel et bien utilisé dans le document mais avec une signification toute autre, ce qui pose la question de la compréhension sémantique des concepts identifiés grâce à la fouille de texte, sur laquelle nous reviendrons un peu plus tard. Toutefois dans ce cas précis ce n'est pas tant l'aspect sémantique qui nous intéresse que l'exploitation des données liées qui s'est peut-être faite. En effet lorsque nous nous intéressons aux enrichissements en anglais, nous pouvons remarquer que parmi les concepts issus du LCSH figure « *Bundling (Marketing)* », s'il

y a une nouvelle fois un problème de compréhension sémantique, il est intéressant de souligner que la recherche sur les termes « *bundling* » et « *marketing* » dans les métadonnées et le texte intégral à l'aide de CTRL+F n'a cette fois-ci donnée aucun résultat. Il est donc envisageable de faire l'hypothèse que ce concept n'a pas pu être identifié par une technologie de fouille de texte, or, nous pouvons également remarquer que dans le détail des enrichissements relatifs à ce concept donné sur Isidore, le concept Rameau « Vente jumelée » figure en tant que concept lié. De plus, pour cette même ressource, Isidore a associé aux enrichissements en espagnol un seul et unique concept issu du vocabulaire, pourtant monolingue, de la Bibliothèque nationale d'Espagne : « *Venta conjuta* » (pouvant être traduit en français par « Vente jumelée »), avec la forme alternative espagnole « Bundling (Marketing) » ; or, tout comme « bundling » et « marketing », les chaînes de caractères « venta » et « conjuta » ne semblent pas figurer dans le texte intégral ou les métadonnées du document. Cependant lorsque nous regardons le code RDF-XML de chacun des trois concepts : « Vente jumelée » pour Rameau, « *Bundling (Marketing)* » pour LCSH et « *Venta conjuta* » pour la Bibliothèque nationale d'Espagne nous pouvons remarquer que des alignements ont été effectués entre ces trois vocabulaires[169]. En effet, alors qu'Isidore n'affiche que l'alignement entre le concept LCSH et celui de Rameau, l'observation du code nous permet d'affirmer que le concept Rameau est pourtant aligné en « *Skos :closeMatch* » aussi bien sur le concept de LCSH que sur celui de la BNE, tandis que « *Venta conjuta* » est également aligné en « *Skos :closeMatch* » sur « *Bundling (Marketing)* » ainsi que sur « Vente jumelée ». Au regard de tous ces éléments, nous pouvons faire l'hypothèse que même si Isidore n'affiche pas les alignements entre vocabulaires ceux-ci sont tout de même exploités par les liens d'équivalence entre concepts afin d'élargir les enrichissements en permettant à des vocabulaires monolingues d'être utilisés pour agrémenter des ressources n'étant pas dans leur langue. Toutefois rappelons que nous formulons-là une simple hypothèse, basée exclusivement sur des observations, et qui n'a pas été vérifiée, notamment en terme de faisabilité sur le plan technique.

Enfin, précisons que la littérature aussi bien professionnelle que scientifique est avare en retours d'expérience sur les raisons ayant menées à l'alignement de différents systèmes

169 cf. annexe 7

d'organisation des connaissances et l'exploitation qui en est faite, préférant se concentrer sur les considérations techniques ayant rendus ces alignements possibles.

I.II Les rebonds

Concernant les rebonds pouvant être effectué à partir des enrichissements d'une notice documentaire sur Isidore sur le reste du catalogue de la plate-forme, contrairement aux résultats pouvant être obtenus en passant par la section « Rebondir ? » qui peuvent faire l'objet du croisement des différentes facettes de la section, la page de résultats accessible à partir de l'onglet « Enrichissements ? » est limitée au seul concept sélectionné. La recherche ayant permis d'arriver au document sur lequel le concept a été sélectionné n'étant pas gardée en mémoire, l'utilisation des facettes pour restreindre le périmètre de la recherche se fait donc directement depuis la page de résultats. Concernant l'ergonomie de l'interface de recherche, bien que ces considérations ne soient pas au cœur des problématiques de notre étude de cas, précisons tout de même qu'une fois un rebond effectué, aussi bien à partir de la section « Rebondir ? » que de la section « Enrichissements ? » le rappel du ou des critères de recherche sélectionnés n'apparaît pas de façon précise dans la nouvelle page de résultats. En effet, si une nouvelle recherche a été lancée à partir de la section « Rebondir ? » de, par exemple, la notice « La scène punk à Poitiers (1976-2016) »[170] en sélectionnant : la même source, le même type de document, l'une des trois disciplines, l'une des trois catégories et quatre des huit « sujets », ou « enrichissements », de cette notice. La page de résultats qui en découle semble proposer des résultats pertinents compte-tenu des critères sélectionnés, toutefois c'est l'intitulé du critère, et non de son contenu, qui est affiché en guise de rappel juste avant la liste des résultats.[171] En outre, si l'intitulé du contenu du critère « types de ressource », en l'occurrence « Colloques et conférences » a bien été automatiquement coché dans la liste de facettes de la section « Affiner » situé à gauche de la page lorsque nous avons lancé la recherche, il n'en va pas de même pour la discipline « Musique, musicologie et arts de la scène » ou la catégorie « Esprit et langage | Représentation » qui n'ont pas été automatiquement cochées bien que faisant partie des facettes proposées par la colonne « Affiner ». Le critère de « sujets » n'est lui jamais présent parmi les facettes dans la colonne « Affiner », probablement car les systèmes d'organisation des connaissances exploités

170 Isidore. La scène punk à Poitiers (1976-2016). *ISIDORE – accès aux données et services numériques des SHS*, [en ligne]. [Consulté le 15 août 2018]. Disponible à l'adresse : https://rechercheisidore.fr/search/resource/?uri=10670/1.aphqrr

171 cf. annexe 7

contiennent *a minima* plusieurs centaines, voire plusieurs milliers, de concepts, il est donc impensable de les proposer comme critères de recherche sélectionnables ; toutefois ils peuvent l'être à partir de la barre de recherche qui fonctionne sur le principe de l'auto-complétion pour les auteurs et les mots-clefs[172]. Il est par ailleurs intéressant d'observer qu'une fois sur la page de résultats le mot-clef sélectionné dans la barre de recherche, par auto-complétion ou non, est explicitement retranscrit juste avant la liste des résultats, sous la forme « Mot-clé : xx »[173]. Ce critère « Mot-clé » semble être issu de la fusion des mots-clefs des notices documentaires telles qu'indexées par les professionnels des organismes producteurs, et des enrichissements de la section éponyme. De plus, les « disciplines », les « catégories » et les « sujets » étant issus de différents référentiels indépendants du projet et non pas du travail de normalisation effectué par Isidore suite au moissonnage des notices documentaires, peut-être que ce critère est à prendre en considération dans l'observation que nous venons d'effectuer.

Notons également que nous avons relancé exactement la même recherche à quelques minutes d'intervalle et que l'ordre de certains des résultats proposés, à commencer par le premier, était différent. Le même phénomène s'est produit lorsque nous avons cliqué une nouvelle fois sur le lien hypertexte de la première recherche que nous avions faite à partir de ces critères : l'ordre des résultats avait de nouveau changé. Cependant, précisons qu'une fois l'un des critères de tri d'affichage sélectionné, par exemple le tri par pertinence, l'ordre des résultats devient fixe. Le même problème se présente lorsque nous cliquons sur les termes préférentiels pour chaque langue sur un concept donné : le nombre de résultats affiché est exactement le même quelle que soit la langue, ce qui semble logique dans la mesure où le concept reste inchangé peu importe le terme qui le retranscrit en langage humain et que c'est l'Uri du concept qui est exploité par le moteur d'Isidore, mais l'ordre des résultats varie totalement d'une langue à l'autre. Nous avons tout d'abord pensé que ces différences étaient dues au fait qu'Isidore faisait remonter en priorité les documents dans la langue sélectionnée ou en lien avec les régions du monde dans lesquels la langue sélectionnée est parlée ; cependant une rapide vérification de la première page de résultats a tôt fait d'invalider ces hypothèses. En revanche, une fois encore, dès que nous sélectionnons un mode de tri sur la

172 cf. annexe 9
173 cf. annexe 9

page de résultats, ces derniers apparaissent exactement dans le même ordre peu importe le terme préférentiel sélectionné.

Enfin, si nous nous intéressons à présent à la section « Pour aller plus loin... », nous avions expliqué, en présentant la constitution d'une notice documentaire sur Isidore, qu'elle est divisée en deux onglets : un consacré aux suggestions de ressources écrites par le même auteur et un autre dédié à des suggestions de ressources sans qu'aucune mention ne soit faite de leur critère de sélection, nous nous étions alors demandé si les enrichissements étaient exploités pour faire ces suggestions. L'analyse de notre corpus semble montrer que le choix des suggestions du même auteur est basé sur les résultats du tri par pertinence de la page de résultats de recherche dédiée à l'auteur en question, sans mention d'aucun autre critère de recherche, toutefois nous ignorons quels sont les informations utilisées pour établir ce tri par pertinence. Concernant le second onglet, « Suggestions de lecture », nous n'avons pas été en mesure d'identifier le moindre critère pouvant expliquer comment ces suggestions sont sélectionnées, les auteurs, mots-clefs, classifications, enrichissements, etc. différant totalement d'une ressource à ses suggestions, un peu comme si qu'elles étaient choisies de façon aléatoire par l'algorithme d'Isidore afin d'éveiller la curiosité du chercheur d'information.

II. Les doublons

En dépit du caractère spécialisé du champ thématique recouvert par chacun des systèmes d'organisation des connaissances exploité, force est de constater que de nombreux doublons et triplons, voire même quadruplons, figurent dans la section « Enrichissements ? » puisqu'un même terme préférentiel peut être utilisé dans différents vocabulaires contrôlés.

Bien que la norme Iso 25964 interdise l'utilisation d'un même terme pour représenter plusieurs concepts, il arrive que cette règle ne soit pas respectée par les gestionnaires de thésaurus. En effet, si nous prenons l'exemple de la notice « Essai de restauration de roselières en marais dulçaquicole »[174] nous pouvons observer que le concept « marais » est présent à quatre reprises dans la section « Enrichissement ? » : une fois en provenance de

174 Isidore. Essai de restauration de roselières en marais dulçaquicole. *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 24 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.7aj5go

Pactols, une autre de Rameau et les deux dernières sont issues de Gemet. Il est intéressant de constater que bien que le terme préférentiel en français soit le même pour les deux concepts, les termes en anglais et en espagnol ainsi que les traductions divergent, l'un de deux concepts fait même l'objet d'un alignement, signifiant bien qu'il s'agit de deux entités distinctes.[175] Ce doublon pourrait être expliqué par la définition vague du concept de « marais » ainsi que par l'ambiguïté existant dans la langue française entre « marais » et « marécage », ce dernier faisant l'objet de définitions divergentes selon les spécialistes[176], ce qui pourrait expliquer pourquoi est-ce que ce concept n'existe pas sur Gemet et que le terme de « marais » aurait donc été privilégier à chaque fois. Toutefois, nous pouvons observer sur Gemet que le terme préférentiel est identique entre les deux concepts « marais » pour dix-sept des 37 langues disponibles[177] tandis que l'un des deux concepts n'existe pas pour deux langues. Il pourrait donc être intéressant de se demander si ce nombre important de doublons est bien dû, comme nous le pensions, à une ambiguïté existant entre plusieurs termes d'une même langue et sur lesquels la communauté scientifique n'a pas tranché, ou bien si ce phénomène est dû à un défaut de construction du concept sur Gemet.

Bien que ce soit les responsables d'Isidore qui décident des référentiels à ajouter à la plate-forme de recherche en sciences humaines et sociales, les producteurs de vocabulaires peuvent tout de même avoir un rôle à jouer sur la probabilité de doublons dans les enrichissements d'Isidore. En effet, si nous prenons l'exemple de GeoEthno, thésaurus initialement entièrement dédiés aux toponymes, il comprend, depuis 2016, également des ethnonymes mais qui ne figurent pas sur Isidore malgré la mise à jour annuelle ou bisannuelle du langage documentaire sur la plate-forme. Cet écart est dû au fait qu'Isidore intégrait déjà Rameau, qui est un vocabulaire à portée presque encyclopédique et comprenant donc déjà de nombreux ethnonymes se trouvant également sur GeoEthno, la gestionnaire du thésaurus nous a donc expliqué ne pas avoir jugé utile d'ajouter cette partie à Isidore, précisant : « je pense que ça poserait plus de problèmes de pertinence que ça n'en résoudrait ». Ainsi, même s'il peut exister une certaine complémentarité entre les

175 cf.annexe 10

176 cf. les définitions différentes du terme « marécage » issues de deux ouvrages de référence :

RAMADE, François. *Dictionnaire encyclopédique des sciences de la nature et de la biodiversité*. Paris : Dunod, 2008. p.360. ISBN 978-2-10-053670-2

TRIPLET, Patrick. *Dictionnaire encyclopédique de la diversité biologique et de la conservation de la nature*. 4ème édition. 2018. p.682. ISBN 978-2-9552171-9-1. Également disponible en ligne à l'adresse : https://www.dropbox.com/s/xdw5unlwancpao0/Dictionnaire%20conservation

177 cf.annexe 10

vocabulaires, le fait qu'Isidore soit restreint aux disciplines des sciences humaines et sociales réduit la diversité thématique des langages documentaires pouvant être utilisés ; par conséquent il n'y a rien d'étonnant à ce que à ce que les référentiels exploités par Isidore aient un certain nombre de concepts, ou plutôt de termes préférentiels, en commun.

II.I Les rebonds à partir de doublons

Il est intéressant de remarquer que pour un même intitulé de terme préférentiel, le rebond vers une nouvelle page de résultats de recherche peut donner un nombre de résultats différent selon le référentiel d'où le terme est extrait. Nous avons remarqué que les concepts issus de Rameau renvoient généralement à moins de résultats tandis que ceux de Pactols et Gemet, restent souvent proches, voire même sont identiques bien qu'il arrive que le concept issu de Pactols donne beaucoup plus de résultats que son homologue sur Gemet, comme par exemple le concept « logiciel » qui sur Pactols donne 26939 résultats contre 13386 pour Gemet. Toutefois nous avons pu remarquer un certain nombre d'exceptions, notamment sur le concept Rameau « Guerre mondiale (1914-1918) » pouvant être considéré comme équivalent au « première guerre mondiale » de Pactols mais permettant de rebondir sur 21176 résultats contre 8011 pour son homologue. Dans ce cas précis nous estimons que la différence de nombre de résultats peut s'expliquer par le fait que le concept de Rameau comprend également quatre formes alternatives en français, là où celui de Pactols est limité à un terme préférentiel en néerlandais et un autre en arabe, ce qui est d'ailleurs surprenant car lorsque nous consultons le thésaurus Pactols directement depuis sa plate-forme de diffusion, nous pouvons constater que le concept est bien disponible dans les sept langues habituelles du thésaurus[178] ; cependant peut-être que cela n'apparaît pas sur Isidore car les quatre autres termes préférentiels auraient été ajoutés après la dernière mise à jour de Pactols sur la plate-forme de recherche en sciences humaines et sociales. En effet, nos observations tendent à montrer que l'algorithme de fouille de texte d'Isidore identifie également les chaînes de caractères des termes alternatifs et les associe au concept auquel ils renvoient ; par conséquent un concept possédant de nombreux termes alternatifs devrait permettre de rebondir sur davantage de ressources qu'un concept ayant le même terme préférentiel mais aucun label alternatif. Néanmoins, le recours à des termes alternatifs pour préciser la portée d'un concept peut poser question en termes de précision et de pertinence.

178 cf.annexe 11

La même réflexion quant à la reconnaissance des chaînes de caractères peut être avancée en ce qui concerne le multilinguisme. En effet, comme nous l'avons expliqué en présentant Isidore, les ressources présentes sur la plate-forme ne sont pas limitées à la langue française, ni même à l'anglais et à l'espagnol, ce sont donc des ressources dans plusieurs dizaines de langues qui sont accessibles sur la plate-forme. D'ailleurs, la présence de documents en texte intégral dans de nombreuses langues qui ne sont disponibles sur aucun des référentiels exploités par Isidore pose question quant au choix des responsables de la plate-forme de moissonner des documents qui ne peuvent être enrichis qu'à partir de certaines de leur métadonnées et non de leur texte intégral, alors même que celui-ci est librement accessible, faisant alors perdre une plus-value certaine à la section « Enrichissements ? ». Pour en revenir à la prise en compte de la reconnaissance des chaînes de caractères dans différentes langues lors de la fouille de texte, celle-ci permettrait d'expliquer pourquoi est-ce que Pactols est bien souvent le référentiel à partir duquel le plus grand nombre de rebonds est possible alors même que les informations sur le concept sont généralement moins riches que sur Gemet ou Rameau notamment pour ce qui concerne les notes et les alignements. Par conséquent, avec ses termes préférentiels généralement disponibles en sept langues, Pactols bénéficierait donc pour quasiment chaque concept d'au minimum sept chaînes de caractères distinctes là où Gemet serait sur un minimum de trois et Rameau et GeoEthno sur un minimum de un. Nous pouvons illustrer notre hypothèse avec l'exemple du concept « sciences humaines » qui, pour Pactols, est composé des termes préférentiels dans les sept langues du thésaurus ainsi que de trois termes alternatifs en français, anglais et italien, pour un total de 47832 résultats par rebond contre 33917 pour Rameau qui possède tout de même trois formes alternatives en français, une note et un concept lié, et 35339 résultats pour Gemet avec le terme préférentiel en trois langues ainsi qu'une définition, pourtant large, mais comme nous le verrons plus en détail dans la sous-partie suivante, celles-ci ne semblent pas prise en considération par l'algorithme d'Isidore. Pourtant, paradoxalement, le concept « plan » issu de Gemet renvoie à un plus grand nombre de résultats que celui de Pactols qui est pourtant davantage traduit (83722 résultats pour Gemet contre 74015 pour Pactols). Partant de ce constat nous pouvons alors formuler deux hypothèses : soit le fait que le concept issu de Pactols ait une définition plus restrictive et en français a eu une incidence sur sa compréhension par l'algorithme d'Isidore, mais cela nous semble peu probable compte-tenu des erreurs de compréhension sémantique que nous avons pu identifier et qui

auraient pu être évitées par l'exploitation des définitions associées aux concepts, soit cette différence est due au fait que ce n'est pas le même terme préférentiel en espagnol qui a été utilisé par Pactols et par Gemet. Néanmoins si nous suivons cette seconde hypothèse, c'est bien le concept issu de Pactols qui devrait avoir le plus rebonds puisque le terme en espagnol est « *plano* », donc une nouvelle chaîne de caractère, alors que celui de Gemet est toujours « plan », soit la même chaîne de caractère que pour les versions françaises et anglaises. Cet exemple est représentatif du fait qu'en l'absence de connaissance du fonctionnement de l'algorithme d'enrichissement d'Isidore, il nous est impossible d'avoir la moindre certitude quant aux informations prises, ou non, en considération.

Nous nous sommes également demandés si la présence d'alignements avec d'autres vocabulaires avait une influence sur le nombre de rebonds. Rien dans nos observations ne nous permet d'établir un lien de causalité entre ces deux éléments et cela n'est pas tellement surprenant compte-tenu du fait qu'aucun des vocabulaires avec lesquels des alignements ont été effectués ne figurent parmi les référentiels exploités par Isidore. En effet, nous pourrions imaginer que le fait de lier les langages documentaires dans la logique du web des données permette par extension de faire des ponts entre les ressources qu'ils indexent ; par exemple nous pourrions supposer qu'une ressource soit indexée avec un concept issu du thésaurus Gemet, que le concept en question soit aligné sur un d'Agrovoc et qu'à partir de là l'utilisateur ait également accès aux ressources indexées à partir de ce concept issu d'Agrovoc. Toutefois la faisabilité de cette hypothèse semble compromise dès lors qu'un vocabulaire est utilisé par plusieurs organismes principalement en raison de la quantité massive de données que cela représente. En effet, si nous prenons l'exemple même d'Isidore aucun des référentiels exploités n'a été conçu pour fonctionner exclusivement avec la plate-forme de recherche en sciences humaines et sociales, par conséquent si nous imaginons qu'un concept d'un vocabulaire x est aligné sur un concept de Rameau, cela signifierait qu'à partir du vocabulaire x, l'utilisateur aurait accès à l'ensemble des ressources indexées avec ce concept, qu'il s'agisse des ressources de la Bibliothèque nationale de France, d'Isidore ou de tout autre organisme en France et à l'étranger ayant utilisé ce concept. Or, même en excluant les considérations techniques, cela poserait un important problème en termes de surcharge cognitive pour l'utilisateur et de pertinence des résultats ;

en effet, un concept d'indexation peut difficilement être appréhendé sans contexte, par exemple le concept « littoral » peut aussi bien être utilisé pour indexer des documents sur la protection de l'environnement que sur l'anthropisation ou l'histoire de territoires en bord de mer. Nous pourrions donc plutôt envisager que l'exploitation de catalogues, liés dans le web des données par l'intermédiaire de leur langage d'indexation, se fasse plutôt à l'échelle de réseaux travaillant sur des thématiques communes, afin de limiter le bruit documentaire. Cependant, gardons à l'esprit qu'une fois encore, cela reste une hypothèse dont nous ignorons la faisabilité sur le plan technique.

Outre le nombre de rebonds, il est également intéressant de s'intéresser aux ressources proposées par les concepts en doublons, en particulier lorsque ceux-ci n'ont pas le même sens. En effet, si nous prenons l'exemple de la ressource « Les ressources en eau et le changement climatique en Provence-Alpes-Côte d'Azur »[179], le concept « territoire » est présent deux fois dans la section « Enrichissements ? », une fois pour Gemet, la deuxième pour Pactols. Or, lorsque nous lisons les définitions associées à ces deux concepts nous pouvons constater qu'elles divergent, donnant au concept deux périmètres bien distincts : pour Gemet celui de la zone géographique dans laquelle un animal vit et pour Pactols celui des circonscriptions juridiques. La prise en compte de la sémantique à travers les définitions devrait donc mener à des rebonds sur des ressources différentes. Cependant comme nous avons pu le constater précédemment, les définitions ne semblent pas prises en compte par l'algorithme d'Isidore, ce qui semble une nouvelle fois se vérifier puisque au minimum les deux premières pages de résultats, triés par pertinence, par rebonds à partir de ces deux concepts sont identiques. Un constat similaire peut être fait sur la même notice avec le concept « culture » issu de Gemet, Pactols et Rameau ; là où Pactols et Rameau proposent une définition similaire, celle de Gemet diffère totalement mais il n'empêche que la première page de résultats de rebonds est identique pour les trois concepts, les erreurs de compréhension sémantique ne semblent donc pas limitées à une poignée de notices isolées. Or, ces erreurs de compréhension de la sémantique des concepts sur une plate-forme telle qu'Isidore, qui exploite le web des données pour créer des liens, supposés de sens, entre des

179 Isidore. Les ressources en eau et le changement climatique en Provence-Alpes-Côte d'Azur. *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 24 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.hb75cp

ressources issues de silos de données divers est problématique dans la mesure où certain des liens créés n'ont pas lieu d'être et contribuent au bruit documentaire au lieu de le diminuer et de faciliter la recherche d'information.

Jusqu'à présent nous avons très peu mentionné GeoEthno car, s'agissant d'un vocabulaire restreint aux entités géographiques, nous n'avons trouvé pratiquement aucun doublons entre des concepts issus de ce thésaurus et ceux des autres référentiels de la section « Enrichissements ? ». En effet, les doublons sur les concepts géographiques se font plutôt entre GeoEthno et GeoNames, le référentiel utilisé dans la section « Espaces géographiques ». Il nous a été confirmé par Isabelle Donze que GeoNames fonctionne sur une logique totalement différente des langages documentaires sus-mentionnés, nous avons donc préféré de pas émettre d'hypothèses sur la façon dont ce référentiel est mis en lien avec les notices sur Isidore. En terme de présentation sur l'interface, la section « Espace géographique » est fortement similaire à « Enrichissements ? » avec la liste des concepts et un détail comprenant : forme préférentielle cliquable permettant de faire des rebonds sur Isidore, formes alternatives et, éventuellement, concept lié. Par conséquent, à première vue, le seul élément distinguant un concept issu de GeoEthno d'un en provenance de GeoNames serait le fait que les concepts de GeoNames sont géolocalisés et indiqués sur une carte. Toutefois , lorsque nous observons les doublons nous pouvons constater que GeoNames permet généralement un plus grand nombre de rebonds malgré le fait que les concept sont moins traduits ; cependant il n'est pas rare que la langue du terme préférentiel issu de GeoNames ne soit pas spécifié, par conséquent le critère du nombre de traductions nous semble ici peu pertinent.

Les entités géographiques issues de GeoNames font elles aussi parfois l'objet d'erreur de compréhension sémantique. Celle que nous avons pu constater concerne les points cardinaux et a été observée sur les notices « Les ressources en eau et le changement climatique en Provence-Alpes-Côte d'Azur » et « ODIT - Section A2 des plans cadastraux napoléoniens de la commune de Gennes ». En effet, dès lors que les termes représentant un concept GeoNames se limitent à un point cardinal, la géolocalisation l'associe au pays « Islande » alors même que celui-ci n'est pas mentionné dans la ressource et n'a aucun lien

avec celle-ci.[180] Isabelle Donze nous a également expliqué qu'Isidore n'exploite pas l'intégralité de GeoNames car le système est trop lourd, nous pouvons donc nous demander si cette erreur, la seule présente dans notre corpus, ne serait pas due au fait qu'Isidore n'utilise pas l'ensemble du référentiel.

III. La compréhension de la sémantique

L'exploitation des différents référentiels est supposée permettre de relier des documents issus de silos isolés suivant la logique du web des données et ainsi permettre à l'utilisateur d'accéder facilement à l'ensemble des ressources présentes sur Isidore sur un sujet donné. Toutefois, la réussite d'un tel projet nécessite une compréhension parfaite de la sémantique de chaque concept de chaque référentiel afin de passer outre la polysémie, la synonymie, et permettre la prise en compte des limites d'applications explicitement mentionnées sous la forme de notes associées aux concepts. Or, comme cela nous a été précisé, une façon de limiter ces erreurs de compréhension sémantique serait d'exploiter la richesse des concepts :

> « on a le terme générique, on a les synonymes, les termes non préférentiels, on a les termes associés, on a des notes, tout ça c'est autour d'un concept. Il faut exploiter cet environnement pour être sûr et certain du contexte qu'il faut appliquer à ces concepts. »[181]

Toutefois rappelons que si, en théorie, un concept peut effectivement être extrêmement détaillé, dans les faits tous ne bénéficient pas du même niveau de précision, bien que la multiplication des gestionnaires de thésaurus informatisés respectant la norme Iso 25964 facilite grandement la saisie de toutes les informations pouvant enrichir un concept.

Un système d'organisation des connaissances, quel qu'il soit, est toujours conçu dans un contexte spécifique et pour un usage particulier. Si cet usage peut être amené à évoluer, notamment par les évolutions technologiques qui ont permis d'imaginer de nouvelles utilisations des langages documentaires dans les applications d'informatique documentaire, adapter un Soc à de nouvelles technologies et/ou à des contextes différents de son contexte d'origine n'est pas anodin. En effet, ces adaptations représentent un coût humain et financier non négligeables puisque du personnel formé doit être mobilisé pour les mener à bien, mais interroge également quant à une éventuelle perte de richesse du système d'organisation des

180 cf. annexe 12
181 cf. annexe 2

connaissances modifié.[182] De plus, le formalisme Skos présente plusieurs limites. La première vient de son essence même de « système simple d'organisation des connaissances » qui en fait un format « basique » incapable de gérer la complexité de certains langages documentaires. En effet, bien que Skos se présente comme un langage adapté à tous types de systèmes d'organisation des connaissances, allant du thésaurus aux classifications en passant par les taxonomies et les listes de vedettes-matière[183], dans la pratique, les professionnels de la documentation trouvent son champ plus restreint. De fait, si nous prenons l'exemple de Rameau qui est, à l'origine, une liste de vedettes-matière au format Intermarc, Thierry Bouchet nous a expliqué que la conversion du langage au format Skos a été faite « *a minima* » et ne sera jamais exploitée par la Bibliothèque nationale de France pour faire de l'indexation. En effet, « Skos ne gère pas la syntaxe de Rameau »[184] dans le sens où ce format ne permet pas d'élaborer un langage pré-coordonné, c'est-à-dire d'assembler des concepts au moment de l'indexation en vue de créer un seul descripteur permettant de décrire l'intégralité du contenu d'une ressource. Toutefois la BNF assure tout de même une exploitation de Rameau au format Skos par l'intermédiaire de leur site web data.bnf sur lequel la totalité des concepts du langage documentaire est librement consultable, téléchargeable et réutilisable. Néanmoins, au regard de ces informations, nous pouvons nous interroger sur la pertinence de convertir un langage pré-coordonné au format Skos pour une organisation de moyenne ou petite taille qui ne disposerait pas de ressources comparables à celles de la Bibliothèque nationale de France.

Une seconde limite du format Skos vient du fait qu'il a été développé en faible collaboration avec les professionnels de l'information et de la documentation. Par conséquent, certains manques ont pu être identifiés, notamment concernant la gestion de l'historicité des termes préférentiels :

> « C'est assez complexe de gérer l'historicité des descripteurs, c'est tout en aspect technique qui n'est d'ailleurs pas pris en compte dans la norme Skos et ça pose d'énormes problèmes parce ce que les gens qui ont conçu cette norme avaient plus un profil d'informaticien que de gestionnaire de thésaurus et donc il y a des choses qu'ils n'ont pas suffisamment détaillé. »[185]

182 MAROYE, Laurence. *op.cit.* p.80
183 W3C. 2012. *op.cit.*
184 cf. annexe 4
185 cf. annexe 3

En effet, dans le cas de GeoEthno, thésaurus géographique adapté au domaine de l'ethnologie, il n'est pas rare que le terme préférentiel associé à un concept change en fonction de la période historique. A titre d'exemple, le terme préférentiel « Paris », correspondant à l'actuelle capitale de la France, possède sur GeoEthno également trois formes renvoyant à des appellations passées : « Parisiis » et la forme latine « *Lutetia* » qui sont utilisés en tant que termes alternatifs, et « Lutèce » qui, sur GeoEthno, est associée à une section propre : « Nom(s) historique(s) ». Le choix de conserver les termes de géographie historique en tant que descripteurs fait effectivement sens dans ce contexte particulier comme l'explique Isabelle Donze : « pour la thématique qu'on a nous ici qui est l'ethnologie, c'est très important parce qu'il y a beaucoup de pays et de territoires qui ont changé de nom », il s'agit donc d'une façon de « situer un petit peu les notices et les documents dans le temps ». De fait nous pourrions imaginer un usager cherchant des ressources relatives à, par exemple, « la population de Paris au II^{ème} siècle » préférer effectuer sa recherche avec le mot-clef « Lutèce » afin de limiter le bruit documentaire qu'il aurait en utilisant « Paris ». Toutefois cette mention de « nom historique » est absente d'Isidore, probablement car absente du format Skos, ceux-ci y figurent donc parmi les termes alternatifs.[186]

Plus spécifiquement, nous avons vu qu'Isidore expose dans les détails de la section « Enrichissements ? » les termes préférentiels et alternatifs dans différentes langues, les notes, les définitions et les concepts liés pour chacun des concepts de l'onglet. Il nous paraît donc judicieux de faire l'hypothèse que chacun de ces champs est exploité par le moteur afin d'établir la sémantique des concepts et donc ne pas se limiter à de la fouille de texte par reconnaissance de chaînes de caractères. Or, dans notre corpus pourtant restreint, nous avons identifié des erreurs de compréhension sémantique dans la section « Enrichissements ? » sur presque un tiers des ressources. Nous pouvons constater que tous les référentiels sont concernés par ses erreurs et nous verrons parmi les exemples que nous avons sélectionnés, que l'influence du niveau de détail d'un concept sur sa bonne compréhension par Isidore semble assez faible.

186 cf. annexe 13

La première catégorie d'erreur d'interprétation sémantique que nous avons identifiée est due à la non-prise en compte de l'homonymie et à la pauvreté des concepts ambigus. En effet, si nous prenons l'exemple du concept « Côtes (anatomie) » issu de Rameau et utilisé pour enrichir la ressource « ODIT - Section A2 des plans cadastraux napoléoniens de la commune de Gennes »[187], nous pouvons remarquer que celui-ci est hors-sujet alors que « côte » dans le sens de « littoral » aurait fait sens. Or, ce concept est faiblement détaillé puisqu'il ne comporte que trois termes alternatifs, dont la forme polysémique « côte », et la mention limitant sa portée à l'anatomie ne figure pas dans une zone dédiée qui permettrait à la machine de le prendre en compte, mais dans l'intitulé du terme préférentiel. Le terme « côte » étant polysémique il n'est donc pas surprenant qu'en l'absence d'informations structurées pour restreindre la portée du concept, Isidore n'ait pas été en mesure d'accéder au sens du terme là où celui-ci apparaît comme une évidence pour l'utilisateur humain.

Un constat similaire peut être fait sur la notice : « Les ressources en eau et le changement climatique en Provence-Alpes-Côte d'Azur »[188] avec le concept « Vis » issu de Rameau et de GeoEthno. Dans le premier cas le seul détail du concept est le terme alternatif « Visserie » qui indique que le concept renvoie à la pièce métallique utilisée pour faire des constructions, dans le second cas nous avons deux formes alternatives « Issa » et « Lissa », qui, couplées au fait que GeoEthno est un thésaurus géographique, permettent de savoir que ce concept renvoie à une île croate. Or, là où ces inférences permettant de déduire le sens des concepts sont faites facilement par des cerveaux humains, il n'en va pas de même pour une machine qui a besoin que les informations soient davantage structurées et formalisées.

Si nous revenons sur l'exemple du concept « côte » de la notice précédente, nous pouvons remarquer que le problème de la polysémie n'entrait pas en ligne de compte pour la version de ce concept existant sur Gemet et Pactols, car ces deux thésaurus sont plus spécialisés que Rameau dont le périmètre est quasiment encyclopédique ; or c'est ce caractère spécialisé qui a restreint la portée du concept au moment même de sa création, « côte » au sens d'« os » ayant moins sa place dans des Soc dédiés à l'environnement ou à l'archéologie que « côte »

187 Isidore. ODIT - Section A2 des plans cadastraux napoléoniens de la commune de Gennes. *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 23 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.evrwyj

188 Isidore. Les ressources en eau et le changement climatique en Provence-Alpes-Côte d'Azur. *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 24 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.hb75cp

au sens de « littoral », il s'agit d'ailleurs du terme préférentiel utilisé sur Pactols, « côte » étant ici un terme alternatif. Par conséquent il semblerait qu'une façon de faire face aux difficultés de compréhension sémantique serait donc de recourir à des vocabulaires très spécialisés dans lesquels l'homonymie n'aurait pas sa place. Toutefois procéder ainsi présenterait de nombreuses contraintes, par exemple cela obligerait un projet comme Isidore à mobiliser de très nombreux référentiels afin d'assurer l'enrichissement de toutes les ressources présentes sur la plate-forme, la thématique des sciences humaines et sociales étant large. Or, il est possible qu'il n'existe pas de système d'organisation des connaissances spécifiques à chaque thème, par conséquent soit certains ne seraient pas enrichis, ce qui irait à l'encontre des ambitions du projet, soit il faudrait recourir à au moins un référentiel plus généraliste mais dans ce cas les difficultés de compréhension sémantique serait à nouveau présentes. Dans le même ordre d'idée, afin d'éviter l'homonymie entre vocabulaires contrôlés, par exemple entre un sur la médecine où « côte » serait présent au sens anatomique et un sur la mer où « côte » renverrait au littoral, il faudrait que les corpus sur lesquels ils sont utilisés soit très restreints sur le plan thématique. Or, si nous reprenons l'exemple que nous venons de citer cela signifie qu'une ressource sur le thème des maladies hydriques ne pourrait pas être indexée par le croisement d'un langage documentaire sur la médecine et d'un autre sur la mer mais nécessiterait un langage propre qui serait à son tour incompatible avec d'autres langages et inutilisable pour un certain nombre de ressources, obligeant les ressources des portails documentaires à être tout aussi spécialisés que les systèmes d'organisation des connaissances qu'ils exploitent. Or, procéder ainsi ne serait pas envisageable pour des ressources dont le niveau de compréhension serait grand public car, par définition, le jargon très spécialisé est inconnu du grand public et plutôt que d'aider à la recherche d'information, le système d'organisation des connaissances risquerait alors plutôt de l'entraver. En outre, plus un système d'organisation des connaissances est spécialisé plus ses chances d'être réutilisable dans un contexte autre que celui de sa création est faible, par conséquent ce type de système d'organisation des connaissances n'aurait pas forcément de raison de s'insérer dans le web des données et d'être interopérable. Par conséquent le recours à des vocabulaires très spécialisés pour lutter contre l'ambiguïté sémantique à laquelle font face les technologies de traitement automatique de la langue ne nous semble pertinent que pour des projets en interne sur des corpus qui sont eux aussi très spécialisés.

La seconde catégorie d'erreur de compréhension sémantique que nous avons pu identifier sur Isidore laisse à penser que l'algorithme du moteur d'enrichissement ne prend pas en considération les définitions associées à un concept. En effet, si nous prenons l'exemple de la notice « L'assassinat de Chantelle. Le notaire Lépine tué par Mme Achet (1890) »[189], la section « Enrichissements ? » contient deux fois le concept « Francs », un issu de Rameau, l'autre de Pactols. Ces deux concepts renvoient au peuple germanique du même nom et non à l'ancienne monnaie de la France, ils sont donc tous les deux hors-sujet et leur présence est donc due à un manque de compréhension sémantique de la part du moteur d'Isidore. Néanmoins il est intéressant de remarquer que ces deux concepts n'ont pas été détaillé de la même façon. En effet, là où celui issu de Rameau comprend exclusivement des termes alternatifs qui, comme nous l'avons constaté dans les exemples précédents, ne permettent pas à Isidore d'effectuer des liens d'inférence, le concept issu de Pactols possède une définition en français : « Peuple germanique, peut-être originaire de la Baltique (Lar.) » dont l'exploitation aurait dû permettre d'exclure le concept des enrichissements de cette notice. Précisons que la langue dans laquelle est écrite la définition ne semble avoir aucune influence sur sa prise en compte par le moteur d'Isidore. Par exemple, le concept « état » issu du thésaurus Gemet et présent parmi les enrichissements de la notice « Détail d'un sismomètre lors de sa réparation sur la plateforme instrumentale sismologique de l'EOST »[190] a une définition en anglais qui le limite au sens de « nation » ce qui le rend inapproprié dans le cas présent.

Dans le même ordre d'idée, les notes ne semblent pas davantage prises en considération par l'algorithme d'Isidore. Par exemple, la ressource sus-mentionnée est également enrichie avec le concept « Croissance » issu de Rameau ; or, là où le document concerne la croissance au sens économique du terme, une note du concept de Rameau spécifie que : « Cette subdivision s'applique aux êtres vivants, parties du corps et sujets noms communs appropriés », ce qui aurait dû limiter la portée du concept et donc ne pas le faire figurer parmi les enrichissements de cette notice. Notons toutefois que de nombreuses notes de concepts issues de Rameau servent à renvoyer à d'autres concepts ou subdivisions de ce

189 Isidore. L'assassinat de Chantelle. Le notaire Lépine tué par Mme Achet (1890). *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 24 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.6dyyfk

190 Isidore. Détail d'un sismomètre lors de sa réparation sur la plateforme instrumentale sismologique de l'EOST. *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 24 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.cxqdqz

vocabulaire, or dans la mesure où le vocabulaire n'est pas librement navigable par l'utilisateur nous pouvons nous interroger sur la pertinence d'une telle information puisque le chercheur d'information n'a aucun moyen de rebondir sur, par exemple, une subdivision indiquée dans une note. En outre, il s'agit également souvent de notes à l'attention des indexeurs, or Rameau étant utilisé par Isidore pour faire de l'indexation automatique mais sans prise en compte de la syntaxe puisque nous avons vu que Skos ne permet pas de faire des chaînes d'indexation pré-coordonnées, nous pouvons nous interroger sur l'utilité d'une note telle que celle du concept « Françaises » précisant :

> « Sous cette vedette, obligatoirement suivie d'une subdivision géographique ou de la subdivision : À l'étranger, directement ou après une subdivision de sujet, on trouve les documents sur les femmes françaises hors de France. Les documents sur les femmes françaises en France se trouvent sous des vedettes-matière telles que : Femmes -- France ; Femmes -- [Subdivision de sujet] – France ; etc. ».[191]

Le même problème s'observe sur des concepts issus de Pactols comme « Histoire » mais cela reste bien plus marginale que sur Rameau où le recours aux notes pour aider l'indexeur est fréquent.

Enfin, la faiblesse de détails dans la description d'un concept peut parfois empêcher de savoir si la sémantique prévue lors de sa création a bien été respectée. Si nous prenons l'exemple de concept « Echelles » issu de Rameau, le fait qu'il n'apparaisse sur Isidore qu'avec cette forme préférentielle laisse à penser qu'il englobe aussi bien l'outil que l'échelle géographique (sens dans lequel il est utilisé pour enrichir la notice « Inventaire du fonds de cartes sur le Brésil de la Cartothèque-Photothèque du Centre de documentation REGARDS-CNRS/ADES »[192]), or, après vérification dans le catalogue des autorités de la BNF, il semblerait que ce concept renvoie bien à l'outil tandis que l'échelle géographique est incarnée par le concept « Cartes – Échelles ».

Rappelons tout de même une nouvelle fois que la taille du corpus sur lequel nous avons basé nos observations est infime par rapport au nombre total de ressources indexées sur Isidore, notre étude ne peut donc en aucun cas prétendre conclure à des généralités sur le

191 cf. annexe 14

192 Isidore. Inventaire du fonds de cartes sur le Brésil de la Cartothèque-Photothèque du Centre de documentation REGARDS-CNRS/ADES. *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 24 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.kq7ph2

fonctionnement de la section « Enrichissements ? » de la plate-forme de recherche. Bien que cela n'enlève rien à la légitimité des analyses que nous venons de présenter, précisons tout de même que la grande majorité des concepts que nous avons pu analyser dans la section « Enrichissement ? » est appropriée aux différentes ressources associées. Par conséquent nous pouvons nous interroger sur le niveau de sémantique exploité par Isidore. Répondre à cette question nécessiterait de savoir quelles sont les techniques de fouilles de texte utilisées par Isidore ainsi que l'exploitation que le moteur fait des informations des concepts qu'il exploite.

III.I La question de la pertinence

En dépit des aspects relatifs à la compréhension sémantique, nous pouvons nous interroger sur la pertinence des concepts présents dans la section « Enrichissements ? ». En effet, le fait que leur sens soit adapté à la ressource qu'ils détaillent n'est pas forcément gage de pertinence et donc de plus-value pour le chercheur d'information. Le travail d'indexation consiste à mettre en exergue les concepts les plus pertinents et représentatifs du contenu d'un document afin que l'utilisateur sache très rapidement si la ressource est susceptible de répondre à son besoin d'information. Or, nous pouvons questionner l'intérêt pour le chercheur d'information de faire figurer le nom du mois auquel s'est déroulé l'événement traité dans une ressource parmi les enrichissements, comme c'est pas exemple le cas avec le concept issu du Rameau « Décembre » pour la ressource « "Ne pillons plus ! Instruisons !" : article-pétition de la féministe Hubertine Auclert pour l'ouverture d'une école arabe pour filles à Alger » ou encore le concept « Octobre » issu du même vocabulaire mais pour la ressource « Poste d'assistant-e diplômé-e à 50% – Chaire d'histoire comparée des religions et de dialogue interreligieux ».[193]

Beaucoup de remarques concernant la pertinence touchent à des concepts issus de Rameau. En effet, s'agissant d'un langage documentaire pré-coordonné, nombre des concepts issu de ce vocabulaire doivent normalement être agrégés à d'autres pour proposer une indexation pertinente, or, comme nous l'avons expliqué, cela est impossible à faire à partir de la version en Skos du langage. En outre, Rameau est un système d'organisation des connaissances plus généraliste que Gemet, GeoEthno ou Pactols, par conséquent la probabilité d'identifier des

193 Isidore. Poste d'assistant-e diplômé-e à 50% – Chaire d'histoire comparée des religions et de dialogue interreligieux. *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 24 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.75fbbf

concepts génériques, peu pertinents est plus élevée. Toutefois la question de la pertinence ne se limite pas au degré de spécificité du concept utilisé. En effet, certaines ressources ayant une portée large, il n'est pas étonnant que des concepts d'ordre généraux soient utilisés pour les indexer, *a contrario* un excès de précision n'est pas non plus souhaitable car cela pourrait mettre en exergue une partie infime du contenu du document et, éventuellement, biaiser son niveau de spécialisation.

Le nombre de concepts affichés dans la section « Enrichissements ? » pose également question dans la mesure où il ne semble pas limité. En effet, pour notre corpus de 24 ressources, la moyenne des concepts présents dans la section francophone « Enrichissements ? » est de 29 avec une valeur minimale de trois[194] et une maximale de 112.[195] Or, nous pouvons nous interroger sur l'utilité d'afficher une centaine ou même une trentaine de concepts qui, de part leur nombre, risquent d'alourdir la lecture de la notice par l'usager et de provoquer une surcharge cognitive ; par conséquent au lieu de faciliter la recherche d'information celle-ci serait rendue plus difficile. Une réflexion similaire peut être faite concernant le détail des concepts de la section « Enrichissements ? ». En effet, lorsque nous avons présenté le thésaurus Gemet, nous avons fait l'hypothèse que les 37 langues le composant n'avaient pas été utilisées par Isidore afin de ne pas alourdir l'interface, or nous avons pu remarquer que le concept « Afrique » issu de GeoEthno affiche sur Isidore plus de 100 termes alternatifs[196], posant des questions en terme d'ergonomie.

Notons d'ailleurs que si, à de nombreuses reprises, nous avons identifier des similitudes entre les mots-clefs des notices originales et les enrichissements, ces derniers ne reprenaient pas forcément l'intégralité des mots-clefs. Cette observation peut être expliquée de différentes façons : la première est que le travail d'indexation n'a pas été réalisé avec les même référentiels que celui d'enrichissement, ainsi il est possible qu'un concept présent dans les mots-clefs ne trouve pas d'équivalent dans les langages documentaires de la section

194 Isidore. Le Matérialisme rationnel. *Isidore – Accès aux données et services numériques de SHS*, [en ligne]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.3acmf1

195 Isidore. Les ressources en eau et le changement climatique en Provence-Alpes-Côte d'Azur. *Isidore – Accès aux données et services numériques de SHS*, [en ligne]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.hb75cp

196 Isidore. Village, région d'Allada (Bénin). *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 24 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.hn8kwk

propre à Isidore, la seconde hypothèse tient au fait que les enrichissements sont réalisés par des procédés d'indexation automatique, l'expertise du professionnel de l'information-documentation quant au choix des termes à utiliser n'entre donc pas en considération.

Dans la mesure où Isidore affiche également les mots-clefs, nous pouvons faire l'hypothèse que la section « Enrichissements ? » n'a pas forcément pour vocation première d'être utilisée par un chercheur d'information humain. En effet, comme nous avons pu l'entrevoir dans notre présentation de la plate-forme, le projet est présenté comme une solution tout-en-un pouvant être consultée directement depuis son interface web public, mais également intégrée à d'autres sites web. Cette intégration peut de se faire de différentes façons : par le biais d'API (*Application programming interface* – interface de programmation)[197] permettant d'accéder aux différents éléments d'une ressource, mais également par le biais de modules un peu plus spécifiques, bien que reposant toujours sur une API, comme par exemple « Isidore Suggestion » développé pour le CMS (*Content Management System* – système de gestion de contenu) Wordpress et qui permet d'ajouter des suggestions de lectures issues d'Isidore à un article publié sur Wordpress à partir de la correspondance entre les mots-clefs utilisés dans l'article et ceux, ainsi que que les enrichissements sémantiques, présents dans les ressources d'Isidore ; ou encore Imoco (*Isidore Motor Contructor*) permettant d'intégrer le moteur de recherche d'Isidore à n'importe quel site web[198], à titre d'exemple Imoco a été intégré à l'interface de consultation du thésaurus GeoEthno. Dernièrement, le projet « Isidore à la demande », actuellement accessible en version bêta, mutualise « un ensemble de services informatiques et d'APIs permettant à des chercheur.e.s ou à des groupes de recherche d'utiliser les outils de traitement sémantique de la plateforme ISIDORE ».[199]

Au regard de ces différentes applications d'Isidore, nous pouvons donc affirmer que l'interface de consultation sur laquelle nous avons fait porter notre étude n'est qu'une des

197 Isidore. API. ISIDORE – Accès aux données et services numériques de SHS, [en ligne]. [Consulté le 27 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/api

198 POUYLLAU, Stéphane. ISIDORE Suggestion, des recommandations de lectures pour les blogs de science. Version auteur. *I2D – Information, données & documents*, 2016/2, Vol. 53. Également disponible en ligne à l'adresse : https://archivesic.ccsd.cnrs.fr/sic_01348561

199 CO.SHS. Entrevue avec Huma-Num sur ISIDORE à la demande. *Produire – Découvrir – Explorer – CO.SHS*, [en ligne]. 10 mai 2018. [Consulté le 03 août 2018]. Disponible à l'adresse : https://co-shs.ca/fr/nouvelles/entrevue-avec-huma-num-sur-isidore-a-la-demande/

briques d'un projet plus vaste de diffusion et d'interconnexion d'information scientifique et technique. Cependant cela ne remet pas en question les observations que nous avons pu faire et peut également poser question quant à l'utilité de développer des services permettant d'exploiter les enrichissements sémantiques, qui sont, certes, la particularité d'Isidore, mais qui comporte à ce jour encore un certain nombre d'imprécisions pouvant rendre leur utilisation parfois contre-productive pour le destinataire final humain.

Conclusion

L'objectif de ce mémoire était de déterminer la plus-value de la compatibilité des systèmes d'organisation des connaissances avec web des données afin d'améliorer l'efficacité de la recherche d'information. L'état de l'art présenté en première partie a soulevé l'ambiguïté des concepts de « web des données » et de « système d'organisation des connaissances », deux notions vastes dont les contours ne font pas l'unanimité, aussi bien dans la communauté scientifique que dans la communauté des professionnels de l'information-documentation. Toutefois, afin que le présent document ne se limite pas à une présentation des différentes définitions de ces notions et des paradigmes qui en découle, bien que cela aurait pu être très intéressant notamment pour mettre en exergue les différences de perception entre les scientifiques et les professionnels de terrain, ou encore entre les différents contextes professionnels, nous avons fait un certain nombre de choix de définition, certes discutables, mais essentiels afin de délimiter le périmètre de notre étude. Nous avons, en particulier, fait le choix de nous concentrer davantage sur une famille de systèmes d'organisation des connaissances précise : les thésaurus. De part leur contexte d'apparition concomitant à celui de l'informatique documentaire et les évolutions qu'ils ont connu ces dernières années avec la parution en 2011 et 2013 de la norme internationale Iso 25964 « Thésaurus et interopérabilité avec d'autres vocabulaires », les thésaurus nous ont semblé cristalliser les enjeux relatifs au web des données pour les Soc, et notamment la question de l'interopérabilité dont nous avions fait l'hypothèse qu'il s'agit d'un enjeu majeur pour les systèmes d'organisation des connaissances dans le web des données.

L'étude de cas autour de la section « Enrichissements ? » présente sur les notices documentaires de la plate-forme Isidore, nous a permis d'apporter une dimension concrète à notre travail et d'ajouter à l'exemple des thésaurus celui des listes de vedettes-matière. Cela nous a permis de constater que le format Skos n'est pas optimal pour tous les systèmes d'organisation des connaissances, mais également que, notamment face aux difficultés de compréhension sémantique, tous les Soc n'ont pas nécessairement besoin d'être interopérables et compatibles avec le web des données.

Comme nous avons eu l'occasion de l'exprimer à plusieurs reprises au cours de notre étude, la principale limite de notre travail vient du fait que nous n'avons pas pu nous entretenir avec un responsable technique du projet Isidore afin de pouvoir connaître le fonctionnement concret de la section « Enrichissements ? », les pistes d'amélioration de la compréhension sémantique, ou encore les ambitions du projet Isidore dans son ensemble à court et long terme. Par conséquent, si cela n'entrave pas la qualité des nos observations, nos analyses sont tout de même à considérer avec prudence et demeurent au stade d'hypothèses qu'il faudrait vérifier.

Si les entretiens que nous avons effectué avec des professionnels travaillant sur des systèmes d'organisation des connaissances utilisés par Isidore n'ont pu qu'être partiellement exploités pour étayer notre étude de cas, ils ont tout de même permis d'alimenter notre réflexion ouvrant des perspectives de prolongement de ce mémoire quant à la façon dont un outil de gestion peut conditionner la création et le développement d'un langage documentaire, la perte ou a contrario le gain d'information permis par différent format d'expression de Soc et la question de l'interopérabilité qui en découle, ou encore celle de l'exploitation de la complémentarité de différents systèmes d'organisation des connaissances dans un même système de recherche d'information. Enfin, une autre piste de prolongement de notre travail concerne les ontologies, sujet trop vaste pour que nous l'abordions en détail dans ce mémoire, mais, dans la mesure où l'ontologie est un type de système d'organisation des connaissances et que certains Soc, notamment les thésaurus, peuvent être utilisés comme fondement pour l'élaboration d'une base de connaissances, nous pourrions nous demander si la vocation d'un langage documentaire compatible avec le web des données ne serait pas d'alimenter une ontologie, ou encore si une ontologie pourrait être utilisée en remplacement des référentiels utilisés dans un projet comme Isidore.

Bibliographie

Bibliographie et webographie de l'état de l'art

ACCART, Jean-Philippe et RIVIER, Alexis. *Mémento de l'information numérique*. Paris : Éditions du Cercle de la Librairie, 2012. 184 p. Collection Bibliothèques. ISBN 978-2-7654-1332-5

AUSSENAC-GILLES, Nathalie. Chapitre 8. Le web sémantique, quel renouvellement pour la recherche d'information ? In : BOUGHANEM, Mohand et SAVOY, Jacques. *Recherche d'information : état des lieux et perspectives*. Paris : Lavoisier, 2008. pp.231-266. Collection Recherche d'information et web. ISBN 978-2-7462-2005-8

BEIGBEDER, Michel. Les temps du document et la recherche d'information. *Document numérique*, 2004/4. Vol.8. pp.55-64. Également disponible en ligne à l'adresse :

https://www.cairn.info/article.php?ID_ARTICLE=DN_084_0055

BERMES, Emmanuelle et POUPEAU, Gautier. Les technologies du web appliquées aux données structurées. In : CALDERAN, Lisette, *et. al*. *Le document numérique à l'heure du web de données. Séminaire Inria Carnac, 1er-5 octobre 2012*. Paris : ADBS Éditions, 2012. pp.41-84. Sciences et techniques de l'information. ISBN 978-2-84365-142-7.

BOUBEE, Nicole et TRICOT, André. *Qu'est-ce que rechercher de l'information ?* Villeurbane : Presses de l'ENSSIB, 2010. 286 p. Également disponible en ligne à l'adresse :

http://books.openedition.org/pressesenssib/799

BOUGHANEM, Mohand. Chapitre 1. Introduction à la recherche d'information. In : BOUGHANEM, Mohand et SAVOY, Jacques. *Recherche d'information : état des lieux et perspectives.* Paris : Lavoisier, 2008. pp.19-44. Collection Recherche d'information et web. ISBN 978-2-7462-2005-8

Centre national RAMEAU. *Guide d'indexation RAMEAU* [en ligne]. 7ème édition. Bibliothèque nationale de France, 2017, 270 p. [Consulté le 09 juin 2018]. Disponible à l'adresse : http://rameau.bnf.fr/docs_reference/pdf/Guide_RAMEAU_2017.pdf#page=7

CHIARAMELLA, Yves et MULHEM, Philippe. La recherche d'information. De la documentation automatique à la recherche d'information en contexte, *Document numérique*. 2007/1, Vol.10, pp.11-38. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-document-numerique-2007-1-page-11.htm

CROZAT, Stéphane. Écrire avec une machine à calculer, écrire pour une machine à calculer, *I2D – Information, données & documents*. 2016/2, Vol.53, pp.62-64. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-i2d-information-donnees-et-documents-2016-2-page-62.htm

CYROT, Catherine et PREUSS, Christian. Réingénierie de thésaurus : une étude de cas, *Documentaliste-Sciences de l'Information*. 2009/3, Vol.46, pp.4-13. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2009-3-page-4.htm

DALBIN, Sylvie, *et al*. *Livre banc ISO 25964-1 – Thésaurus pour la recherche documentaire* [en ligne]. AFNOR, 2013. 39 p. [Consulté le 12 juin 2018]. Disponible à l'adresse : http://dossierdoc.typepad.com/files/iso25964-1-livre-blanc-janvier-2013-vfinale.doc

DELESTRE, Nicolas et MALANDAIN, Nicolas. *Du web des documents au web sémantique.* Bois-Guillaume : Éditions KLOG, 2017. 199 p. ISBN 979-10-92272-18-5.

DEROCHE, Frédéric. Recherche d'information (recherche documentaire). *Enssib* [en ligne]. Mise à jour le 19 août 2015. [Consulté le 22 avril 2018]. Disponible à l'adresse : http://www.enssib.fr/le-dictionnaire/recherche-dinformation-recherche-documentaire

DESFRICHES DORIA, Orélie et ZACKLAD, Manuel. Améliorer la recherche d'information à l'aide de thésaurus « ad hoc ». Expérimentations et réflexions méthodologiques, *Document numérique*. 2010/2, Vol.13, pp.13-40. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-document-numerique-2010-2-page-13.htm

DEXTRE CLARKE, G Stella. *Thesauri, interoperability and the role of ISO 25964*, [en ligne]. 38 diapositives. [Consulté le 02 juin 2018]. Disponible à l'adresse : http://www.docslides.com/thesauri-interoperability-and-the-role-of-iso-25964

DINET, Jérôme. *La recherche d'information dans les environnements numériques*. Londres : ISTE Editions, 2014. 134 p. Collection systèmes d'information, web et informatique ubiquitaire. ISBN 978-1-78405-018-4

FEYLER, François. Vocabulaires contrôlés. *Savoirs CDI* [en ligne]. Avril 2009. [Consulté le 05 juin 2018]. Disponible à l'adresse : https://www.reseau-canope.fr/savoirscdi/societe-de-linformation/tic-et-documentation/veille-technologique/formats-normes-et-standards/vocabulaires-controles.html

GANDON, Fabien, FARON-ZUCKER, Catherine et CORBY, Olivier. *Le web sémantique : comment lier les données et les schémas sur le web ?* Paris : Dunod, 2012. 206 p. ISBN 978-2-10-058140-5.

GNOLI, Claudio. Chapitre 2 : Des métadonnées représentant quoi ?. In : EL HADI, Widad Mustafa. *L'organisation des connaissances : dynamisme et stabilité*. Paris : Lavoisier, 2012. pp.51-66. Traité des sciences et techniques de l'information. ISBN 978-2-7462-3227-3

GREFENSTETTE, Gregory, *et al*. Chapitre 3 : L'utilisation de la sémantique dans les applications basées sur la recherche d'information. In : GRIVEL, Luc. *La recherche d'information en contexte : outils et usages applicatifs*. Paris : Lavoisier, Hermès Sciences, 2011. pp. 97-118. Traité des sciences et techniques de l'information. ISBN 978-2-7462-2581-7.

HENDLER, James. The dark side of the semantic Web, *IEEE Intelligent Systems*. Janvier/Février 2007, Vol.22, n°1, pp.2-4. Également disponible en ligne à l'adresse : https://www.computer.org/csdl/mags/ex/2007/01/x1002.html

HODGE, Gail . *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Council on Library and Information Resources, 2000. ISBN 1-887334-76-9. 1. Knowledge Organization Systems: An Overview. Également disponible en ligne à l'adresse : https://www.clir.org/pubs/reports/pub91/1knowledge/

HUDON, Michèle. ISO 25964 : pour le développement, la gestion et l'interopérabilité des langages documentaires, *Documentation et bibliothèques*. 2012(a)/3, Vol.58, pp.130–140. Également disponible en ligne à l'adresse :

https://www.erudit.org/fr/revues/documentation/2012-v58-n3-documentation01721/1028903ar.pdf

HUDON, Michèle. Chapitre 13 : ISO 25964 : vers une nouvelle norme pour l'organisation et l'accès à l'information et aux connaissances. In : EL HADI, Widad Mustafa. *L'organisation des connaissances : dynamisme et stabilité*. Paris : Lavoisier, 2012(b). pp.207-219. Traité des sciences et techniques de l'information. ISBN 978-2-7462-3227-3

IKOJA-ODONGO, Robert et MOSTERT, Janneke. Information seeking behaviour: A conceptual framework, *South African Journal of Libraries and Information Science*. 2006/3, Vol.72, pp.145-158. Également disponible en ligne à l'adresse : http://sajlis.journals.ac.za/pub/article/view/1112

ISAAC, Antoine. Entre thésaurus et ontologies : une affaire d'interopérabilité et d'alignement, *Documentaliste-Sciences de l'Information*. 2011/4, Vol.48, pp.48-49. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2011-4-page-42.htm

KELLER, Michael. Linked Data: a way out of the information chaos and toward the semantic web, *EDUCAUSE* [en ligne]. 21 juillet 2011. [Consulté le 14 novembre 2017]. Disponible à l'adresse :https://er.educause.edu/articles/2011/7/linked-data-a-way-out-of-the-information-chaos-and-toward-the-semantic-web

KEMBELLEC, Gérald. Que voit réellement Google de la sémantique des pages web ? , *I2D – Information, données & documents*. 2016/2, Vol.53, p.65. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-i2d-information-donnees-et-documents-2016-2-page-65.htm

MAHE, Sylvain, *et. al.* Gestion des connaissances et systèmes d'organisation de connaissances : premier modèle et retours d'expérience industriels, *Document numérique*. 2010/2, Vol.13, pp.57-73. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-document-numerique-2010-2-page-57.htm

MANIEZ, Jacques. *Actualité des langages documentaires : fondements théoriques de la recherche d'information*. Paris : ADBS Edition, 2002. 395 p. Collection Sciences de l'Information Série Études et techniques. ISBN 2-84365-060-7

MAROYE, Laurence. ISO 25964 : de la distinction formelle concept/terme préconisée par la norme pour la création et la gestion des thésaurus, *I2D – Information, données & documents*. 2015/1, Vol.52, pp.72-80. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-1-page-72.htm

MASTORA, Anna, *et. al*. SKOS Concepts and Natural Language Concepts: an Analysis of Latent Relationships in KOSs, *Journal of Information Science*. 2017/4, Vol.43, pp.1-19. Également disponible en ligne à l'adresse :

https://www.researchgate.net/publication/303322229_SKOS_Concepts_and_Natural_Language_Concepts_an_Analysis_of_Latent_Relationships_in_KOSs

MENON, Bruno. Journée d'étude ADBS. Optimiser l'accès à l'information, une opportunité pour les langages documentaires ?, *Documentaliste-Sciences de l'Information*. 2007(a)/6, Vol.44, pp.385-388. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-6-page-385.htm

MENON, Bruno. Les langages documentaires. Un panorama, quelques remarques critiques et un essai de bilan, *Documentaliste-Sciences de l'Information*. 2007(b)/1, Vol.44, pp.18-28. Également disponible en ligne à l'adresse : https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-18.htm

MERCEDES MARTÍNEZ-GONZÁLEZ, M. et ALVITE-DÍEZ, María-Luisa. On the evaluation of thesaurus tools compatible with the Semantic Web, *Journal of Information Science*. 2014/6, Vol.40. pp.1-13. Également disponible en ligne à l'adresse : https://www.researchgate.net/publication/271731043_On_the_evaluation_of_thesaurus_tools_compatible_with_the_Semantic_Web

MONNIN, Alexandre. Du cycle de vie des données au cycle de vie des objets. In : CALDERAN, Lisette, *et. al*. *Le document numérique à l'heure du web de données. Séminaire Inria Carnac, 1er-5 octobre 2012*. Paris : ADBS Éditions, 2012. pp.221-228. Sciences et techniques de l'information. ISBN 978-2-84365-142-7.

PEDAUQUE, T. Roger. *La redocumentarisation du monde*. Toulouse : Cépaduès, 2007. ISBN 978-2-85428-728-8. Chapitre 6 : Sur des aspects primordiaux du web sémantique, pp. 99-116.

POUCHOT, Stéphanie et EPRON, Benoît. Chapitre 8 : Classer numérique. In : SALAÜN, Jean-Michel et HABERT, Benoît (dir). *Architecture de l'information : méthodes, outils, enjeux*. Louvain-la-Neuve : de boeck, 2015. pp.161-182. Information & Stratégie. ISBN 978-2-8041-9140-5.

POUPEAU, Gautier. Petite histoire du Web sémantique, *Les petites cases* [en ligne]. 15 août 2011. [Consulté le 21 novembre 2017]. Disponible à l'adresse :

http://www.lespetitescases.net/petite-histoire-du-web-semantique

POUPEAU, Gautier. Histoire(s) de notices. In : CALDERAN, Lisette, *et. al. Le document numérique à l'heure du web de données. Séminaire Inria Carnac, 1er-5 octobre 2012*. Paris : ADBS Éditions, 2012. pp.25-40. Sciences et techniques de l'information. ISBN 978-2-84365-142-7.

RIGOUSTE, Loïs, *et al*. Chapitre 4 : Analyse sémantique et moteurs de recherche, apport des entités nommées. In : GRIVEL, Luc. L*a recherche d'information en contexte : outils et usages applicatif*. Paris : Lavoisier, Hermès Sciences, 2011. pp. 119-140. Traité des sciences et techniques de l'information. ISBN 978-2-7462-2581-7.

SALAÜN, Jean-Michel. Du document à la donnée et retour : la fourmilière ou les Lumières. In : CALDERAN, Lisette, *et. al. Le document numérique à l'heure du web de données. Séminaire Inria Carnac, 1er-5 octobre 2012*. Paris : ADBS Éditions, 2012. pp.9-24. Sciences et techniques de l'information. ISBN 978-2-84365-142-7.

SUOMINEN, Osma, *et. al*. *Publishing SKOS vocabularies with Skosmos*. Manuscript submitted for review. 2015, 24 p. Disponible en ligne à l'adresse : http://skosmos.org/publishing-skos-vocabularies-with-skosmos.pdf

W3C. *SKOS Simple Knowledge Organization System* [en ligne]. Mise à jour le 13 décembre 2012. [Consulté le 10 juin 2018]. Disponible à l'adresse : https://www.w3.org/2004/02/skos/

W3C. Linked Data. *World Wide Web Consortium (W3C)* [en ligne]. 2015. [Consulté le 09 mars 2018]. Disponible à l'adresse : https://www.w3.org/standards/semanticweb/data

ZACKLAD, Manuel. *Introduction aux ontologies sémiotiques dans le Web Socio Sémantique*. 2005. 12 p. Également disponible en ligne à l'adresse :

https://archivesic.ccsd.cnrs.fr/file/index/docid/62630/filename/sic_00001479.pdf

ZACKLAD, Manuel. Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (ROI). In : ARSENAULT, Clément et DALKIR, Kimiz. (dir.). *CAIS/ACSI 2007, Actes du 35e Congrès annuel de l'Association Canadienne des Sciences de l'Information. Partage de l'information dans un monde fragmenté : Franchir les frontières*. Montréal, 2007. pp.1-15. Également disponible en ligne à l'adresse : https://archivesic.ccsd.cnrs.fr/sic_00202440/document

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

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

ZARGAYOUNA, Haïfa, ROUSSEY, Catherine et CHEVALLET, Jean-Pierre. Recherche d'information sémantique : état des lieux, *Traitement automatique des langues*. 2015/3, Vol.56. pp.49-73. Également disponible en ligne à l'adresse : http://www.atala.org/sites/default/files/2.Zargayouna-56-3.pdf

Bibliographie et webographie de l'étude de cas

Bibliothèque Eric-de-Dampierre. Thésaurus géographique GeoEthno – Bref historique du thésaurus. *Le site de la bibliothèque Eric de Dampierre*, [en ligne]. Mise à jour le 30 mars 2012. [Consulté le 16 août 2018]. Disponible à l'adresse :

http://www.mae.u-paris10.fr/bibethno/spip.php?article20

BOUCHET, Thierry. Le vocabulaire Rameau en SKOS. *Journée nationale Rameau*, [en ligne]. 30 mai 2008. 13 diapositives. [Consulté le 14 août 2018]. Disponible à l'adresse : http://rameau.bnf.fr/informations/pdf/journee2008/rameau_skos.pdf

CANTIE, Philippe. Transition bibliographique : en avant marche ! *Bulletin des bibliothèques de France*, [en ligne]. 23 septembre 2015. [Consulté le 14 août 2018]. Disponible à l'adresse : http://bbf.enssib.fr/tour-d-horizon/transition-bibliographique-en-avant-marche_65461

CAPELLI, Laurent, *et. al. Les guides de bonnes pratiques : comment contribuer à Isidore avec ses données numériques ?*, [en ligne]. Paris : TGIR Huma-Num, 2014. 29 p. [Consulté le 04 août 2018]. Disponible à l'adresse : https://www.huma-num.fr/sites/default/files/guide-isidore.pdf

Centre national RAMEAU. Statistiques d'accroissement du référentiel RAMEAU. *BNF – RAMEAU*, [en ligne]. Décembre 2017. [Consulté le 14 août 2018]. Disponible à l'adresse : http://rameau.bnf.fr/informations/chiffres.htm

Centre national RAMEAU. Réformer RAMEAU. *BNF – RAMEAU*, [en ligne]. Mise à jour le 03 mai 2018. [Consulté le 14 août 2018]. Disponible à l'adresse : http://rameau.bnf.fr/chantier_syntaxe/intro.html

CO.SHS. Entrevue avec Huma-Num sur ISIDORE à la demande. *Produire – Découvrir – Explorer – CO.SHS*, [en ligne]. 10 mai 2018. [Consulté le 03 août 2018]. Disponible à l'adresse :

https://co-shs.ca/fr/nouvelles/entrevue-avec-huma-num-sur-isidore-a-la-demande/

Direction de l'information scientifique et technique – CNRS. *Livre blanc : une science ouverte dans une République numérique*, [en ligne]. Marseille : OpenEdition press, 2016. 195 p. Collection Laboratoire d'idées. Disponible à l'adresse :

https://books.openedition.org/oep/1548

Eionet. About Gemet. *European Environment Information and Observation network – Eionet*, [en ligne]. Mise à jour le 29 août 2017. [Consulté le 14 août 2018]. Disponible à l'adresse : http://www.eionet.europa.eu/gemet/fr/about/

FRANTIQ. Le thésaurus. *FRANTIQ : Fédération et ressources sur l'Antiquité*, [en ligne]. [Consulté le 16 août 2018]. Disponible à l'adresse : https://www.frantiq.fr/fr/thesaurus

Isidore. API. *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 27 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/api

MENARD, Florence. *Rapport du groupe de travail national sur la syntaxe de Rameau : préconisations et pistes d'évolution*, [en ligne]. Version 1.0. Paris : Centre national Rameau – Bibliothèque nationale de France, 05 mai 2017. 41 p. Disponible à l'adresse : http://rameau.bnf.fr/chantier_syntaxe/pdf/rapport_final_syntaxe_rameau.pdf

MENARD, Florence et ROUSSEAUX, Olivier. Quand Rameau se greffe au programme national. *Arabesques*. 2017, Vol. 87. p.12. Également disponible en ligne à l'adresse : http://www.abes.fr/Publications-Evenements/Arabesques/Arabesques-n-87

POUYLLAU, Stéphane. *Isidore : signaler, enrichir et valoriser les documents, données, informations numériques des sciences humaines*, [en ligne]. Paris : TGIR Huma-Num, 2013. 53 diapositives. [Consulté le 02 août 2018]. Disponible à l'adresse : https://www.rnbm.org/supports_anf/cirm2013/plateforme-ISIDORE.pdf

POUYLLAU, Stéphane. ISIDORE Suggestion, des recommandations de lectures pour les blogs de science. Version auteur. *I2D – Information, données & documents*, 2016/2, Vol. 53. Également disponible en ligne à l'adresse : https://archivesic.ccsd.cnrs.fr/sic_01348561

RAMADE, François. *Dictionnaire encyclopédique des sciences de la nature et de la biodiversité*. Paris : Dunod, 2008. 760 p. ISBN 978-2-10-053670-2

Telplus. RAMEAU subject headings as SKOS linked data. *STITCH Project*, [en ligne]. [Consulté le 28 août 2018]. Disponible à l'adresse : https://www.cs.vu.nl/STITCH/rameau/index-fr.html

TGIR Huma-Num. Isidore speaks English, sino también español et toujours en français. *Le blog d'huma-num et de ses consortiums*, [en ligne]. Mise à jour le 19 mai 2015. [Consulté le 03 août 2018]. Disponible à l'adresse : https://humanum.hypotheses.org/921

TRIPLET, Patrick. *Dictionnaire encyclopédique de la diversité biologique et de la conservation de la nature*. 4ème édition. 2018. 1096 p. ISBN 978-2-9552171-9-1. Également disponible en ligne à l'adresse :

https://www.dropbox.com/s/xdw5unlwancpao0/Dictionnaire%20conservation

Table des illustrations

Illustration 1: processus en « U » de recherche d'information.....15
Illustration 2: pile du web sémantique.....20
Illustration 3: enrichissements d'une notice sur Isidore.....46

Table des annexes

Annexe 1 : les trois zones d'une notice documentaire sur Isidore.....95

Annexe 2 : retranscription de l'entretien téléphonique avec Miled Rousset, responsable « Têtes de réseaux documentaires », Maison de l'Orient et de la Méditerranée Jean Pouilloux / responsable informatique OpenTheso.....96

Annexe 3 : retranscription de l'entretien téléphonique avec Isabelle Donze, bibliothécaire et responsable du thésaurus GeoEthno, bibliothèque Eric-de-Dampierre.....102

Annexe 4 : retranscription de l'entretien téléphonique avec Thierry Bouchet, en charge de la conversion de Rameau au format Skos, Centre national Rameau.....111

Annexe 5 : exemple d'alignement entre un concept Rameau et un de LCSH.....115

Annexe 6 : exemple d'alignement entre un concept Rameau et un de LCSH.....116

Annexe 7 : liens entre les concepts « Vente jumelée » de Rameau, « *Bundling (Marketing)* » de LCSH et « *Venta conjut*a » de la BNE sur Isidore et dans leur code XML-RDF.....117

Annexe 8 : détail des rebonds sur Isidore.....119

Annexe 9 : exemple d'auto-complétion dans la barre de recherche d'Isidore.....120

Annexe 10 : doublons du concept « marais » sur Gemet.....121

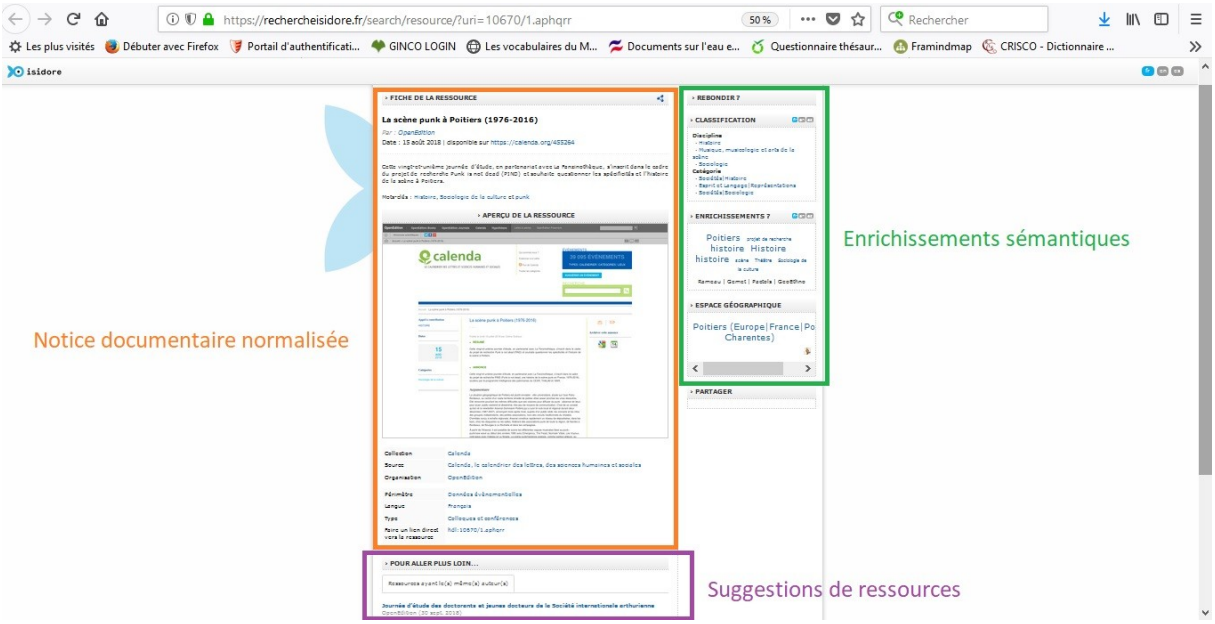
Annexe 11 : concept « première guerre mondiale » sur Isidore et Pactols, écarts de traductions entre les deux versions Pactols.....123

Annexe 12 : représentation du concept « Sud » issu de GeoNames sur Isidore.....124

Annexe 13 : concept GeoEthno « Paris » sur Isidore et sur GeoEthno.....125

Annexe 14 : détail du concept Rameau « Françaises » sur Isidore.....126

Annexe 1 : les trois zones d'une notice documentaire sur Isidore



Source : Isidore. OpenEdition, La scène punk à Poitiers (1976-2016). *ISIDORE – accès aux données et services numériques des SHS*, [en ligne]. [Consulté le 15 août 2018]. Disponible à l'adresse : https://rechercheisidore.fr/search/resource/?uri=10670/1.aphqrr

Annexe 2 : retranscription de l'entretien téléphonique avec Miled Rousset, responsable « Têtes de réseaux documentaires », Maison de l'Orient et de la Méditerranée Jean Pouilloux / responsable informatique OpenTheso.

Date : 16 juillet 2018

Durée : 22 minutes 21 secondes

C.B : Tout d'abord, est-ce que vous pourriez me présenter vos activités ?

M.R : Je suis informaticien de formation, je me suis spécialisé petit à petit dans l'informatique documentaire, et donc actuellement je gère le réseau Frantiq et je suis responsable d'une plate-forme technologique qui est spécialisée justement dans l'informatique documentaire. Essentiellement je développe des applications et je travaille avec les documentalistes sur tout ce qui est normes de catalogage, d'indexation, d'où mon intérêt, mon implication dans tout ce qui est indexation sémantique, thésaurus.

De ce que j'ai compris Pactols et Opentheso ont été créé en même temps, est-ce que vous pourriez me parler un petit peu de l'origine du projet ?

Non en fait. Je vais vous expliquer un peu comment ça s'est fait en fait. Pactols est un thésaurus qui a été créé au fil des années, ça fait déjà au moins 30 ans, avec la création du réseau Frantiq. A l'origine c'était que des petits catalogues, enfin pas petits, mais je veux dire des catalogues gérés par chaque petite bibliothèque de recherche. Chacun gérait sa bibliothèque comme il pouvait et petit à petit il y a des mots-clefs qui ont commencé à apparaître, à devenir des mots spécialisés dans l'archéologie, et donc après quelqu'un a, à un moment donné, commencé à réfléchir à la conception, création d'un thésaurus spécialisé dans l'archéologie. Donc ça ça remonte vraiment à très loin, après moi quand je suis arrivé à peu près en 2000, je suis arrivé dans le réseau Frantiq et autre, donc là j'ai vu qu'il y avait des besoins du réseau. Il y avait un gestionnaire de thésaurus qui était payant, qui était toujours intégré à un système intégré de gestion de bibliothèque, et en 2005 à peu près, le logiciel qui gérait les bibliothèques du réseau Frantiq était obsolète et donc le gestionnaire de thésaurus n'était plus valable. Donc là on avait le choix : ou on achetait un gestionnaire de thésaurus, il n'y en avait pas ou c'était privé et c'était très cher, ou de développer. Je l'ai fait moi-même,

j'ai commencé en 2005, c'était la première version en java pur, après en 2012 à peu près j'ai pu passer sur le logiciel qu'on a appelé « Opentheso », en full web. Et voilà, c'est comme ça que c'est né et ça continue à vivre.

Concernant les évolutions d'Opentheso, comme ça se passe ? Est-ce que c'est Opentheso qui évolue et donc Pactols qui évolue, ou c'est plutôt Pactols qui a envie d'évoluer et donc Opentheso qui s'adapte ?

En fait ce sont deux choses différentes. Dès l'origine en fait, quand j'ai développé Opentheso c'était pour qu'il soit générique, parce que c'était pas un besoin qui était que pour Frantiq, pour les Pactols, donc c'est pour ça que je l'avais orienté dès le départ pour qu'il soit générique, qu'il puisse gérer plusieurs thésaurus indépendamment d'un thésaurus en particulier. Donc le logiciel il respecte la norme, donc chaque thésaurus, entre autres les Pactols, tant qu'ils respectent la norme ils peuvent être exploités et gérés par Opentheso. Donc si y a des évolutions à apporter à Opentheso, ce serait dans le cadre de la norme et non pas dans le cadre des Pactols.

Aujourd'hui, des gestionnaires de thésaurus gratuits comme Opentheso il y en a quand-même pas mal, je pense par exemple à Ginco, à VocBench, ce genre de logiciels. Comment est-ce que vous positionneriez Opentheso par rapport à ces concurrents ?

Opentheso c'était le premier. Après moi je connais Ginco parce que j'ai travaillé à la conception du cahier des charges ; les collègues du ministère de la Culture m'ont contacté pour éventuellement prendre Opentheso et le faire évoluer, après ils ont eu des choix différents, donc voilà c'est parti sur Ginco sur un autre chemin. Donc Ginco, il gère les thésaurus, il respecte la norme, mais moi je trouve qu'il est assez complexe et pas très intuitif pour ceux qui ne sont pas très spécialisés dans les thésaurus. Donc l'objectif c'était de faire un gestionnaire très simple d'utilisation et convivial et qui évolue facilement, c'est-à-dire que le code source est ouvert et il est développé dans une norme, enfin dans des technologies très accessibles à tout le monde pour que ça évolue facilement ; donc moi je pense que c'est là qu'il y a vraiment ce qui différencie Opentheso par rapport aux autres.

Vous en êtes à la version 4 actuellement, c'est bien ça ?

Oui.

Est-ce que vous avez déjà des idées, des projets pour la version 5 ?

Oui. Là on est en train de préparer la version 5. Ça va être une version qui va être la refonte complète de l'interface graphique, c'est un projet qu'on a avec Huma-Num, avec le réseau Frantiq, etc. ; donc il y a plusieurs personnes qui sont impliquées dans la réflexion, la conception, la construction des pages pour l'adapter au maximum aux besoins de tout le monde. Ça c'est pour l'interface graphique, pour la partie on va dire « cachée », ça va être l'intégration du NoSQL côté DB [base de données NDLR] pour qu'il y ait une recherche très très rapide, presque instantanée. Après, l'intégration d'un *triplestore*, et après il y a plein de choses qui vont avec, c'est-à-dire, déjà, il y a l'alignement automatique et je vais le faire évoluer encore plus pour le rendre encore plus générique, plus rapide dans le sens où on pourrait comparer des thésaurus entièrement, c'est tout ces mécanismes là qu'on est en train de changer.

Vous me disiez qu'il y a un projet d'ajouter un *triplestore*, j'ai un peu de mal à voir à quoi ça servirait d'avoir un *triplestore* dans un gestionnaire de thésaurus…

Ça serait très intéressant dans le sens où vous n'avez pas à rediffuser votre thésaurus en *triplestore*, parce qu'actuellement, c'est le cas de pratiquement tout le monde, si vous prenez le thésaurus de l'Unesco, Agrovoc, y en a plein, enfin pratiquement tous ils proposent un *triplestore* pour la diffusion, pour qu'on puisse s'aligner avec, pour qu'on puisse les consulter d'une autre façon que pour l'humain, pour les machines. Donc on aurait nativement un *triplestore*, c'est-à-dire que quand vous construisez votre thésaurus, il est automatiquement mouliné ou reversé dans un *triplestore* qui serait accoler ou qui va avec Opentheso.

J'ai pu voir dans la présentation que vous avez à Semweb pro en 2017 que vous aviez pour projet d'ajouter des identifiants Handle, et du coup je me demandais, pourquoi faire le choix d'avoir à la fois des Handle et des Ark ?

Alors, pourquoi ?… Parce que ark c'était le premier choix d'Opentheso, c'était un besoin pour beaucoup de partenaires, beaucoup de personnes qui utilisaient, utilisent actuellement Opentheso et qui ont choisi cet identifiant de type Ark, et après, en s'associant avec Huma-Num, comme Huma-Num utilise Handle, donc c'était un de leur souhait, un de leur besoin qu'OpenTheso puisse gérer les Handle aussi. Donc c'était dans ce sens là, et maintenant c'est fait, donc c'est en traduction en tout cas.

J'ai pu voir aussi que sur Opentheso on peut rechercher directement dans le thésaurus de deux façons possibles : la façon que je dirais « classique » et une autre qui est « sans mise à jour de l'arbre », en fait j'ai pas très bien compris la différence entre les deux...

La recherche classique c'est la recherche où on peut filtrer, c'est-à-dire qu'on peut chercher sur tel domaine, un mot qui commence par... se termine par... enfin, c'est ce que j'appellerais la recherche classique. Il y a une autre recherche qui fait un pointage rapide des concepts dans l'arbre, c'est-à-dire que pour les professionnels qui gèrent les thésaurus, ils n'ont pas à aller chercher à chaque fois dans la recherche classique, ils perdent du temps, donc ils vont dans cet index rapide, donc il tape le mot qui commence par… et on va avoir dans les résultats tout ce qui contient ce mot qu'on a tapé. C'est vraiment plus un pointage qu'une recherche, on vérifie seulement si les concepts existent mais on se positionnera plus tard dessus.

Dans le contexte de mon stage actuel je travaille sur Ginco, et au niveau de l'interface public de Ginco on peut avoir accès aux documents de notre portail documentaire qui ont donc été indexé avec le thésaurus ; est-ce que c'est par exemple un développement que vous souhaiteriez ajouter à Opentheso ?

C'est déjà fait. C'est en fonction pour le réseau Frantiq, y a tous les documents indexés avec les mots-clefs, ils sont déjà alignés, et on a un lien pour y accéder. Il y a même plusieurs

façons de lier des bases de données, on peut les lier par Z39-50 c'est-à-dire le format des bibliothèques, par des bases de données… et tout ça c'est paramétrable, c'est-à-dire que l'utilisateur peut le faire sans passer par un développement particulier.

J'ai vu aussi que parmi les mises à jour qui ont été faite sur Opentheso, il y a notamment maintenant la possibilité d'aligner les concepts du thésaurus avec des coordonnées GPS, comment vous est venue ce besoin, cette idée de rajouter cette fonctionnalité ?

C'était une demande d'un partenaire, ou enfin d'un réseau, qui m'a demandé si l'on pouvait adapter Opentheso pour qu'il accepte ce genre de trace. Donc moi j'ai vu que c'était pas très compliqué, qu'il suffisait juste de rajouter un petit module et voilà, je l'ai fait dans ce sens là. Je fais toujours une petite enquête avant auprès des utilisateurs, et finalement c'était un besoin pour tout le monde, donc ce qui a fait accélérer les choses c'est que c'était pas que pour une ou deux personnes, c'était quelque chose d'important.

Par exemple sur Isidore j'étais sur une notice, à propos du roman Fantomas il me semble, et qui avait notamment été écrit par un certain « Pierre Quelquechose », et dans les enrichissements qui ont été proposés, notamment au niveau de Pactols ils avaient proposé le concept « roche » alors que la notice ne traitait absolument pas de roche ou de pierre. Moi je pense que la confusion a été faite parce que l'auteur s'appelait « Pierre » et donc le lien sémantique ne s'est pas fait entre la notice et l'enrichissement, et du coup je me demandais si, j'en sais absolument rien, mais si au niveau de la structuration des données ce genre de choses peut être évité, s'il existe des solutions pour réussir à améliorer un peu la compréhension sémantique ?

Oui effectivement, c'est plutôt au niveau de Isidore qu'il faudrait arriver à cadrer mieux ce genre d'indexation ou d'enrichissements. On pourrait facilement, enfin facilement, on pourrait éviter ce genre de confusion par le fait qu'un concept il est très riche, moi je les fourni d'une façon assez riche, c'est-à-dire qu'on a le terme générique, on a les synonymes, les termes non préférés, on a les termes associés, on a des notes, tout ça c'est autour d'un concept. Il faut exploiter cet environnement pour être sûr et certain du contexte qu'il faut appliquer à ces concepts.

J'aurais voulu savoir aussi si vous aviez connaissances de moteurs de recherche autres que Isidore et que celui du catalogue indexé du réseau Frantiq, qui exploitent également des langages développés sur Opentheso ?

Autre que Isidore ? Nous on exploite Koha. En fait on intègre toute l'indexation qui arrive d'Opentheso pour enrichir les notices et permettre l'indexation en temps réel. Après y a aussi l'Opac, on utilise pas l'Opac de Koha, on utilise un autre Opac qui est développé sous Drupal et qui, pareil, qui est indexé avec le thésaurus qui arrive par Opentheso par le web service. Après y a d'autres façons de faire, y en a qui utilisent des bases de données xml pour faire de l'édition numérique, y en a d'autres qui travaillent sous Wordpress ; donc y a différentes façons de faire pour exploiter ces données-là.

Annexe 3 : retranscription de l'entretien téléphonique avec Isabelle Donze, bibliothécaire et responsable du thésaurus GeoEthno, bibliothèque Eric-de-Dampierre.

Date : 19 juillet 2018

Durée : 40 minutes 12 secondes

I.D : Donc en fait GeoEthno c'est complètement différent de Opentheso, d'un point de vue technique ça fonctionne pas du tout de la même façon, parce qu'en fait à proprement parler Opentheso c'est un logiciel, qui a été développé par un informaticien, qui gère spécifiquement un référentiel de type thésaurus qui se conforme à la norme des thésaurus, alors que GeoEthno c'est une application qui est beaucoup plus « maison » on va dire, c'est-à-dire qui a été développée sans moyen informatique et qui est une base de données XML native, donc sur une dtd assez ancienne et donc c'est pas à proprement parler un logiciel, si vous voulez. C'est différent d'un point de vue informatique, c'est beaucoup plus artisanal comme fonctionnement GeoEthno.

C.B : Pour parler un petit peu de l'origine du projet GeoEthno, j'aurais voulu savoir pourquoi est-ce que vous avez fait le choix, au début du moins, de limiter le thésaurus uniquement aux concepts géographiques ?

Alors en fait c'est parce que à l'origine quand j'ai commencé à m'occuper de l'indexation documentaire, donc ça c'était en 2000-2001, je travaillais sur la base de données de la bibliothèque, donc le catalogue de la bibliothèque qui comporte plusieurs champs d'indexation : donc y a un champ d'indexation thématique, un champ d'indexation des noms d'ethnies et un champ d'indexation des noms géographiques, chaque champ est bien séparé, et donc ça génère, ça générait parce qu'il y a plusieurs logiciels de SIGB qui ont été utilisés à la bibliothèque, et à l'origine ça ne tournait pas sur le même SIGB qu'aujourd'hui, Koha, et donc y avait création, au fur et à mesure qu'on indexait des documents avec des mots-clefs, y avait la création de ce qu'on appelle un index, et donc moi je suis partie de ces index pour créer le thésaurus. J'ai démarré par la géographie tout simplement parce qu'au départ, un peu naïvement, je supposais que ce serait plus simple pour démarrer, pour mettre au point des techniques, que d'attaquer d'emblée l'indexation thématique ; parce qu'en fait le projet

c'est aussi, surtout, de faire l'indexation thématique qui n'est pas du tout structurée pour l'instant. C'est un chantier qui a une dimension beaucoup plus importante parce que là il y a un index relativement énorme, ça c'est beaucoup accumulé, et donc je pensais démarrer par quelque chose de plus « limité » on va dire pour mettre au point, au départ, des techniques. En fait ça s'est avéré plus complexe que prévu, donc ça a pris beaucoup de temps etc. de fil en aiguille on a développé des technologies pour ça, donc après on s'est retrouvé à coopérer avec différents projets ; donc du coup, voilà, ça a pris une dimension qui n'était pas du tout prévue au départ. L'aspect géographique était sensé être beaucoup plus modeste comme projet au départ, et finalement c'est devenu un petit peu un terrain d'expérimentations un peu technologique, donc voilà, ça a pris une dimension beaucoup plus importante en fait.

J'imagine que c'est donc ce contexte d'expérimentation qui fait que, par exemple, la visualisation en arborescence est pour l'instant limitée uniquement à la partie géographique. Est-ce que vous pensez l'étendre à la partie ethnologique aussi ?

Alors ça c'est un autre aspect, un peu différent, c'est plus réellement un problème lié à de l'expérimentation, parce qu'évidemment l'aspect ethnonymes reprend les mêmes technologies que l'aspect toponyme, donc d'un point de vue technique y a pas vraiment de problème, vu que les techniques ont été développées pour l'aspect géographique ; mais pour l'instant c'est un choix politique de ne pas créer cette arborescence parce que la manipulation des ethnonymes était plus complexe que celle des toponymes. Ça pose pas les mêmes types de problèmes et donc pour l'instant on permet seulement l'interrogation par la base de données mais effectivement il n'y a pas d'arborescence, et je ne pense pas d'ailleurs que ça changera dans un futur proche. Cela dit on indique les grandes sections qui ont été basées sur un système qui répertorie des familles linguistiques, ce système est indiqué, on peut le trouver etc. donc les gens peuvent avoir accès quand-même au système de base, à l'organisation de base en fait. Ça s'appelle le « système lingosphère », c'est un système qui répertorie des familles linguistiques.

D'accord. Et donc c'est un système que vous utilisez également pour les toponymes ?

Non, c'est pour les familles linguistiques donc ça va être que pour les ethnonymes. Mais tout ça, si vous voulez, c'est indiqué sur le site du thésaurus ; les gens qui cherchent ils peuvent s'apercevoir que derrière y a effectivement une base de données qui a été développée, qui s'appelle le « système lingosphère », etc. avec différentes structures.

J'ai remarqué aussi que vous aviez intégré une ontologie dans le thésaurus, une ontologie des objets, comment vous avez eu cette idée ?

A l'origine le but du projet c'était pas tellement de faire un thésaurus, c'était il y a une quinzaine d'années, c'était plutôt de faire une ontologie, mais toutes ces questions étaient très nouvelles à l'époque et donc il n'y avait pas vraiment de logiciel qui était développé, c'était un peu compliqué. Donc finalement nous on s'est un peu rabattu sur l'aspect thésaurus, finalement la technique du thésaurus est plutôt bien adaptée. Mais du coup j'ai voulu quand-même conserver cette idée de l'ontologie, et je me suis dit que je pouvais l'utiliser pour, en quelque sorte, documenter la logique du système ; finalement c'est un petit peu l'organisation de la dtd qui se retrouve là sous la forme d'ontologie. Donc ça reste très très formaliste, c'est juste une sorte de description de l'organisation logique du système, c'est une ontologie de haut niveau, dedans on peut pas trouver les termes eux-mêmes. Ça reste uniquement un système de description logique d'assez au niveau. J'ai fait ça parce que j'avais la capacité de le faire, c'était une démarche qui m'intéressait et puis que je trouvais intéressant à mettre en ligne aussi pour les gens qui peuvent éventuellement faire des recherches là-dessus, même dans un but pédagogique d'ailleurs. Moi je m'en sers personnellement quand je sais plus exactement, je ne connais pas la dtd par cœur et on fait aussi des petites modifications donc des fois j'utilise ça pour rechercher de la documentation en fait sur les différents éléments disponibles, etc. C'est assez commode et puis c'est une façon aussi de faire des choses un petit peu expérimentales sur le système. Mais ça ne va pas plus loin que ça !

D'accord, donc vous ne souhaitez pas l'étendre à tous les champs du thésaurus par exemple ?

Alors, les champs du thésaurus, non, parce qu'en fait là si vous voulez c'est plutôt une description toponymique en fait, qui correspond à la façon dont on décrit le réel, c'est-à-dire là en l'occurrence, la toponymie, l'objet terrestre ; donc pour l'instant ça reste uniquement à ce niveau là, oui. Le thésaurus sinon, lui, il a son système de documentation dans la petite rubrique « En savoir plus », c'est là qu'il y a toutes les explications plutôt basées sur la notion de thésaurus, toutes les caractéristiques habituelles du thésaurus ; donc voilà, j'ai pris ce système-là. Après tout ça ça évolue, de toute façon on change tout le temps de logiciel, de système, etc. donc à chaque fois il faut redécouvrir, c'est pas très stable toutes ces choses.

Du coup je me demandais : le thésaurus GeoEthno vous le faites évoluer à quelle fréquence à peu près ?

Alors en fait il y a trois niveaux. Il y a la base de données elle-même, donc c'est une base de données Xml qui n'est donc pas en ligne, c'est en quelque sorte la zone de chantier ; donc ça ça évolue en permanence en fonction des besoins, c'est-à-dire, moi si vous voulez, mon travail principal c'est l'indexation documentaire du fonds de la bibliothèque Eric-de-Dampierre, régulièrement, tous les jours je fais des séances d'indexation, et y a régulièrement du nouveau vocabulaire, des nouveaux lieux etc. Donc après il faut arbitrer, savoir s'il faut créer un nouveau descripteur, est-ce que ça vaut le coup, etc. donc il y a différents critères. Si je décide de créer un nouveau descripteur c'est très rapide, je vais sur la zone de chantier et c'est immédiat, donc ça c'est rentré dans mon système. Ce fichier xml, par période, j'attends qu'il y ait un certain nombre de modifications, en gros tous les ans il y a un chargement d'une mise à jour sur le site web, et puis tous les ans, tous les deux ans, il y a aussi une mise à jour dans le système Isidore. Sachant que moi, pour mon travail quotidien, les informations évoluent de façon immédiate, ma base de données elle est tout le temps à jour.

Est-ce que vous avez déjà des projets ou des idées pour les développements à venir sur ce thésaurus ?

Pour l'instant au niveau de GeoEthno on est plutôt dans une phase relativement stable. La priorité c'est d'améliorer les choses au niveau du système Isidore, c'est-à-dire en fait de travailler sur la pertinence du référentiel, et donc régulièrement j'essaie d'aller faire des tests sur Isidore pour voir comment les choses répondent, etc. En plus il y a un développement international d'Isidore en ce moment, assez orienté vers le langage espagnol, etc., donc voilà pour moi c'est un peu prioritaire d'améliorer la qualité du référentiel dans Isidore, parce que là il y a beaucoup de choses qui sont… Sur GeoEthno j'ai essayé de faire un assez gros travail au niveau de la pertinence, donc maintenant c'est correct on va dire. C'est vrai que les autres référentiels des fois il y a des choses qui ne sont pas géniales, mais bon, ce sont des grosses machineries qui ne sont pas faciles à mettre à jour non plus, donc chacun avance un petit peu à son rythme. Donc voilà, améliorer la qualité sur Isidore ça c'est quand-même important, et puis sinon moi ma priorité actuelle c'est de travailler sur l'aspect thématique, l'indexation thématique, donc là on sort de GeoEthno, c'est tout un chantier qui s'est mis en route et donc voilà, j'essaie de construire une sorte de gros thésaurus pour le vocabulaire que nous avons, donc voilà ce sont un petit peu nos deux priorités actuelles.

GeoEthno là c'est plutôt une mise à jour, c'est-à-dire par exemple là pour vous donner une idée, je me tiens un petit peu au courant de l'actualité par rapport à ça et j'ai appris qu'il y a un pays qui s'appelle la « Macédoine », c'est un ex-pays de la Yougoslavie, y avait un conflit depuis plusieurs années maintenant, sur le nom de ce pays parce qu'en fait il y a une rivalité avec la Grèce par rapport au nom « la Macédoine », ça dure depuis plus de dix ans ; et là apparemment ils auraient trouvé un terrain d'entente, donc ils auraient trouvé un nom pour ce pays qui s'appellerait je crois « la Macédoine du Nord » quelque chose comme ça. Donc j'ai vu passer cette information dans la presse et c'est quelque chose qui est important parce que c'est un nom structurel, c'est un nom de pays donc on est sur un niveau de base en fait, un niveau structurel dans le système. Donc j'attends de voir les mises à jour dans les systèmes de l'Onu, parce que c'est basé sur les bases de données de l'Onu, Wikipédia aussi, je surveille un petit peu comment est-ce qu'ils vont mettre ça à jour, et puis le moment venu il faudra faire le changement dans le système qui sera là assez important puisqu'il s'agit d'un nom de pays. C'est un suivi pour essayer de faire en sorte que l'information soit pas trop

obsolète trop vite, sachant qu'en fait, évidemment, c'est impossible de tout suivre puisque moi je ne fais pas que ça non plus, donc de toute façon le thésaurus est en permanence un petit peu en retard par rapport à l'actualité. Il y a des changements continuels un peu partout dans le monde, ce sont des choses qu'il faut suivre pour essayer de ne pas trop laisser s'introduire des choses qui ne sont pas à jour ; donc régulièrement je balaye les pays dans lesquels j'ai vu qu'il y avait des changements pour voir s'il y a des choses à modifier.

Maintenant je vais vous parler un petit peu plus du fonctionnement avec Isidore. J'ai parcouru quelques ressources indexées dans Isidore et il me semble que, pour l'instant, il n'y a que les toponymes, il n'y a pas la partie ethnonymes, c'est bien ça ?

Oui, il n'y a pas la partie ethnonymes. En fait ce qui se passe c'est que dans Rameau il y a un certain nombre d'ethnonymes parce que dans Rameau il y a tout, c'est un système un peu encyclopédique, par contre ils ont plus ou moins débranché l'aspect géographique, il y a beaucoup moins de toponymes qu'avant puisque donc il y a GeoEthno qui est là, donc c'est moins utile, mais il y a des ethnonymes qui répondent de temps à autre quand-même et pour l'instant c'est un peu la même problématique que pour l'arborescence, dans l'immédiat j'ai pas prévu de rajouter ça dans Isidore. Après c'est peut-être une question à revoir mais dans l'immédiat c'est pas d'actualité, je pense que ça poserait plus de problèmes de pertinence que ça n'en résoudrait en fait. Après c'est une question de justement connaître le fonctionnement du système d'Isidore pour travailler sur la pertinence, c'est une question complexe en fait, donc pour l'instant les ethnonymes ne sont pas chargés.

Je me demandais aussi, dans les enrichissements d'Isidore il y a donc la partie « enrichissement » avec les thésaurus, et il y a aussi une partie « espace géographique » qui est, elle, alimentée uniquement par GeoNames, et donc j'ai pu remarquer qu'assez souvent il y a des doublons entre ce qui ressort de GeoNames et de GeoEthno, mais du coup comme c'est pas dans la même partie des enrichissements je me demandais si, je sais pas, si c'est vous qui avez fait la démarche de vouloir être dans les enrichissements et pas dans les espaces géographiques ou si…

Alors non, en fait c'est une politique d'Huma-Num, c'est-à-dire que quand ça s'est construit il y a maintenant pas mal de temps, donc au départ ça s'appelait même « Adonis », la première chose qu'ils ont fait ça a été de se brancher sur GeoNames pour avoir la géolocalisation puisque c'est un petit peu le système de base que tout le monde utilise dans tous les systèmes. GeoNames est complètement différent comme logique d'organisation des données et ils se sont aperçu que c'était trop riche, trop complexe et que donc ça générait beaucoup trop de bruit. Donc du coup ils ont réduit le niveau d'agrégation des données c'est-à-dire que maintenant ce n'est pas la totalité de GeoNames qui est chargée dedans, ce qui n'était pas le cas au départ. Au début ils avaient mis tout GeoNames et puis en fait ça a fait exploser tout le système ; donc là ils sont revenus à seulement une partie de GeoNames, et parallèlement comme ils voulaient aller chercher des référentiels ils ont récupéré GeoEthno comme ça, ce qui fait qu'au final après plusieurs années de tests de tout ça ils sont arrivés à un certain équilibre entre les deux. C'est pas inintéressant de comparer les deux : à chaque fois qu'on a des réponses est-ce qu'on touche bien la même chose ? Des fois il y a des erreurs de chaque côté, c'est pas mal d'avoir les deux ; eux l'avantage c'est qu'il y a la géolocalisation derrière, mais des fois ils ont aussi des problèmes. C'est pas simple ces systèmes-là.

Du coup la géolocalisation, c'est pas quelque chose que vous souhaiteriez ajouter dans GeoEthno ?

Si, le cas échéant, d'ailleurs il y en a déjà un petit peu au niveau des capitales d'états, c'est ce qui permet d'avoir un début d'orthographie dans le site web parce qu'en fait il y a des cartes et ça permet déjà d'interroger sur les capitales de pays et à partir de là on peut rebondir sur les bases de données de la bibliothèque. Mais ça reste effectivement au niveau des capitales

parce que sinon il faudrait aller faire un match entre les termes et GeoNames et ça c'est ce qu'ils sont entrain de faire chez Frantiq avec les Pactols mais ils ont embauché une équipe spéciale de l'Inist pour faire ce travail qui est en cours et qui est un travail semi-manuel, donc il y a des gens qui ne font que ça en ce moment à l'Inist. Maintenant je me dis que comme maintenant sur Isidore on a GeoNames qui est couché parallèlement, c'est plus une priorité ; à un moment donné j'y pensais mais là du coup maintenant c'est pas vraiment une priorité en fait. Ça se fera peut-être, je ne sais pas du tout. Après il y a peut-être d'autres projets qui seraient plus intéressants pour nous, ce serait de travailler sur une micro-zone : on aurait par exemple tout un corpus documentaire très spécialisé et là on pourrait géolocaliser des zones géographiques très très précises, et ça interrogerait derrière ces fameuses ressources documentaires ; ce serait plus un travail comme ça, dans la précision. Donc voilà, il y a plusieurs idées comme ça, après c'est une question de priorité.

Sur la documentation de GeoEthno il est expliqué que vous avez gardé les descripteurs qui ne sont plus utilisés, et du coup je me demandais déjà pourquoi est-ce que vous avez voulu les garder, et est-ce que vous utiliser un système pour les exclure lorsque vous mettez GeoEthno sur Isidore ou est-ce que vous chargez tout sur Isidore peu importe si le descripteur n'est plus utilisé ?

Alors en fait il y a deux sortes de choses. C'est assez complexe de gérer l'historicité des descripteurs, c'est tout en aspect technique qui n'est d'ailleurs pas pris en compte dans la norme Skos et ça pose d'énormes problèmes parce ce que les gens qui ont conçu cette norme avaient plus un profil d'informaticiens que de gestionnaires de thésaurus et donc il y a des choses qu'ils n'ont pas suffisamment détaillé. C'est un aspect important la gestion de l'historique, en quelque sorte du *versioning*, du terme. Il y a plusieurs cas de figures : il y a des termes qui sont toujours valables mais qui ont eu un problème d'orthographe, etc., donc on change de terme, l'ancien terme change de statut et devient un « non-descripteur » donc là il bascule en « employé pour » donc ce terme là on ne l'utilisera plus. Il y a un autre type de terme, ce sont les termes de géographie historique, donc ça se sont des entités géographiques réelles qui changent de nom ; c'est pas le terme lui-même qui a eu un problème et on se dit qu'il faut le changer parce qu'il y a quelque chose qui ne convient pas, là c'est l'actualité, typiquement ça va être le cas avec la Macédoine : on va générer un terme,

alors je crois que ça s'appelle actuellement « ex-république yougoslave de Macédoine », donc ce descripteur-là va devenir un terme de géographie historique. Le choix que j'ai fait par rapport à ça c'est de conserver ces termes-là comme étant des descripteurs, alors d'un statut un peu différent des autres, mais qu'on puisse quand-même mettre dans l'indexation de la base de données pour situer un petit peu les notices et les documents dans le temps, c'est-à-dire par exemple tous les documents qui ont été écrits sur la Macédoine dans les années 2000 eh bien ils auront ce nom « ex-république Yougoslave de Macédoine » parce qu'à l'époque ça s'appelait comme ça. Donc voilà, la gestion de la géographie historique c'est quelque chose d'assez complexe à gérer mais ça permet, pour la thématique qu'on a nous ici qui est l'ethnologie, c'est très important parce qu'il y a beaucoup de pays et de territoires qui ont changé de nom notamment pendant la période post-coloniale, avec la décolonisation les pays ont changé de nom, en Afrique il y a beaucoup de pays qui ont changé de nom et donc c'était important de conserver l'ancien nom avec toute l'explication qui correspondait à cette période-là. Donc il y a un peu deux styles de gestion de l'historique des termes. Les termes historiques, eux, ils restent en tant que descripteurs, par contre dans Isidore, à l'heure actuelle, ils sont basculés en tant que non-descripteurs ; mais ça c'est pareil ça peut typiquement faire partie des choses pour améliorer la pertinence : le travail sur ces termes historiques qu'on peut essayer de mettre plus en valeur ou pas, après c'est une question de choix, si on veut les faire ressortir ou pas.

Annexe 4 : retranscription de l'entretien téléphonique avec Thierry Bouchet, en charge de la conversion de Rameau au format Skos, Centre national Rameau.

Date : 07 août 2018

Durée : 18 minutes 12 secondes

T.B : Il y a un premier projet qui a eu lieu, la date exacte ça doit être 2008-2009 je pense, avec Antoine Isaac qui a fait une première transformation de Rameau en Skos, vous connaissez ce langage ?

C.B : Oui.

D'accord. Skos c'est vraiment basique, ça permet de créer des liens sémantiques, des notes d'application, de définition, on peut faire des équivalences dans d'autres langues. Si vous voulez la transformation elle est vraiment a minima, le *mapping* et le transfert d'information de Rameau, qui est en Intermarc, en Skos est vraiment *a minima*, et puis surtout Skos ne gère pas la syntaxe de Rameau.

Le système de vedettes ?

Oui voilà, c'est compliqué pour faire des chaînes, spécifier des subdivisions géographiques, de périodes ; c'est assez minimal.

Après, le traitement au niveau de la plate-forme Isidore je ne le connais pas, je pense qu'ils récupèrent le Rameau en Skos depuis le nouveau catalogue, depuis data.bnf. Sur ça je ne sais pas du tout comment se fait la transformation, enfin, il y a une interface Sparql et donc avec les bonnes requêtes on peut extraire les vedettes Rameau de data.bnf.

Par rapport à la transition bibliographique je ne pourrais pas vous en dire beaucoup plus parce que j'ai été absent deux mois et en fait les documents de référence vont paraître au mois de septembre.

Après je sais qu'il y a des outils pour faire des alignements entre vocabulaires. Moi j'ai fait un alignement de la partie thermodynamique de Rameau en Skos avec le vocabulaire de l'Inist, c'était une étude de cas ; en gros, j'ai fait des scripts en Python sur les chaînes de caractères. L'Inist avait son langage en Skos, ses vedettes-matières, mais pas les liens sémantiques, donc

ça leur a permis d'avoir des liens sémantiques ainsi que des sources puisqu'ils n'avaient pas de sources non plus.

Voilà, je ne sais pas trop ce que je peux vous dire d'autres, à moins que vous ayez des questions assez précises ?

Oui, j'ai des questions un petit peu plus précises. Pour revenir au premier projet en 2008-2009, le premier projet de conversion de Rameau, c'était donc dans le cadre du projet européen « TelPlus » si j'ai bien compris, est-ce que vous pouvez m'en dire un petit peu plus sur les ambitions que vous aviez pour ce projet ?

En fait c'était plutôt une étude de faisabilité et une proposition de la part d'Antoine Isaac de la Vrije université à Amsterdam parce qu'il a fait partie du comité au W3C qui développait Skos. Skos était encore à l'état de *draft*, donc lui ça lui a permis de développer la spécification du langage ; nous en fait on ne l'a jamais exploité, son travail n'a jamais été exploité au sein de la BNF, dans un workshop pour les utilisateurs lambdas. Lui avait mis ça sur un serveur avec une interface de requête basique.

Et donc du coup, la raison pour laquelle vous ne l'exploitiez pas à la BNF c'était parce que la conversion ne prenait pas toute la complexité du langage, c'est ça ?

Non, c'est juste qu'en fait je pense que data.bnf n'était pas encore lancé, parce que par la suite quand on a créé data.bnf il y a eu un ensemble d'outils pour extraire ça en RDF, sous différents types de langages, en Json ou des choses comme ça. Mais dans tous les cas Skos n'arrivait pas à exploiter la syntaxe, donc même en l'état actuel la syntaxe n'est pas exploitée par Skos dans data.bnf et d'ailleurs elle ne le sera jamais. C'est pas le but de Skos de faire un langage pré-coordonné.

Mais justement, dans le cadre de la transition bibliographique, l'impression que j'ai d'assez loin c'est que le langage Rameau est en train de se « dé-coordonner » si je peux dire ça comme ça…

Oui, en fait on va surtout le simplifier parce qu'on a atteint un niveau de complexité où on a des listes de subdivisions pour les ethnonymes, pour les produits chimiques, etc. et en fait on va permettre aux gens d'utiliser du texte libre avec différents types de concepts pour qui soient utilisables pour les personnes, à tout type de noms communs et à des concepts géographiques. En fait on va réduire la complexité du langage.

Même une fois que cette simplification sera faite, vous pensez quand-même ne pas partir sur du Skos ?

C'est-à-dire que le Skos il sera toujours utile pour que les gens puissent faire des requêtes dans data.bnf, mais Skos ne pourra jamais être utilisé pour créer des chaînes d'indexation. C'est pas le but du langage au départ. En fait Skos ça peut être utilisé pour faire un dictionnaire, la chose la plus simple, comme pour faire un thésaurus et comme dans un thésaurus il n'y a pas de notions de chaînes construites, ça n'ira pas plus loin que ça.

J'avais vu aussi que dans le cadre de ce premier projet il y a notamment des alignements qui ont été faits avec le langage de la bibliothèque du Congrès aux États-Unis, et en fait ces équivalences de langues je sais que sur Isidore notamment on ne les retrouve pas du tout dans les enrichissements proposés par Rameau, est-ce que vous savez pourquoi ? est-ce que c'est une volonté de la BNF de ne pas les avoir gardé dans le fichier Rameau ?

Je ne sais pas du tout. Le fait est que nous les alignements avec le vocabulaire LCSH de la bibliothèque du Congrès, on les a faits en intermarc. On a aussi des alignements avec le MeSH [*Medical Subject Headings*, NDLR], avec le RVM [Répertoire de vedettes-matière, NDLR] de Laval au Canada, pourquoi ça ne ressort pas dans Isidore ? je ne sais pas.

Vous me dîtes que ces alignements ils ont été fait en intermarc ?

Je pense pas qu'Antoine Isaac puisse s'attaquer au LCSH pour les récupérer, je vois pas comment il aurait pu faire. Nous on a fait les alignements à la main, donc en intermarc.

Ce que vous me disiez c'est que, pour la version Skos de Rameau, vous ne l'utilisez et ne comptez pas continuer à expérimenter là-dessus, si ?

Ah si si, le langage sera toujours disponible en Skos !

Oui, mais juste pour pouvoir fonctionner avec data.bnf, c'est bien ça ?

Oui. En fait Skos permet d'avoir un langage du web sémantique pour récupérer des métadonnées dans une page chez data.bnf. Si vous avez une chaîne construite avec une vedette et deux subdivisions, vous allez récupérer trois entrées au format Skos.

Je voulais vous demander aussi, qui exploite Rameau au format Skos, il y a donc data.bnf, Isidore, est-ce que vous savez si d'autres moteurs l'utilisent ?

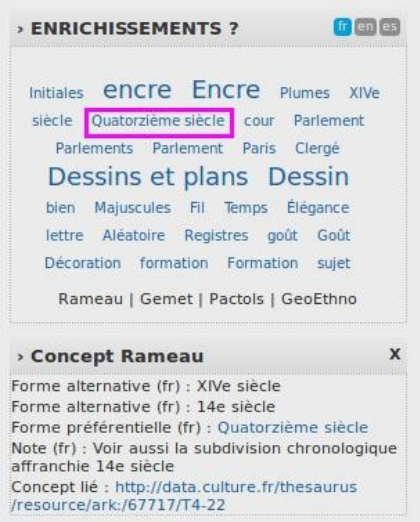
Je ne sais pas du tout. Les données sont libres d'utilisation donc n'importe qui peut faire son site web avec un moteur d'interrogation qui va utiliser notre langage en attaquant data.bnf derrière ; on peut ne même pas être au courant.

Comme projet officiel il y a les alignements qu'on a fait avec l'Inist, moi sur la thermodynamique, et j'ai une collègue qui a fait des alignements sur les poissons avec l'Inist qui avait une taxonomie des poissons assez importante. C'est un projet qui est toujours en cours, et là on est sur des volumes qui sont entre 5000 et 7000 peut-être. Là par contre c'est vraiment intéressant parce que Skos se prête très bien aux taxonomies.

Vous me disiez, au début, que vous ne savez pas du tout comment Isidore a récupéré Rameau, c'est ça ?

Non, je ne peux pas vous le dire. Comment ils ont fait ? Il suffit d'aller sur l'interface Sparql de data.bnf et de faire une moulinette et de récupérer toutes les notices. Nous, nos données elles sont libres d'accès, après les gens font ce qu'ils veulent.

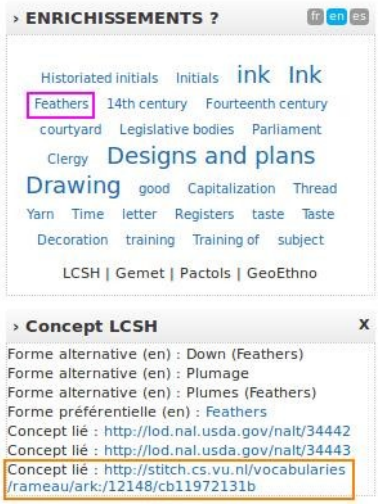
Annexe 5 : exemple d'alignement entre un concept Rameau et un de LCSH



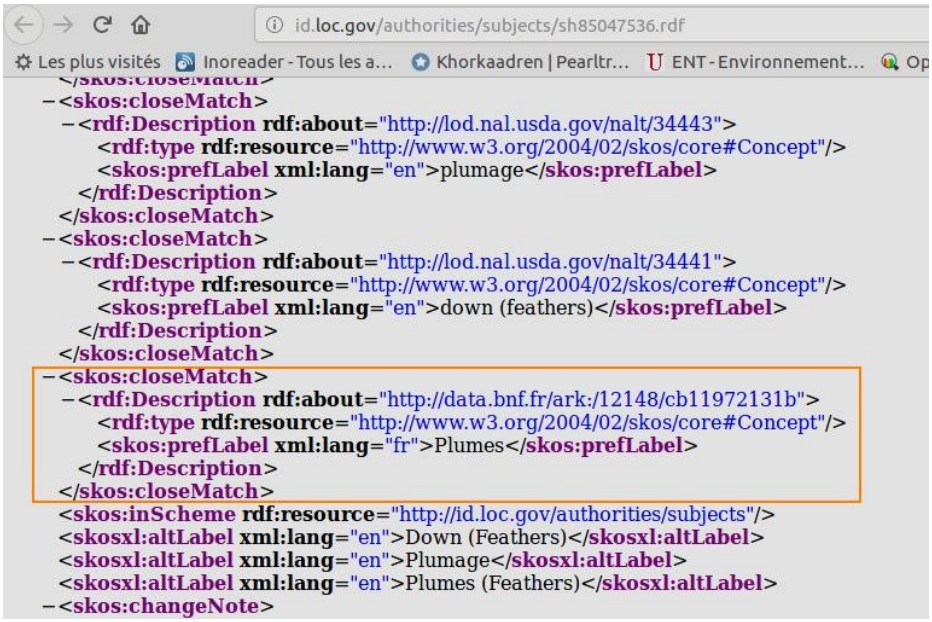
Source : Isidore. Morgat-Bonnet Monique, Des initiales ouvragées à l'encre et à la plume. *Isidore – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 19 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.zarvk5

Source : Bibliothèque nationale de France. Quatorzième siècle. *data.bnf*, [en ligne]. Mise à jour le 29 juin 2018. [Consulté le 19 août 2018]. Disponible à l'adresse : http://data.bnf.fr/11965802/quatorzieme_siecle/rdf.xml

Annexe 6 : exemple d'alignement entre un concept Rameau et un de LCSH

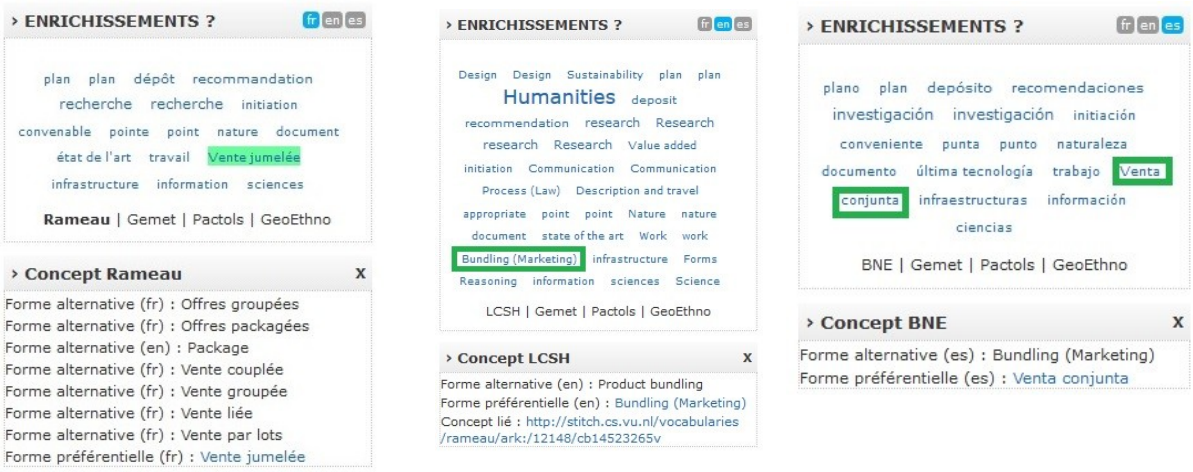


Source : Isidore. Morgat-Bonnet Monique, Des initiales ouvragées à l'encre et à la plume. *Isidore – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 19 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.zarvk5

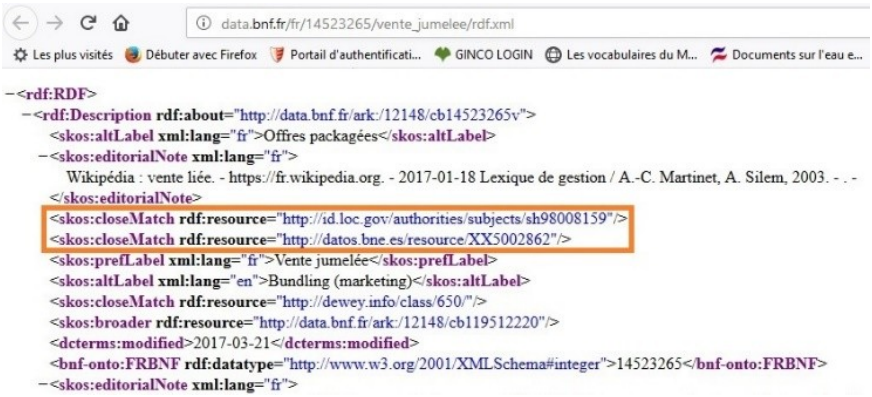


Source : Library of Congress. Feathers. *Library of Congress*, [en ligne]. [Consulté le 19 août 2018]. Disponible à l'adresse : http://id.loc.gov/authorities/subjects/sh85047536.rdf

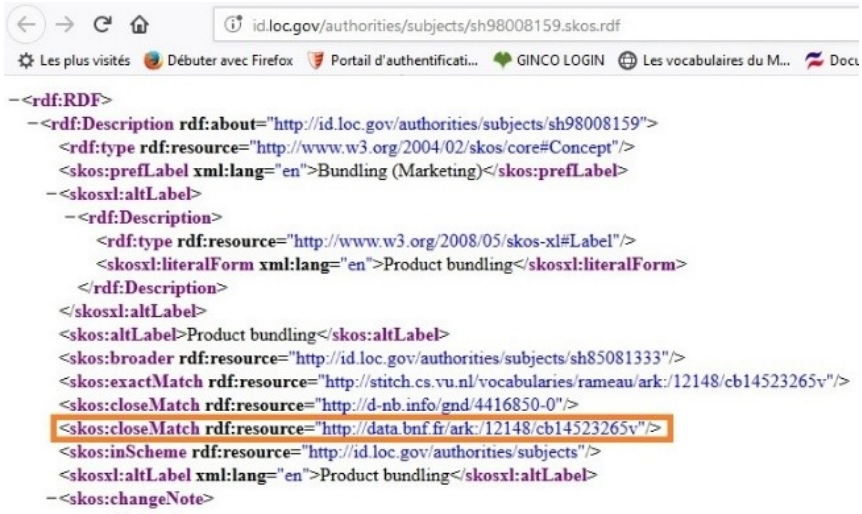
Annexe 7 : liens entre les concepts « Vente jumelée » de Rameau, « Bundling (Marketing) » de LCSH et « Venta conjuta » de la BNE sur Isidore et dans leur code XML-RDF



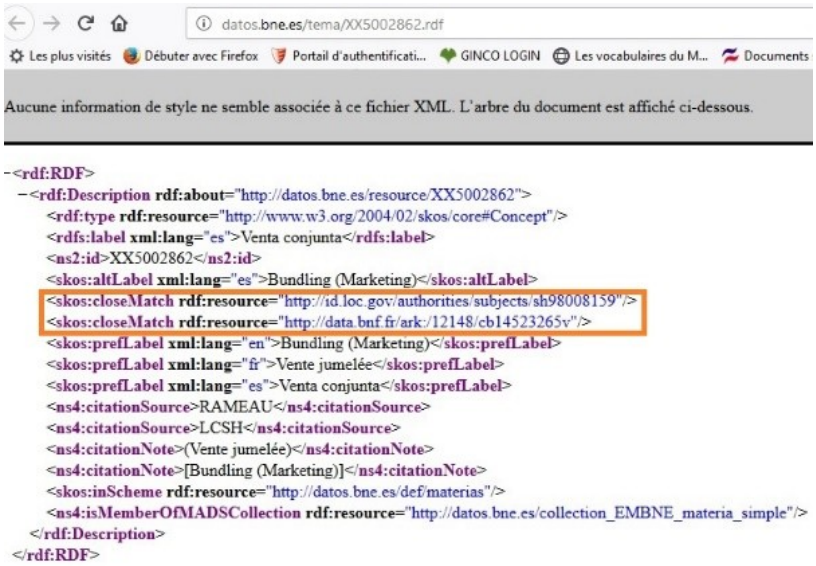
Source : Isidore. D 7.2 Design and Sustainability Plan for an Open Humanities Data Platform. *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 20 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.wd8i17



Source : Bibliothèque nationale de France. Vente jumelée. *data.bnf*, [en ligne]. Mise à jour le 29 juin 2018. [Consulté le 20 août 2018]. Disponible à l'adresse : http://data.bnf.fr/fr/14523265/vente_jumelee/rdf.xml

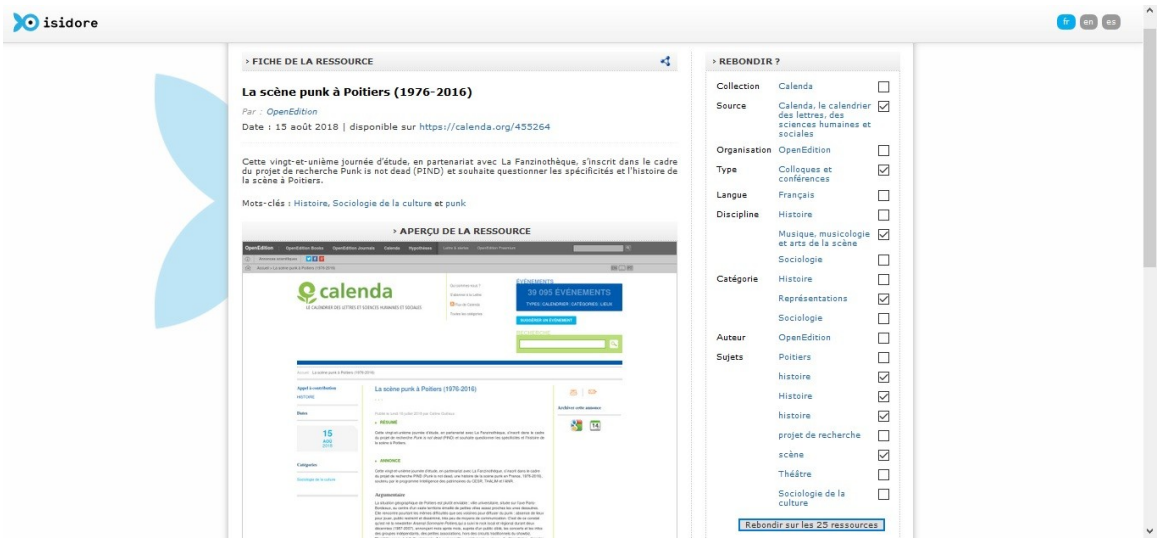


Source : Library of Congress. Bundling (Marketing). *Library of Congress*, [en ligne]. [Consulté le 20 août 2018]. Disponible à l'adresse : http://id.loc.gov/authorities/subjects/sh98008159.Skos.rdf

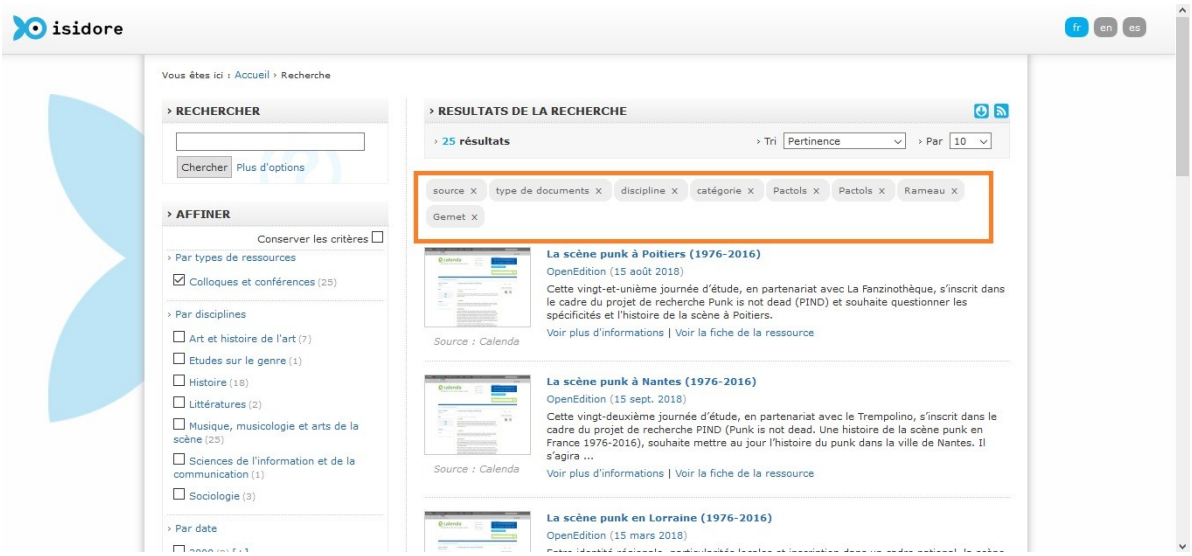


Source : Biblioteca nacional de España. Temas. *dato.bne.es*, [en ligne]. [Consulté le 20 août 2018]. Disponible à l'adresse : http://datos.bne.es/tema/XX5002862.rdf

Annexe 8 : détail des rebonds sur Isidore

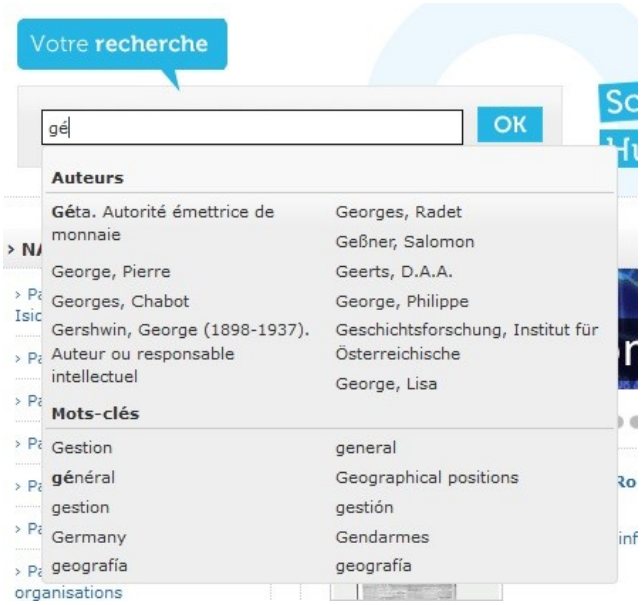


Source : Isidore. OpenEdition, La scène punk à Poitiers (1976-2016). *ISIDORE – accès aux données et services numériques des SHS*, [en ligne]. [Consulté le 15 août 2018]. Disponible à l'adresse : https://rechercheisidore.fr/search/resource/?uri=10670/1.aphqrr

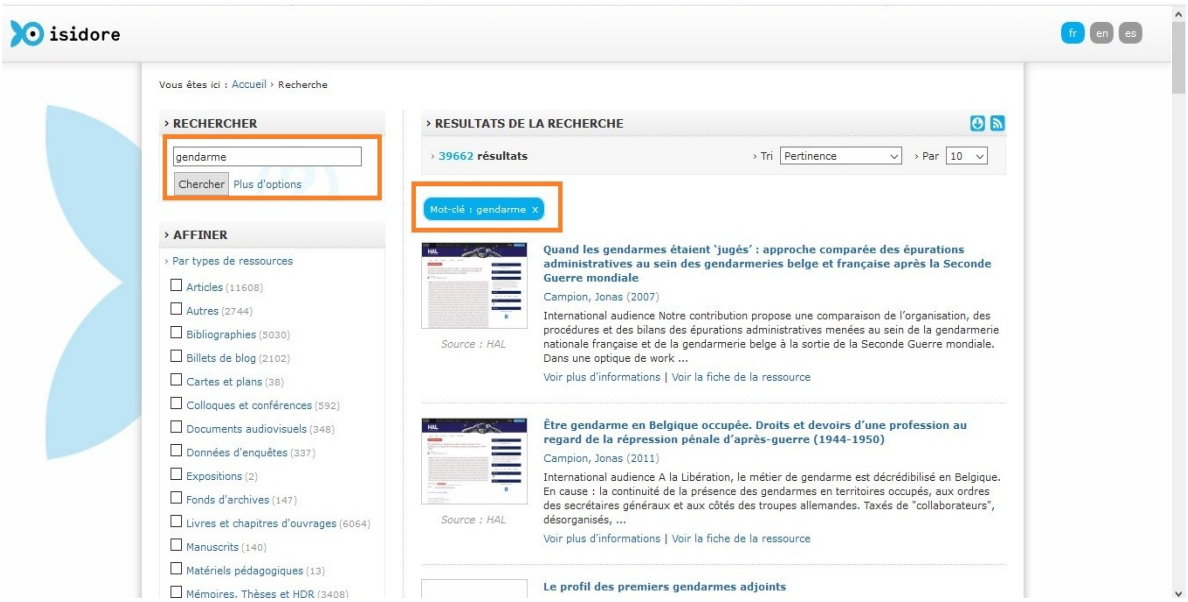


Source : Isidore. Résultats de recherche. *ISIDORE – accès aux données et services numériques des SHS*, [en ligne]. [Consulté le 15 août 2018]. Disponible à l'adresse : https://rechercheisidore.fr/search/?source=10670/2.6anse6&type=http://www.rechercheisidore.fr/ontology%23conference&subject=http://aurehal.archives-ouvertes.fr/subject/shs.musiq&taxo=http://calenda.org/categories.rdf%23categorie276&pactols=http://ark.frantiq.fr/ark:/26678/pcrtJHV6SKuS8I&rameau=http://data.bnf.fr/ark:/12148/cb119344445&gemet=http://www.eionet.europa.eu/gemet/concept/3983&pactols=http://ark.frantiq.fr/ark:/26678/pcrtUomA7N7mHk

Annexe 9 : exemple d'auto-complétion dans la barre de recherche d'Isidore



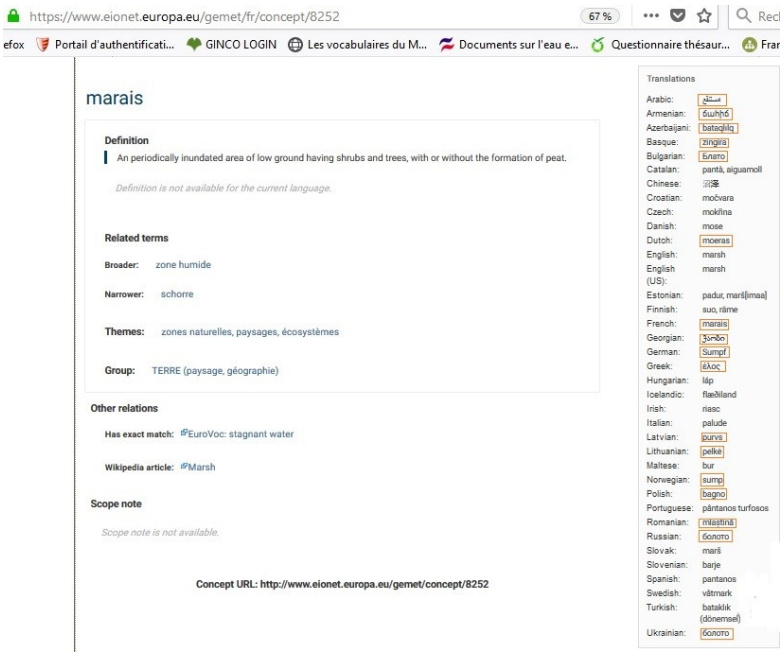
Source : Isidore. *ISIDORE – accès aux données et services numériques des SHS*, [en ligne]. [Consulté le 18 août 2018]. Disponible à l'adresse : https://rechercheisidore.fr/



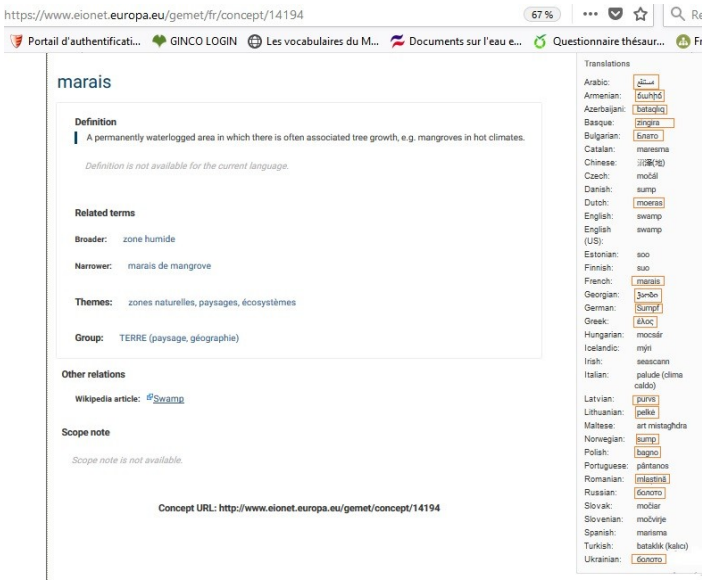
Source : Isidore. Résultats de la recherche. *ISIDORE – accès aux données et services numériques des SHS*, [en ligne]. [Consulté le 18 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/?q=gendarme

Annexe 10 : doublons du concept « marais » sur Gemet

Source : Isidore. Essai de restauration de roselières en marais dulçaquicole. *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 24 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.7aj5go



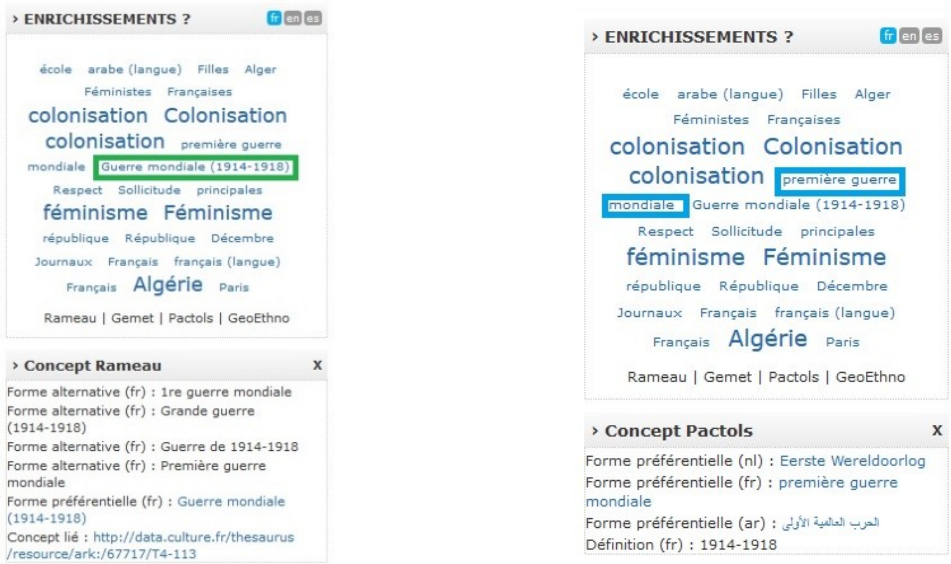
Source : Eionet. Marais. *European Environment Information and Observation network – Eionet*, [en ligne]. Mise à jour le 16 août 2018. [Consulté le 24 août 2018]. Disponible à l'adresse : https://www.eionet.europa.eu/gemet/fr/concept/8252



Source : Eionet. Marais. *European Environment Information and Observation network – Eionet*, [en ligne]. Mise à jour le 16 août 2018. [Consulté le 24 août 2018]. Disponible à l'adresse : https://www.eionet.europa.eu/gemet/fr/concept/14194

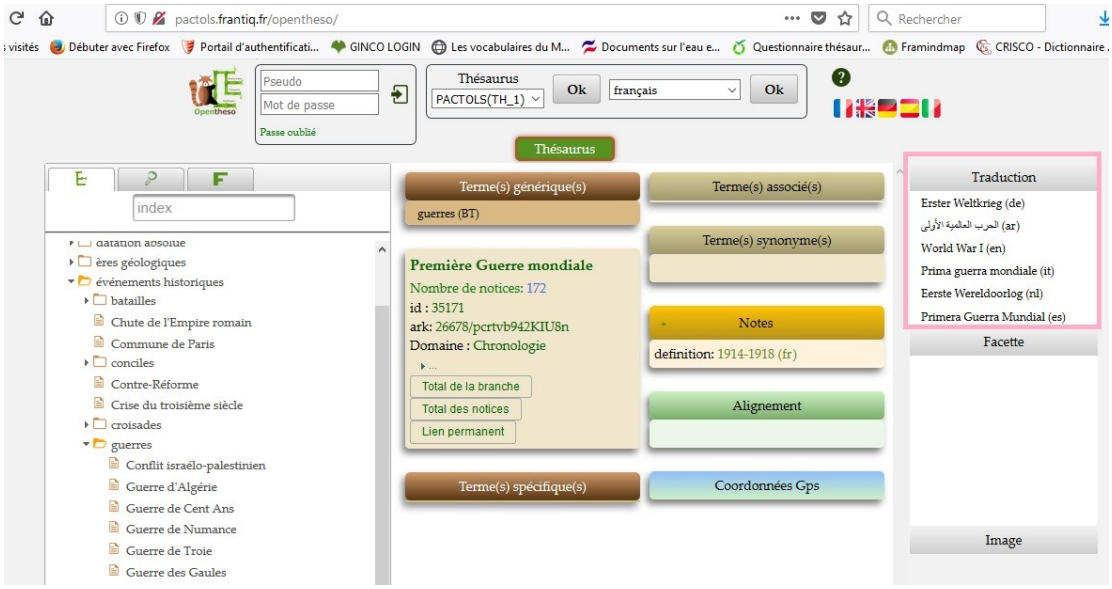
Légende : terme préférentiel identique sur les deux concepts

Annexe 11 : concept « première guerre mondiale » sur Isidore et Pactols, écarts de traductions entre les deux versions Pactols



Source : Isidore. "Ne pillons plus ! Instruisons !" : article-pétition de la féministe Hubertine Auclert pour l'ouverture d'une école arabe pour filles à Alger. I*SIDORE – Accès aux données et ressources numériques de SHS*, [en ligne]. [Consulté le 20 août 2018]. Disponible à l'adresse :

https://www.rechercheisidore.fr/search/resource/?uri=10670/1.wwsm1o



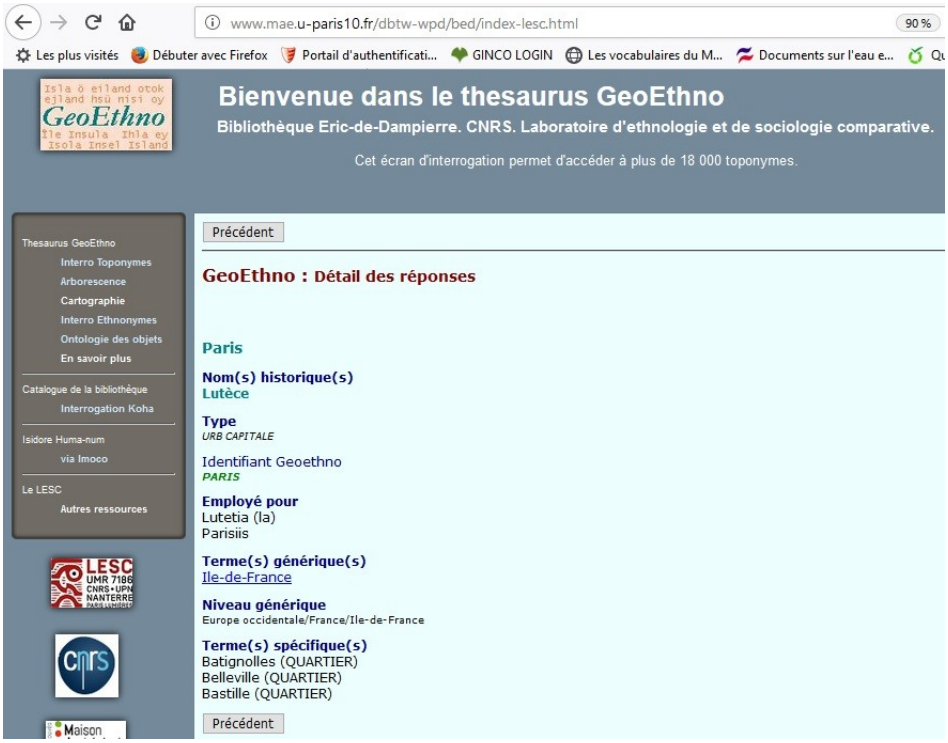
Source : Réseau Frantiq. Première Guerre mondiale. *pactols.frantiq.fr/opentheso*, [en ligne]. [Consulté le 20 août 2018]. Disponible à l'adresse : http://pactols.frantiq.fr/opentheso/

Annexe 12 : représentation du concept « Sud » issu de GeoNames sur Isidore

Source : Isidore. Les ressources en eau et le changement climatique en Provence-Alpes-Côte d'Azur. *ISIDORE – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 24 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.hb75cp

Annexe 13 : concept GeoEthno « Paris » sur Isidore et sur GeoEthno

Source : Isidore. Morgat-Bonnet Monique, Des initiales ouvragées à l'encre et à la plume. *Isidore – Accès aux données et services numériques de SHS*, [en ligne]. [Consulté le 19 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.zarvk5



Source : Bibliothèque Eric-de-Dampierre. CNRS, Laboratoire d'ethnologie et de sociologie comparative. Maison Archéologie et Ethnologie, René-Ginouvès. Paris. *Thesaurus GeoEthno*, [en ligne]. [Consulté le 24 août 2018]. Disponible à l'adresse : http://www.mae.u-paris10.fr/dbtw-wpd/bed/index-lesc.html

Annexe 14 : détail du concept Rameau « Françaises » sur Isidore

Source : Isidore. "Ne pillons plus ! Instruisons !" : article-pétition de la féministe Hubertine Auclert pour l'ouverture d'une école arabe pour filles à Alger. *ISIDORE – Accès aux données et ressources numériques de SHS*, [en ligne]. [Consulté le 20 août 2018]. Disponible à l'adresse : https://www.rechercheisidore.fr/search/resource/?uri=10670/1.wwsm1o

Table des matières

Remerciements **3**

Table des abréviations **4**

Introduction **7**

Partie 1 : état de l'art **10**

I. La recherche d'information 10
I.I La recherche d'information : entre sciences de l'information et informatique 10
I.II Des recherches d'information 12
I.III Les systèmes de recherche d'information 15
II. Du web sémantique au web des données 19
II.I Le projet initial : le web sémantique 19
II.II La notion de « sémantique » : entre centralité et ambiguïté 22
II.III Le web des données, un web sémantique 23
I.IV Les entraves à l'essor du web des données 25
II.V La recherche d'information dans le web de données 26
III. Les systèmes d'organisation des connaissances dans le web des données 28
III.I Les systèmes d'organisation des connaissances 28
III.II Les langages documentaires 31
III.III Skos : un format pour intégrer les systèmes d'organisation des connaissances au web des données 35
III.IV Iso 25964, une norme pour rendre les thésaurus compatibles avec le web des données 37

Partie 2 : méthodologie de l'étude de cas **41**

I. Présentation de la plate-forme de recherche Isidore 41
I.I Les enrichissements sémantiques sur Isidore 43
I.II Présentation de la section « Enrichissements ? » 45
I.III Les quatre langages documentaires exploités dans les enrichissements sémantiques francophones d'Isidore 47
Présentation de Rameau 47
Présentation de Gemet 48
Présentation de Pactols 49
Présentation de GeoEthno 49
II. L'analyse des enrichissements sémantiques 50
III. Les entretiens 55

Partie 3 : étude de cas : présentation des résultats de l'observation **58**

I. Créer du lien entre les données : rebonds et alignements 58
I.I Les alignements 58
I.II Les rebonds 62

II. Les doublons 64
II.I Les rebonds à partir de doublons 66
III. La compréhension de la sémantique 71
III.I La question de la pertinence 78

Conclusion **82**

Bibliographie **84**

Bibliographie et webographie de l'état de l'art 84
Bibliographie et webographie de l'étude de cas 90

Table des illustrations **93**

Table des annexes **94**