Des grandes classifications au Web de données et l'émergence de l'indexation sémantique : le cas du tagging sémantique dans le portail histoiredesarts.culture.fr

Domingos Ruiz Lepores

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

École Management et Société-Département CITS

INTD

MÉMOIRE pour obtenir le

Titre professionnel "Chef de projet en ingénierie documentaire" INTD

niveau I

Présenté et soutenu par

Domingos Ruiz Lepores

le 8 décembre 2011

Des grandes classifications au Web de données et l'émergence de l'indexation sémantique:

le cas du tagging sémantique dans le portail histoiredesarts.culture.fr

Jury
Adriana Lopez Uroz, directrice de mémoire et responsable du centre de documentation de l'INTD
Bertrand Sajus, responsable de stage et chef de projet utilisateur du portail histoiredesarts.culture.fr

Promotion XLI

À mon père

Remerciements

Je tiens à remercier Adriana Lopez Uroz pour m'avoir fait confiance depuis le départ de cette nouvelle étape de mon parcours professionnel.

Je remercie également l'équipe pédagogique de l'INTD : Brigitte Guyot, Nadia Raïs et Claire Scopsi et aussi l'équipe administrative, Anne Christine Pinchemel, Carole Briend et Geneviève Courtial, pour rendre moins ténébreuse la bureaucratie française.

Un grand et vif merci à Bertrand Sajus pour le partage de ses remarquables connaissances, pour la conception de cet innovateur projet de tagging sémantique, pour me faire redécouvrir le plaisir dans le travail et pour la dégustation quotidienne de crus de chocolat.

Merci à Alexandre Monnin pour ses conseils et indications bibliographiques.

Je remercie aussi les collègues de formation Assia Abed, Florence de Conninck Olivier, Danielle Gillot et Pascal Lafitte qui ont toujours dit : courage !

Merci encore à Luciano Brito et à Jean-François Alfaro pour les bonnes énergies.

Et finalement, je remercie Domingos Ruiz, dont la présence absente a été indispensable pour l'exécution de ce travail.

Notice

RUIZ LEPORES Domingos. Des grandes classifications au Web de données et l'émergence de l'indexation sémantique : le cas du tagging sémantique dans le portail Histoire des Arts. 2011. 118 pages. Mémoire, ingénierie documentaire, INTD-CNAM, 2011.

Ce mémoire a pour objectif présenter une nouvelle approche pour l'indexation documentaire à l'ère du Web de données. La première partie est consacrée à tracer l'évolution des langages documentaires vis-à-vis les nouvelles technologies du Web Sémantique. Il est également analysé comment les modalités d'indexation collaborative, telles que le tagging et les folksonomies ont contribué à faire émerger l'indexation sémantique. Deuxièmement, une étude de cas de tagging sémantique de ressources culturelles est réalisée. Enfin, plusieurs pistes et perspectives, liées à la recherche d'information, sont envisagées comme conséquence d'une nouvelle approche d'indexation sémantique qui profite de l'information structurée et interopérable, disponible grâce à l'application des technologies du Web Sémantique.

Folksonomie, indexation, interopérabilité, langage documentaire, métadonnée, patrimoine culturel numérique, recherche d'information, ressource pédagogique, structuration de données, web sémantique, web 3.0.

Table des matières

Introduction

« En effet le savoir lui-même est pouvoir ». Si l'on adapte la célèbre formulation du philosophe anglais Francis Bacon à nos jours numérisés, on pourrait affirmer que l'information c'est pouvoir.

Et comment fait-on pour accéder et exercer ce pouvoir ?

Atteindre l'information précise est le but premier des sciences de l'information. C'est pour cela que plusieurs méthodes d'organisation des systèmes d'information sont mises en place.

Les techniques documentaires sont au service de cet aboutissement, spécialement les langages documentaires, qui permettent l'intermédiation entre les besoins des utilisateurs et l'information elle-même.

À l'ère du numérique, où une quantité immensurable d'informations circule sur le Web, trouver l'information souhaitée plus aisément est devenu impératif dans un monde où la recherche d'information est une pratique banale et quotidienne, non plus réservée à des spécialistes.

C'est aux experts de l'information, néanmoins, de concevoir des nouvelles procédures, capables de restituer l'information aux utilisateurs dans les conditions les plus simples et rapides.

L'arrivée de l'informatique, contrairement à ce que l'on pourrait croire, a permis de placer les langages documentaires, y compris ses pratiques d'indexation, en évidence parmi l'ensemble de techniques documentaires.

C'est donc pour cela que réaliser une étude de l'évolution des langages documentaires, permet d'éclairer la façon selon laquelle les nouveaux systèmes d'organisation de connaissances comme le tagging collaboratif et les ontologies ont été constitués.

Le développement de nouvelles technologies a permis aussi l'avènement de systèmes de représentation du savoir inédits, ouvrant la voie au renforcement des langages documentaires, qui se mettent à jour constamment, dans son objectif premier de rendre l'information le plus exploitable possible.

Ainsi, comprendre les mécanismes de structuration des données sur le Web est une mission essentielle aujourd'hui, ce qui nous amène à être complètement d'accord avec

le chercheur Fabien Gandon pour qui « celui qui contrôle les métadonnées, contrôle le Web »[1].

[1] "He who controls metadata, controls the web."

Première partie
Historique des langages documentaires et émergence du Web sémantique

1 Les langages documentaires

Le langage documentaire est un « **langage contrôlé et normalisé** utilisé dans un système documentaire **pour l'indexation et la recherche**. Un langage documentaire permet de **représenter de manière univoque les notions identifiées dans les documents** et dans les demandes des utilisateurs, en prescrivant une liste de termes ou d'indices, et leurs règles d'utilisation. »[2]

Les langages documentaires peuvent êtres **combinatoires ou postcoordonnés**, dont les éléments peuvent alors être combinés entre eux *a posteriori* lors de l'indexation ou de la recherche, comme dans les thésaurus et dans les classifications à facettes, ou **précoordonnés**, contenant des combinaisons de notions établies *a priori,* comme dans les classifications hiérarchiques et les listes de vedettes-matières.

1.1 Les langages de type classificatoire

1.1.1 Classifications

« Les langages classificatoires (ou classifications documentaires, ou encore classifications bibliographiques) sont des instruments de travail liés aux nécessités de fonctionnement d'une bibliothèque ou d'un centre de documentation. Ils permettent de représenter de façon synthétique le sujet d'un document, et de regrouper les ouvrages sur les rayons par affinité de contenu » [16, Maniez, 1987, p. 22].

Les classifications bibliographiques[3] ont été les premiers outils d'organisation thématique des ouvrages.

Un domaine de la connaissance est découpé en sous-domaines, ce qui conduit d'une part, à élaborer un système d'organisation des connaissances qui témoigne, pour une époque donnée, de l'état d'avancement des connaissances et d'autre part, à proposer un outil permettant d'étiqueter et de classer les documents, en rapport avec un domaine de la connaissance.

Il existe différentes classifications encyclopédiques et par domaines d'activité ou disciplines utilisées à travers le monde dans diverses bibliothèques, services d'archives,

[2] http://www.adbs.fr/langage-documentaire-17593.htm?RH=OUTILS_VOC

[3] http://fr.wikipedia.org/wiki/Classification_(science_de_l%27information)

musées ou services documentaires. La mise à disposition de ressources sur internet fait évoluer fortement les outils d'accès de type classificatoire.

Selon le Vocabulaire de la doc de l'Association des professionnels de l'information et de la documentation – ADBS, la classification est le « langage documentaire permettant l'organisation d'un ou plusieurs domaines de la connaissance en un système ordonné de classes et sous-classes. Celles-ci ainsi que leurs relations peuvent être représentées par les indices d'une notation. Ces indices sont explicités par un libellé textuel. »4

Différents critères peuvent être choisis pour différencier les classifications: la notation décimale ou alpha-numérique ; la couverture des domaines de la connaissance traités (classifications encyclopédiques comme la classification Décimale universelle - CDU et la classification de Dewey) et le mode de division et de structuration (classification à facettes).

1.1.1.1 La classification décimale Dewey – CDD

La classification décimale Dewey - CDD[5] – a été créée par le bibliothécaire américain Melvil Dewey en **1876** au sein de la bibliothèque du Ahmerst College où il travaillait comme aide-bibliothécaire. En 1883, il devient bibliothécaire du Columbia College et fonde l'école de bibliothéconomie de Columbia, qui est la plus ancienne institution de formation de bibliothécaires au monde. Cette école, qui connaît un grand succès, déménage à Albany en 1890 et devient la New York State Library School. Dewey participe aussi à la fondation de la American Library Association – ALA.

La classification Dewey comporte dix **classes et chaque classe est elle-même subdivisée en dix autres classes et ainsi de suite**. Sa notation est donc très simple, ce qui en fait un outil facile à manipuler et à adapter :

000 - Informatique, information, ouvrages généraux

100 - Philosophie, Parapsychologie et Occultisme, Psychologie

200 - Religion

300 - Sciences sociales

---

[4] http://www.adbs.fr/classification-16541.htm?RH=OUTILS_VOC

[5] http://fr.wikipedia.org/wiki/Classification_d%C3%A9cimale_de_Dewey

400 - Langues

500 - Sciences de la nature et Mathématiques

600 - Technologie (Sciences appliquées)

700 - Arts. Beaux-arts et arts décoratifs. Loisirs et Sports

800 - Littérature (Belles-Lettres) et techniques d'écriture

900 - Géographie, Histoire et disciplines auxiliaires

Exemple de divisions successives:

600 - Techniques. (l'indice est 600, car il faut au moins trois chiffres)

640 - Vie domestique. (l'indice est 640, car il faut au moins trois chiffres)

641 - Alimentation

641.5 - Cuisine. (un point sépare le troisième et le quatrième chiffres)

641.57 - Cuisine pour les collectivités

C'est la classification la plus connue et la plus diffusée dans le monde : environ 135 pays l'ont adoptée. Il existe aujourd'hui une trentaine de traductions, ce qui en fait un système vivant et ouvert sur le monde. La CDD est depuis 1988 un nom déposé par Online Computer Library Center – OCLC[6] -, qui a acheté les droits à la Forest Press Foundation, créée par Melvil Dewey pour poursuivre son œuvre.

La classification Dewey est néanmoins objet de plusieurs critiques. Pour ses détracteurs, cette classification reflète toujours l'organisation générale du savoir telle qu'on la concevait aux États-Unis à la fin du XIXème siècle. Ils lui reprochent également sa macrostructure, vu qu'elle ne convient pas à des institutions dont leurs fonds est constitué d'une collection très spécialisée ou pointue dans un domaine donné. A titre d'exemple, la philosophie et la religion, qui représentaient environ 10% de la production éditoriale à l'époque de sa création, ont aujourd'hui encore une position disproportionnée dans l'ensemble de la classification.

[6] http://www.oclc.org. Organisation aussi responsable pour la production et mise à jour de la base de données bibliographiques en ligne WorldCat.org. Ce catalogue contient les données relatives à plus de 10.000 bibliothèques publiques et privées du monde.

Enfin, **les notions d'interdisciplinarité et pluridisciplinarité sont peu présentes**, voire inexistantes. Dans la classification Dewey, il n'est pas possible de se documenter sur un sujet sans savoir très précisément à quelle discipline le rattacher. Or cela pose problème lorsqu'un ouvrage traite précisément du lien entre deux disciplines. La bio-informatique, par exemple, sera-t-elle à chercher dans la section Sciences de la nature et Mathématiques ou Technologie (Sciences appliquées) ? C'est encore à examiner ces manques et incongruences pour que la classification Dewey devienne un outil complet et efficient.

1.1.1.2 La classification décimale universelle - CDU

La classification décimale universelle – CDU[7] - est un système de classification de bibliothèque développé par Paul Otlet et Henri La Fontaine, deux juristes belges fondateurs de l'Institut International de Bibliographie en **1895**, à partir de la classification décimale de Dewey – CDD -, et avec l'autorisation de Melvil Dewey. Elle a connu plusieurs éditions depuis 1905 et a été traduite en 40 langues.

Également à la CDD, elle **répartit les connaissances humaines en neuf catégories** notées de 1 à 9, le 0 étant réservé aux généralités « en général », tel comme se suit :

Classe 0 - Sciences et connaissance. Organisation. Informatique. Information. Documentation. Bibliothéconomie. Institutions. Publications

Classe 1 - Philosophie et psychologie

Classe 2 - Religion. Théologie

Classe 3 - Sciences sociales. Statistique. Économie. Commerce. Droit. Gouvernement. Affaires militaires. Assistance sociale. Assurances. Éducation. Folklore

Classe 4 - Inoccupée

Classe 5 - Mathématique. Sciences exactes et naturelles

Classe 6 - Sciences appliquées. Médecine. Technologie

Classe 7 - Arts. Divertissements. Sports

Classe 8 - Langue. Linguistique. Littérature

[7] http://fr.wikipedia.org/wiki/Classification_d%C3%A9cimale_universelle

Classe 9 - Géographie. Biographie. Histoire

Chaque catégorie est elle-même normalement divisée en dix parties (toutes les classes et sous-classes ne sont pas utilisées). Un zéro terminant un indice indique toujours qu'il s'agit de généralités, par exemple 750 signifie « généralités dans le domaine de la peinture ».

Il faut se rappeler que ces indices sont en fait des nombres décimaux, dont on a enlevé le 0, puisque ces deux signes seraient communs à tous les indices. Ainsi, 541 doit être classé avant 61 (un peu de la même façon qu'un sous-titre numéroté 5.4.1 vient avant un titre numéroté 6.1).

Destinée à l'origine à établir une bibliographie universelle exhaustive, elle permet la réalisation d'indices très complexes. Les indices ont par la suite eu tendance à se simplifier, l'expérience ayant prouvé que ce qui était à l'origine une volonté de précision ne facilitait pas la recherche documentaire et rendait la classification décimale universelle très difficile à maintenir.

La CDU est très peu connue en Amérique du Nord et d'une manière générale on peut dire qu'elle est exclusivement utilisée dans les pays européens non anglophones, comme la France, la Belgique et l'Espagne, malgré son **but original d'être multilingue**. En France, la CDU a été utilisée dans la plupart des bibliothèques universitaires, mais elle régresse depuis la fin des années 1980, au profit de la CDD. Elle reste encore en usage, dans une version généralement simplifiée, dans les centres de documentation et d'information des établissements scolaires du secondaire (essentiellement dans les lycées) et dans de nombreuses bibliothèques publiques ou privées.

La CDU a pour avantage d'avoir un système de notation assez simple et flexible, ce qui lui permet d'être employée tant dans des domaines très spécialisés comme dans des bibliothèques publiques ou universitaires et elle convient pour classer des collections d'objets ou de documents.

Cette classification n'est disponible complètement que sous forme informatisée – et sous licence - auprès de l'UDC Consortium[8], organisme responsable pour sa maintenance. Bien qu'elle soit dotée d'un niveau de précision plus élevé que la CDD et la classification de la librairie du Congrès – LCC – (voir ci-dessous), la CDU connaît depuis quelques années une certaine régression par rapport à son utilisation, due à l'irrégularité de ses mises à jour. Par ailleurs, elle ne bénéficie pas de caution particulière, comme le fait d'être l'outil de référence

[8] http://www.udcc.org/

d'une grande bibliothèque, ce qui lui permettrait d'asseoir sa légitimité. Les moyens financiers investis pour cette classification ne sont pas non plus suffisants pour envisager des études lui assurant une certaine pérennité.

1.1.1.3 La classification de la Bibliothèque du Congrès - LCC

La classification de la Bibliothèque du Congrès – LCC[9] (en anglais Library of Congress Classification) – a été conçue **entre 1890 et 1901**, ayant pour objectif organiser le gros volume de documents détenus par la Bibliothèque du Congrès des États-Unis. Elle est aussi utilisée par plusieurs bibliothèques universitaires aux États-Unis et dans le monde.

Elle est répartie sur 21 classes représentées chacune par une lettre de l'alphabet[10] :

A : Généralités

B : Philosophie, psychologie, religion

C : Sciences auxiliaires de l'histoire

D : Histoire générale et histoire de l'Europe

E : Histoire de l'Amérique (généralités et États-Unis)

F : Histoire de l'Amérique (autres pays d'Amérique)

G : Géographie, anthropologie, loisirs

H : Sciences sociales

J : Sciences politiques

K : Droit

L : Éducation

M : Musique et musicologie

N : Beaux arts

---

[9] http://www.loc.gov/catdir/cpso/lcco/

[10] Les lettres I, O, W, X et Y ne sont pas utilisées.

P : Langage et littérature

Q : Sciences

R : Médecine

S : Agriculture

T : Technologie

U : Science militaire

V : Science navale

Z : Bibliographie, Sciences de l'information et des bibliothèques

Chacune de ses catégories peut être précisée par une ou deux lettres supplémentaires, puis par une série de chiffres, comme dans l'exemple suivant :

C : Sciences auxiliaires de l'histoire

CJ : Numismatique

CJ 1-4625 : Pièces de monnaie

La LCC est donc une classification alphanumérique vu que pour les subdivisions, les lettres sont accompagnées de chiffres, permettant de construire un système bien précis et plutôt convivial. Elle remporte un certain succès aux États-Unis, mais est très peu utilisée dehors ce pays, ce qui peut s'expliquer par le fait que la classification **ne s'appuie sur aucune source comme une encyclopédie ou autre outil de connaissance fiable** contrairement à la Dewey, qui avait bénéficié de toute la richesse culturelle de son auteur. On lui reproche un certain américanisme car elle reflète avant tout une vision de la connaissance tel comme elle se présente aux États-Unis : l'existence d'une lettre dédiée à d'autres pays de l'Amérique, exclu les États-Unis, montre clairement son axe d'orientation.

Cette classification ne possède pas de contrôle terminologique et syntaxique. On lui reproche aussi certaines parties qui sont considérées comme obsolètes et qui mériteraient une révision totale; l'absence de notation hiérarchique et par conséquent l'organisation du savoir n'en est que partielle; l'index, qui aurait été un outil de navigation supplémentaire, est totalement inexistant. Son évolution s'avère très lente et le corpus des classes reste stable, même s'il est possible de rajouter des termes à travers de chiffres.

Au début des années 1980, des travaux d'automatisation avaient été entrepris mais l'arrivée de l'internet a obligé l'équipe de maintenance à tout revoir et à s'interroger sur les processus de compatibilité et d'interopérabilité, à commencer par celle entre la classification elle-même et le répertoire de vedettes-matière Library of Congress Subject Headings – LCSH[11] - qui provient de la même institution. Pour le moment, la classification de a Librairie du Congrès n'est disponible qu'en sa version anglaise.

1.1.1.4 La Classification Bibliothéco-Bibliographique - BBK

La Classification Bibliothéco-Bibliographique (Bibliotečno-Bibliografičeskaja Klassifikacija - BBK)[12], est un système de classification de bibliothèque créé lors de l'existence de l'Union Soviétique et élaborée à partir de la CDU avec l'objectif de la remplacer.

Cette classification essaie d'intégrer le marxisme-léninisme dans le système de classement des documents. Les sciences dites « objectives » sont donc favorisées dans le classement, au détriment des sciences humaines. Son système est alpha-décimal. En effet, à chacune des 28 classes est affectée une lettre de l'alphabet cyrillique, et cette lettre est suivie de nombres exprimés en fractions décimales inspirées de la CDU.

Son structure actuelle est la suivante[13] :

1. Généralités scientifiques et interdisciplinaires

2. Sciences naturelles (physique et mathématiques, chimie, sciences de la terre, biologie)

3. Technique, sciences techniques

4. Gestion agricole et forestière, sciences agricoles et forestières

5. Protection de la santé, sciences médicales

6. Sciences sociales

7. Culture, science, éducation

---

[11] http://id.loc.gov/authorities/subjects.html

[12] http://fr.wikipedia.org/wiki/BBK_(classification)

[13] http://www.indiana.edu/~libslav/slavcatman/bbkover.html

8. Lettres, théologie, philosophie, psychologie

9. Littérature au contenu universel (bibliographies)

La BBK est encore en vigueur dans de nombreuses bibliothèques de l'ex-bloc soviétique, dans une version parfois adaptée.

1.1.2 Taxonomies ou taxinomies

La taxinomie[14] (du grec *taxis* « placement », « classement », « mise en ordre » et *nomos* « loi », « règle ») ne s'intéressait à l'origine qu'à la classification des organismes vivants en bactériologie, en botanique et en zoologie, mais son utilisation s'étend aujourd'hui à d'autres sciences, telles les sciences humaines et les sciences de l'information.

La taxinomie est la science qui a pour finalité de décrire des objets et de les regrouper en entités appelées taxons dans l'intention de les identifier puis les nommer, et enfin les classer. Elle complète la systématique qui est la science qui organise le classement des taxons et leurs relations.

Par extension, le mot **taxinomie** est utilisé pour **désigner des systèmes ou des méthodes de classification hiérarchiques permettant d'inventorier des objets, des concepts, des informations d'un domaine donné** en vue de : décider du comportement à adopter face à un objet donné, prédire le comportement d'un objet et comprendre un phénomène dans l'objectif de pouvoir ensuite agir.

La taxonomie est régie par des relations de subsomption, c'est-à-dire qu'un élément fait forcément partie d'un autre. On peut les représenter sous forme d'un arbre où chaque nœud est un taxon :

[14] http://wiki.univ-paris5.fr/wiki/Taxinomie

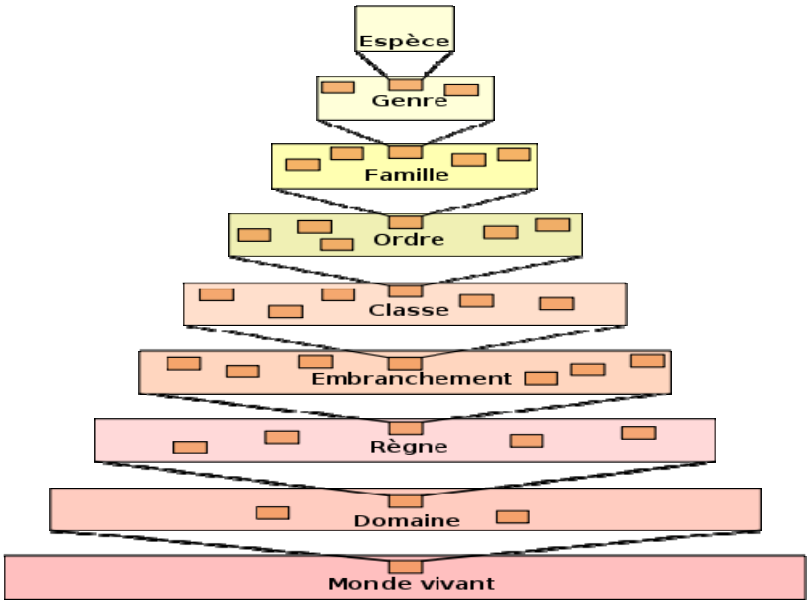

Figure 1 : Hiérarchie taxonomique du vivant[15]

L'indexation des ressources documentaires a recours à des taxonomies, entre autres[16], pour les répertorier dans le domaine concret dont elles relèvent.

1.1.2.1 Taxonomie ou Taxinomie ?

Le terme taxonomie a été créé en 1813, sous l'orthographe de « taxonomie », par le botaniste suisse Augustin Pyrame de Candolle (1778-1841) dans son ouvrage « *Théorie élémentaire de la botanique ou exposition des principes de la classification naturelle et de l'art de décrire et d'étudier les végétaux* », pour désigner dans sa « théorie des classifications » à la fois la méthode et ce qu'il a qualifié de « bases de la botanique philosophique ».

Émile Littré, dans son *Dictionnaire de la langue française* (version 1872-1877) précisait que le mot « taxinomie » pouvait aussi être utilisé, formé sur l'étymon grec taxis (l'ordre). Le Grand dictionnaire terminologique confirme que taxinomie est recommandé par plusieurs auteurs considérant « taxonomie » comme « un calque de l'anglais *taxonomy* ».

Le terme taxinomie ne dérive pas du mot taxon, ce dernier étant un concept apparu bien plus récemment : le botaniste Herman Lam a créé le mot taxon en 1948. Ainsi,

---

[15] Source : http://fr.wikipedia.org/wiki/Fichier:Taxonomic_hierarchy.svg

[16] Classifications et ontologies, par exemple.

la taxinomie n'est pas, étymologiquement, l'étude des taxons mais bien les lois sur l'ordre, donc les règles de la classification. Cependant, le mot taxinomie est le plus souvent utilisé pour nommer la science de la description des taxons.

Le terme est devenu d'usage courant aujourd'hui, soit dans la graphie originale, mais étymologiquement contestée[17], de taxonomie, soit sous la graphie corrigée par Émile Littré de taxinomie, mais l'autre graphie reste néanmoins très répandue, notamment en raison de sa conservation dans la traduction anglaise, ***taxonomy***.

1.1.2.2 Élargissement du concept de taxonomie

Le concept de taxonomie dépend de son contexte d'utilisation. D'un point de vue documentaire, on lui reconnaît déjà plusieurs sens mais elle peut tout à fait être envisagée dans d'autres disciplines. En ce qui concerne les Sciences de l'information, Alan Gilchrist [10, Gilchrit] propose **cinq acceptions** :

Taxonomie **comme répertoire de site Web** : Elle est souvent utilisée sur internet et en augmentation constante sur les intranets. Le répertoire **Dmoz**[18] est un très bon exemple de cet usage fait des taxonomies. Il n'y a pas de niveau hiérarchique et les termes peuvent être répétés indéfiniment d'une classe à l'autre, ce qui offre à l'utilisateur plusieurs voies de navigation.

Taxonomie **pour appuyer l'indexation automatique** : Elle peut être cachée derrière une classification supportant des algorithmes comprenant eux-mêmes des ensembles de mots, des syntagmes, des synonymes ou encore des instructions. Cette base de règles est utilisée pour extraire automatiquement les termes d'indexation qui peuvent ou non être présents dans les documents. Cette approche est intéressante dans la mesure où la prolifération des informations rend impossible une indexation 100% manuelle.

Taxonomie **créée par catégorisation automatique** : Cette définition rejoint la première dans la mesure où il est question de générer automatiquement des catégories pour classer de grandes quantités de documents et la présentation ressemble à celle du répertoire Dmoz. Les logiciels utilisés pour aboutir à ce résultat sont globalement les mêmes, bien que dans ce cas présent, l'intervention humaine permet d'augmenter le nombre de résultats.

---

[17] Sur la controverse étymologique, voir: FISCHER, Jean-Louis et REY, Roselyne. De l'origine et de l'usage des termes taxinomie-taxonomie. Documents pour l'histoire du vocabulaire scientifique, 1983, vol. V, p. 97-113. Et aussi : http://www.btb.termiumplus.gc.ca/chroniq-srch?lang=eng&srchtxt=taxonomie&i=&i=1&lettr=&cur=1&nmbr=&comencsrch.x=0&comencsrch.y=0 [Consulté le 22 sepetembre 2011].

Taxonomie **comme filtre de recherche** : Dans cette acception, une taxonomie est créée, importée ou utilisée dans une formule de requête. Les homographes peuvent être débarrassés de leur ambiguïté et les synonymes regroupés entre eux. L'utilisateur peut naviguer dans les hiérarchies et passer d'un terme à l'autre. Cette taxonomie est finalement un thésaurus qui a été adapté aux besoins des utilisateurs et leur permet une navigation fonctionnelle.

Taxonomie **dans l'entreprise** : La taxonomie est une des solutions envisageables pour les entreprises de permettre l'accès de leurs informations à leur personnel. Dans un contexte professionnel où les informations proviennent de sources variées, externes ou internes, le besoin d'ordonner pour en faciliter l'accès est fondamental. Ces taxonomies peuvent être perçues comme un système hybride de thésaurus et d'ontologies, un nouveau type de taxonomies dont l'appellation anglophone exacte est « corporate taxonomies ». Ce nouveau procédé pourrait être adapté au fonctionnement des moteurs de recherche et servir de modèle pour l'indexation.

Il est important affirmer que la taxonomie utilise à la fois les techniques des classifications et des thésaurus mais ces trois notions ne doivent pas être confondues. Il existe des similarités entre chaque technique dans la manière de procéder à l'indexation, mais les degrés de granularité font la différence.

Pour Jean Delahousse[19], les taxonomies peuvent être des organisations très formelles de classes d'objets ou d'êtres vivants avec des applications les sciences de la vie, la biologie. Dans le domaine de la gestion du contenu, le mot de taxonomie est utilisé pour désigner un **plan de classement ad hoc**.

Ce type de taxonomie ou plan de classement composé d'une seule hiérarchie repose sur un modèle de plan de classement indiquant le contenu des différentes branches et niveaux et les référentiels utilisés pour chaque niveau. Les logiciels de gestion de terminologie permettent la génération et contrôle de ces taxonomies en fournissant les différents référentiels utilisés.

Et encore : Le terme de taxonomies peut également désigner un ou plusieurs listes hiérarchiques de sujets permettant de filtrer le contenu dans un site web ou un intranet. On parlera alors de **taxonomies de navigation**.

---

[18] Open Directory Project

Les taxonomies de navigation sont souvent créées à partir de la terminologie métier de l'entreprise (parfois adaptée à la cible clientèle) et de listes de référence (comme noms de marques, noms de produits). Les taxonomies métier ont comme objectifs d'offrir un système simple et clair de filtrage des contenus sur un site web en utilisant plusieurs axes de filtrage. Ces taxonomies font ainsi parties des ressources terminologiques de l'entreprise.

D'après le Vocabulaire de la Doc de l'ADBS[20], d'un point de vue structurel, les taxonomies (de termes, de classes, de concepts) désignent la hiérarchie ou l'arborescence autour de laquelle sont construits différents types d'instruments, comme les thésaurus, les réseaux sémantiques ou les ontologies ; d'un point de vue fonctionnel, une taxonomie est un cadre d'organisation pour des ressources numériques de toute nature (et pas seulement documentaires), destiné à en permettre une présentation ordonnée et y donnant accès par navigation hypertextuelle.

Synthétiquement, pour l'éditeur de logiciels sémantiques Lingway[21], la taxonomie est le réseau sémantique dans lequel la seule relation est la relation hiérarchique (générique-spécifique)[22], ce qui permet de la différencier des thésaurus, dont les possibilités des relations sont plus riches.

1.1.3 Ontologies

1.1.3.1 Ontologie : mot polysémique

Le terme « ontologie[23] » couvre au moins trois champs de la science. En philosophie, l'ontologie est la branche de la métaphysique concernant l'étude de l'être ; en médecine, l'ontologie s'intéresse à la genèse des maladies ; **en informatique, une ontologie est un système de représentation des connaissances.**

---

[19] http://mondeca.wordpress.com/2007/10/10/gestion-d%E2%80%99une-terminologie-d%E2%80%99entreprise-utilisations-enjeux-et-differentes-formes-d%E2%80%99organisation/#more-73

[20] http://www.adbs.fr/taxonomie-58346.htm?RH=OUTILS_VOC

[21] http://www.lingway.com/index.php?option=com_content&task=view&id=37&Itemid=83

[22] http://www.lingway.com/content/view/22/62/lang,fr/

[23] http://fr.wikipedia.org/wiki/Ontologie

Ce n'est pas vraiment évident de comprendre l'emploi du terme ontologie dans le champs des sciences de l'information et de l'informatique, possiblement due à son appropriation de la philosophie.

En philosophie, l'ontologie (du grec *onto*, « étant », participe présent du verbe « être ») est l'étude de l'être en tant qu'être, c'est-à-dire l'étude des propriétés générales de ce qui existe. Par analogie, le terme a été repris en informatique et en sciences de l'information, où une ontologie est **l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations**, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances.

Le terme renvoie alors à la « théorie de l'existence », c'est-à-dire la théorie qui tente d'expliquer les concepts qui existent dans le monde et comment ces concepts s'imbriquent et s'organisent pour **donner du sens**.

1.1.3.2 L'ontologie des connaissances

L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts. Elle est employée pour raisonner à propos des objets du domaine concerné.

Dans une phrase: "**L'ontologie est aux données ce que la grammaire est au langage**".[24]

Les concepts sont organisés dans un graphe dont les relations peuvent être :

- des **relations sémantiques**,

- des relations de subsomption (inclusion).

L'objectif premier d'une ontologie est de **modéliser** un ensemble de connaissances dans un domaine donné, qui peut être réel ou imaginaire :

[24] http://fr.wikipedia.org/wiki/Ontologie_(informatique)

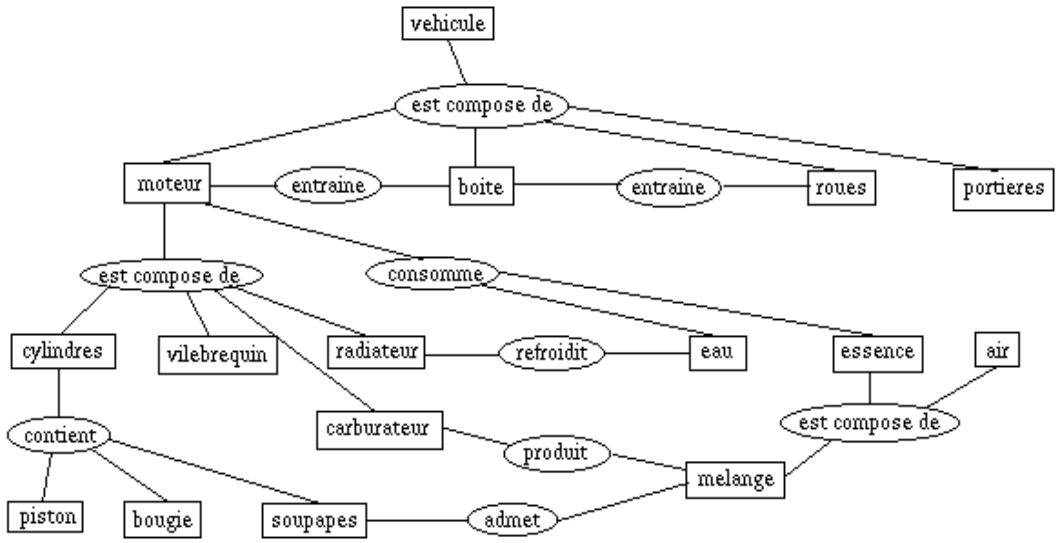

Figure 2 : Exemple d'ontologie pour un véhicule comportant des relations complexes entre termes.[25]

Contrairement à l'être humain, **la connaissance pour un système informatique se limite à la connaissance qu'il peut représenter**. Chez l'être humain, les connaissances représentables (c'est-à-dire l'univers du discours) sont complétées par des connaissances non exprimables (sensations, perceptions, sentiments non verbalisables, connaissances inconscientes, connaissances tacites, etc.). Ces éléments non représentables participent néanmoins aux processus de raisonnement et de décision, qui sont des processus cognitifs en gestion des connaissances. Les performances cognitives d'un agent informatique vont donc en partie reposer sur le champ des représentations auquel il aura accès, c'est-à-dire concrètement au champ des représentations qui aura été formalisé.

Les ontologies informatiques sont des outils qui permettent précisément de représenter un corpus de connaissances sous une forme utilisable par un ordinateur.

Les ontologies sont employées dans l'intelligence artificielle, le **Web sémantique**, le génie logiciel, l'informatique biomédicale et l'architecture de l'information comme une forme de représentation de la connaissance au sujet d'un monde ou d'une certaine partie de ce monde.

[25]Source :http://liris.cnrs.fr/alain.mille/enseignements/IGC_M2_2008/session8/rapc2/Rapc_Session2_Cas_base_de_cas.html

L'appropriation du concept informatique d'ontologie par les langages documentaires s'explique par le fait que les ontologies servent à **normaliser un corpus de termes** et à **expliciter les liens sémantiques qui les attachent**.

Les ontologies décrivent généralement :

- individus : les objets de base,

- classes : ensembles, collections, ou types d'objets,

- attributs : propriétés, fonctionnalités, caractéristiques ou paramètres que les objets peuvent posséder et partager,

- relations : les liens que les objets peuvent avoir entre eux,

- événements : changements subis par des attributs ou des relations.

Les ontologies peuvent aussi être distinguées selon plusieurs niveaux, selon le domaine modélisé et éventuellement les tâches pour lesquelles elles sont conçues[26] :

- les ontologies d'application ont un domaine de validité restreint et correspondent l'exécution d'une tâche.

- les ontologies de domaine ont un faisceau plus large, une bonne précision et ne sont pas propres à une tâche particulière.

- les ontologies générales ne sont pas propres à un domaine. Leur précision est moyenne.

- les ontologies supérieures représentent des concepts généraux comme l'espace, le temps ou la matière. Elles sont universelles et les concepts des trois autres types d'ontologie peuvent y faire référence.

Pour Gruber [11, Gruber, p. 13], « une ontologie est la spécification d'une conceptualisation d'un domaine de connaissance », *a view of a world from a particular perspective*. Ainsi, une ontologie est la conceptualisation d'un domaine, c'est-à-dire un choix quant à la manière de décrire un domaine. C'est la spécification de cette conceptualisation, sa description formelle. Elle est aussi une base de formalisation des connaissances. Elle se situe à un certain niveau d'abstraction et dans un contexte particulier. Elle est encore une représentation d'une conceptualisation partagée et consensuelle, dans un domaine

[26] http://websemantique.org/Ontologie

particulier et vers un objectif commun; elle classifie en catégories les relations entre les concepts.

Selon Jacques Chaumier [4, Chaumier], « une ontologie fournit le vocabulaire spécifique à un domaine de la connaissance et, selon un degré de formalisation variable, fixe le sens des concepts et des relations qui les unissent ». Il complète cette première définition en stipulant que les composants d'une ontologie comprend : une ou plusieurs taxonomies, ordonnées en classes et sous-classes composées d'instances représentant les individus ou objets ; les types d'attributs ou propriétés qui peuvent être attachés à ces objets ; les types de relations entre les concepts d'une taxinomie; des axiomes ou des règles d'inférence permettant de définir les propriétés de ces relations.

Il s'agit, donc, de normaliser les termes d'un fonds lexical d'un domaine/métier donné, d'analyser d'un point de vue linguistique ce fonds, de l'uniformiser sémantiquement et d'établir les relations entre les mots ou les syntagmes.

Cela dit par Jean Delahousse[27] :

« L'ontologie a pour objectif dans un domaine métier donné, de disposer d'un **ensemble de concepts** métiers **non ambigus** et de leur organisation par des **relations hiérarchiques ou** des relations **sémantiques**. Une ontologie métier s'inscrit dans une modélisation qui permet de définir les classes de concepts (termes, personnes, projets, molécules, hébergements…), leurs attributs descriptifs (nom, acronyme, définition, taille, âge, date de début, localisation…) ainsi que les types de relations sémantiques pouvant les relier (est un sous concept de, travaille avec, est une filiale de, a une interaction médicamenteuse avec…). Une ontologie bien réalisée permet d'effectuer des inférences et du raisonnement automatisé (ex: trouver les sociétés produisant des molécules ayant une interaction médicamenteuse avec un des produits de la société X). »

Il ajoute que « en raison de ses capacités de modélisation de différents organisation de termes, les ontologies offrent une grande souplesse pour gérer : des ressources terminologiques organisées sous forme de listes de référence, de thésaurus… ; des taxonomies de classification ou de navigation; des représentations des connaissances en reliant les concepts métier par des relations sémantiques ; des **relations d'équivalence ou**

---

[27] http://mondeca.wordpress.com/2007/10/10/gestion-d%E2%80%99une-terminologie-d%E2%80%99entreprise-utilisations-enjeux-et-differentes-formes-d%E2%80%99organisation/#more-73

de correspondance entre des référentiels devant être alignés pour des besoins d'interopérabilité. »

En bref, pour les sciences de l'information, **une ontologie est un réseau sémantique qui regroupe un ensemble de concepts décrivant complètement un domaine. Ces concepts sont liés les uns aux autres par des relations taxinomiques (hiérarchisation des concepts) d'une part, et sémantiques d'autre part.**

En Europe, la norme qui fait actuellement l'objet d'une attention particulière est une norme permettant notamment de décrire les ontologies sur le patrimoine culturel immatériel (bibliothèques, musées et archives…). C'est la norme ISO 2112732[28] : « ontologies nécessaires à la description des données concernant le patrimoine culturel », élaborée à partir du travail de normalisation effectué dans le cadre de la définition du patrimoine culturel immatériel effectuée par l'UNESCO. Cette norme a été publiée en 2006 et décrit en particulier les métadonnées nécessaires à la structuration des ontologies.

Parmi les outils qui ont été conçus pour l'élaboration et déploiement des ontologies, Protégé se met en évidence. Il est le plus connu et le plus utilisé des éditeurs d'ontologie. Open-source, développé par l'université Stanford, il a évolué depuis ses premières versions (2000) pour intégrer à partir de 2003 les standards du Web sémantique et notamment le langage de représentation d'ontologies Web Ontology Language (OWL), qui fournit les moyens pour définir des ontologies web structurées, avec des interfaces graphiques.

1.2 Les langages de type combinatoire

1.2.1 Le répertoire de vedettes-matière

1.2.1.1 Définition

Un répertoire de vedettes-matière[29] – RDV - est une liste encyclopédique dont les termes sont liés les uns aux autres par une syntaxe particulière. Il est avant tout conçu pour les bibliothèques dans un objectif de catalogage de leur fonds. Il a été créé et il est met à jour par la Section du Répertoire de vedettes-matière de la Bibliothèque de l'Université Laval au Québec. Il contient plus de 270 000 notices d'autorité-matière.

[28] http://www.iso.org/iso/catalogue_detail.htm?csnumber=34424

[29] http://fr.wikipedia.org/wiki/R%C3%A9pertoire_de_vedettes-mati%C3%A8re

Plus précisément, selon l'AFNOR[30], le répertoire de vedettes-matière[31] est un « ensemble d'un ou plusieurs descripteurs [termes] exprimant et précisant le **sujet** d'un document. Chaque vedette-matière correspond à un seul sujet, simple ou complexe. Un même document peut avoir plusieurs sujets donnant lieu à la rédaction de plusieurs vedettes-matières. »[32]

Les répertoires de vedettes-matières les plus connus sont celui de la Bibliothèque de l'Université de Laval[33], le Library of Congress Subject Headings[34] - LCSH - de la Bibliothèque du Congrès et le Répertoire d'autorité matière encyclopédique et alphabétique unifié - RAMEAU[35] - de la Bibliothèque Nationale de France.

Outil d'indexation et de recherche par sujet, le RVM est utilisé par des bibliothèques francophones du Québec et du Canada ainsi que par des bibliothèques d'organismes internationaux, de gouvernements et d'universités ailleurs dans le monde. Son équivalent en France est RAMEAU, élaboré en 1980 à partir du RVM Laval.

Le RVM Laval est une adaptation partielle du LCSH et complète des Canadian Subject Headings – CSH - de Bibliothèque et Archives Canada. Il contient aussi des vedettes-matière dites originales, c'est-à-dire sans équivalent anglais, qui rendent compte de réalités ignorées par les listes anglaises. Le RVM Laval est une norme nationale canadienne en indexation en français depuis 1974.

1.2.1.2 Le RVM de la Bibliothèque de l'Université de Laval

Le RVM de la Bibliothèque de l'Université de Laval a été édité pour la première fois en 1962, après 18 ans d'élaboration. Il s'agit en gros d'une adaptation francophone du LCSH qui est adopté par d'autres bibliothèques que celle de l'Université, à l'instar de la Bibliothèque et Archives Canada[36] et de la Bibliothèque nationale du Luxembourg[37].

---

[30] Association française de Normalisation

[31] AFNOR, 1996b, NF Z 44-070, p. 441

[32] http://www.ebsi.umontreal.ca/termino/00000246.htm

[33] https://rvmweb.bibl.ulaval.ca/

[34] http://id.loc.gov/authorities/subjects.html

[35] http://rameau.bnf.fr/

[36] http://www.collectionscanada.gc.ca/index-f.html

Le succès du RVM Laval s'explique en partie par la grande rigueur et la cohérence mobilisée pour créer et mettre à jour cet outil. Une équipe de bibliothécaires assure son développement et travaille indifféremment sur chaque vedette, avec mises à jour hebdomadaires.

En raison des coûts documentaires importants mais aussi dans un souci de partage des ressources, l'équipe a traduit les vedettes du LCSH. Ensuite, elle a considéré plus intéressant de les adapter aux besoins de la Bibliothèque et de la culture canadienne que de les traduire tout simplement. Ainsi, environ quatre mille vedettes n'ont pas d'équivalents en anglais, soit 20% du RVM, afin de répondre aux besoins d'indexation spécifique de la documentation en français. Le Canada reproche au LCSH un certain américano-centrisme, déjà visible dans la Classification de la Bibliothèque du Congrès – LCC.

L'équipe a mis en place un système de traduction automatique de l'anglais au français, ce qui rend le travail moins fastidieux, vu que 75% des vedettes peuvent êtres traduites automatiquement.

Le RVM Laval se présente comme un outil en constant évolution, surtout car les abonnés peuvent suggérer de nouvelles entrées de vedettes-matière et les mises à jour sont hebdomadaires, véritablement exploitable par ses utilisateurs.

1.2.1.3 Le Répertoire d'autorité matière encyclopédique et alphabétique unifié – RAMEAU

Le Répertoire d'autorité matière encyclopédique et alphabétique unifié – RAMEAU[38] est élaboré depuis 1980, de façon autonome, en relation avec le RVM Laval, qui s'appuie dans la liste de vedettes-matière de la Bibliothèque du Congrès – LCSH.

RAMEAU est le répertoire de vedettes-matière utilisé, en France, par la Bibliothèque nationale de France, les bibliothèques universitaires, de nombreuses bibliothèques de lecture publique ou de recherche ainsi que plusieurs organismes privés.

Le langage d'indexation précoordonné RAMEAU se compose d'un vocabulaire de termes reliés entre eux et d'une syntaxe indiquant les règles de construction des vedettes-matière à l'indexation. L'ensemble des notices d'autorité - dont le noyau est formé de noms communs et de noms géographiques - constitue les autorités RAMEAU. Elles sont

---

[37] http://www.bnl.public.lu/fr/bibliotheque/outils-bibliotheconomiques/catalogage-et-indexation/index.html

[38] http://rameau.bnf.fr/informations/rameauenbref.htm

complétées par un Guide d'indexation RAMEAU qui en assure le bon usage. Le Journal des créations et des modifications informe par ailleurs les utilisateurs, deux fois par an, des enrichissements et des évolutions du langage d'indexation.

A la différence d'un thésaurus, l'outil RAMEAU n'est pas constitué a priori mais enrichi au fur et à mesure des besoins de l'indexation, à partir des propositions formulées par le réseau de ses utilisateurs grâce à un Fichier national des propositions RAMEAU – FNPR.

RAMEAU est structuré sur trois niveaux : terminologique, sémantique et syntaxique. Il y a les termes retenus, nommés vedettes-matières, et les termes exclus. Cet ensemble constitue une « grappe terminologique »[39] qui fait la richesse du langage RAMEAU.

Néanmoins, sa précoordination assez sophistiquée, instituée pour exprimer des notions assez complexes, a posé quelques problèmes pour l'indexation et a conduit le centre national RAMEAU à réviser sa politique syntaxique afin de la simplifier car le volume croissant de documents obligerait à revoir l'indexation de milliers de documents.

L'actuel défi pour RAMEAU, en profitant de son langage contrôlé et hiérarchisé, est de modéliser des ontologies entre domaines et sous-domaines, de manière à utiliser de relations sémantiques qui existent déjà entre les termes, au sein d'une liste d'autorités qui resterait commune, dont le caractère homogène serait préservé, et qui finirait par constituer elle-même une manière d'ontologie encyclopédique, en raison du réseau des liens sémantiques établi entre les vedettes.

1.2.2 La classification à facettes

Une classification à facettes[40] est un système de classification de bibliothèque analytico-synthétique, et aussi pragmatique, dont les critères de classification ne sont pas homogènes.

Elle a été élaborée en 1924 par le mathématicien et bibliothécaire indien Shiyali Ramamrita Ranganathan[41] pour résoudre le problème posé par l'écrivain argentin Jorge Luis

---

[39] http://bbf.enssib.fr/consulter/bbf-2005-05-0038-001

[40] http://fr.wikipedia.org/wiki/Classification_%C3%A0_facettes

[41] http://unesdoc.unesco.org/images/0013/001333/133325fb.pdf

Borges : « Comment ranger les livres dans une bibliothèque quand on sait qu'il y en a des grands et des petits, des livres d'histoire et des romans, des auteurs qui ont écrit les deux et des collections reliées qui traitent de tout et que l'on doit y ajouter les dossiers correspondant aux différents sujets ? »

Sont célèbres les cinq lois de Ranganathan[42] :

1. Les livres sont faits pour être utilisés

2. À chaque lecteur son livre

3. À chaque livre son lecteur

4. Épargnons le temps du lecteur

5. Une bibliothèque est un organisme en développement

Le bibliothécaire indien a finalisé la première classification à facettes, nommée la Classification Colon - CC, en 1933. Composée de 42 classes, la notation utilise des lettres, des chiffres, des caractères grecs et différents signes de ponctuation, comme le « colon », terme en français qui désigne le signe « : », utilisé pour concaténer les différents indices servant à indexer le document.

L'objectif était de trouver une alternative aux classifications classiques hiérarchisées, comme la CDD ou la LCC. Dans le cas de la CC, les documents sont représentés en quelque sorte par une combinaison de concepts. Le bibliothécaire a conçu un système de déclinaison pour les documents en cinq facettes, appelées PMEST :

1. Personnalité : le concept principal du document

2. Énergie : l'opération ou action subie par l'objet

3. Matière : une substance ou une propriété

4. Espace : la localisation géographique

5. Temps : la localisation chronologique et temporelle.

Les grands concepts, comme l'Informatique ou la Médecine, se déclinent en listes prédéterminées d'objets, d'actions, et de propriétés. L'espace et le temps sont communs à

[42] http://www.libraryjournal.com/article/CA512179.html

tous ces concepts. D'autres concepts peuvent être ajoutés, ce qui étend les possibilités de cette classification. L'obligation d'enchaîner les propriétés toujours dans le même ordre permet d'aboutir à une notation homogène afin de classer les livres.

D'après l'étude de Michèle Hudon [12, Hudon, p. 94], cette classification comporte de nombreux avantages : Elle établit des liens fonctionnels entre les sujets et cette vision pluridimensionnelle semble apporter plus de cohérence que les schémas classiques qui reposent sur la hiérarchie et la synonymie. C'est cette absence de hiérarchie qui permet de visualiser les concepts indépendants uns des autres.

Du à non homogénéité des ses critères, tous les sujets peuvent être envisagés dans ce système, Grâce à la concaténation des indices, n'importe quel sujet peut être exprimé de manière simple et explicite. Par ailleurs, ce langage, à la base en mode post-coordonné, peut aussi bien fonctionner en pré-coordonné. La flexibilité pour ordonnancer les termes, les facettes ou les indices de classement en font, donc, un outil très complet pour représenter les sujets.

La CC possède l'atout d'être universelle.

Cependant, la CC intéresse assez peu les professionnels de la bibliothéconomie. Le système de notation mixte et les procédures d'application les découragent, même s'ils jugent que c'est un outil plutôt stimulant sur le plan intellectuel. Selon Jacques Maniez [17, Maniez, p. 253], « la formule PMEST est le point faible de ce système et le statut hybride qu'il réserve aux facettes a contribué dès le départ à brouiller la perception de ce concept en documentation ».

Même dans son pays de création, la CC est très peu répandue. La septième et dernière révision date de 1989[43]. Elle n'est quasiment pas remise à jour et de manière très sporadique. En conséquence, elle ne bénéficie pas de soutien institutionnel qui lui permettrait de vivre, ou plutôt de survivre.

Les classifications à facettes ont était l'objet de plusieurs études, notamment celui de Brian Campbell Vickery et Paule Salvan[44].

---

[43] http://www.amazon.com/Colon-classification-7th-practical-introduction/dp/817000103X

[44] Campbell, Vickery ;Salvan, Paule. La classification à facettes : guide pour la construction et l'utilisation de schémas spéciaux. Paris, Gauthier-Villars, 1963.

Jacques Maniez [17, Maniez] fait le rapprochement entre les facettes et les thésaurus et Jean Aitchison est connue pour avoir créé le néologisme « thesaurofacet[45] » pour nommer l'union de ces deux langages d'indexation.

1.2.3 Le thésaurus

1.2.3.1 Origine et définition

Un thésaurus[46] est une **liste organisée de termes représentant les concepts d'un domaine de la connaissance.**

C'est un **langage contrôlé utilisé pour l'indexation et la recherche de ressources documentaires** dans des applications informatiques spécialisées, d'où le nom de 'langage documentaire'. Les termes sont reliés entre eux par des relations de synonymie (terme équivalent), de hiérarchie (terme générique et terme spécifique) et d'association (terme associé); chaque terme appartient à une catégorie ou domaine.

Le thésaurus est un outil linguistique qui permet de **mettre en relation le langage naturel des utilisateurs et celui contenu dans les ressources.** Cette technique pallie les limites du langage naturel, très riche mais aussi souvent ambigu. Le thésaurus évite ainsi les risques induits par les synonymies, les homonymies et les polysémies présentes dans le langage naturel. Contrairement à un dictionnaire auquel il est souvent rapproché, un thésaurus ne fournit qu'accessoirement des définitions, les relations des termes et leur sélection l'emportant sur la description des significations.

Le terme thésaurus vient du grec ancien et signifie « trésor ». En latin, on écrit thesaurus, sans accent, qui signifie « recueil », « répertoire ». Les deux orthographes - thésaurus et thesaurus, un xénisme qui reprend directement la forme latine - sont admises, mais la première est la plus employée dans la littérature francophone. Les anglo-saxons privilégient la version latine et aussi au pluriel : thesaurus/thesauri. En suivant cette approche, on écrit soit un thesaurus, des thesauri, soit un thésaurus, des thésaurus.

Le thésaurus apparaît au XVIème siècle grâce à publication en 1531 du dictionnaire *Thesaurus lingua latinae* du lexicographe français Robert Estienne, dont l'objectif était de faire un panorama de la latinité depuis ses débuts jusqu'à les études étymologiques d'Isidore de Séville, le saint-patron des informaticiens et de l'Internet. Au XVIIème siècle, l'évêque

[45] AITCHISON, Jean. The thesaurofacet : a multipurpose retrieval language tool. Journal of Documentation, Vol. 26 Iss: 3, 1970. Pages 187 – 203.

[46] http://fr.wikipedia.org/wiki/Th%C3%A9saurus

John Wilkins, enthousiaste d'un « langage philosophique universel », c'est-à-dire, d'un système d'écriture basé non sur un alphabet, mais sur un système idéographique compréhensible internationalement, créa un type de dictionnaire amélioré constitué de relations entre les mots et destiné à organiser les objets du référentiel d'histoire naturelle de la Royal Society à Londres. En 1852, le lexicographe anglais Peter Mark Roget, basé sur l'oeuvre philosophique de Leibniz, a conçu une liste classée de mots connexes, son *Thesaurus of English Words and Phrases*, d'où est sorti le Thésaurus de Roget, qui peut être vu comme un **système de classification** et **qui a été employé pour la constitution des schémas d'arborescence de catégories de Wikipedia.**

Seulement à partir de la décennie 1970, des normes pour le thésaurus ont été édités et il prend de l'importance par l'informatisation dès les années 1990.

C'est un système de recherche privilégié notamment en raison de l'utilisation des opérateurs booléens qui permet alors de faire des équations logiques de recherche.

Parmi les plusieurs définitions de « thésaurus », dans la littérature francophone ou anglophone, toutes sont d'accord, d'une manière générale, qu'il s'agit d'un **vocabulaire contrôlé, outil d'indexation et d'aide à la recherche, comprenant des relations sémantiques entres les termes.**

On peut retenir la définition publié par l'ADBS[47], d'après celle élaborée par Danièle Dégez et Dominique Ménillet [5, Dégez ; Ménillet] :

« Liste organisée de termes contrôlés et normalisés (descripteurs et non-descripteurs) servant à l'indexation des documents et des questions dans un système documentaire. Les descripteurs sont reliés par des relations sémantiques (génériques, associatives et d'équivalence), exprimés par des signes ou symboles conventionnels. Les synonymes (non-descripteurs ou termes interdits) sont reliés aux descripteurs par une seule relation d'équivalence. »

Plus synthétique est la version anglaise du Webster's Online Dictionnary[48] : « Un langage documentaire contrôlé et dynamique contenant des termes connexes de façon sémantique et générique et qui traite exhaustivement un domaine spécifique de la connaissance. »

---

[47] http://www.adbs.fr/thesaurus-18864.htm?RH=OUTILS_VOC

[48] http://www.websters-online-dictionary.org/definitions/thesaurus

Un thésaurus est, donc, un type particulier de langage documentaire. Il est constitué d'un ensemble structuré de concepts représentés par des termes, pouvant être utilisés pour l'indexation de documents dans une base de données à des fins de recherche documentaire. **L'utilisation du thésaurus permet de pallier les imperfections du langage naturel dans un but d'indexation.** Le langage naturel, soit notre langage quotidien, contient de nombreux soucis de polysémie et de synonymie. Le thésaurus est un outil d'indexation combinatoire à vocabulaire contrôlé, c'est-à-dire, les termes qui le constituent sont sélectionnés et ne peuvent être modifiés, sauf lors des mises à jour. C'est un langage post coordonné car les descripteurs définissant les concepts peuvent être combinés ou associés *a posteriori* lors de la recherche d'information.

1.2.3.2 Aspects constitutifs :

L'indexation en langage documentaire grâce au thésaurus permet une homogénéité du mode d'indexation qui ne dépend alors plus de la culture de l'indexeur. Le thésaurus est ambivalent. Il est utilisé à l'entrée et la sortie de la chaîne documentaire : lors de la phase d'indexation et lors de la phase d'interrogation par l'usager. La capacité de recherche via le thésaurus est importante puisque ce dernier utilise un langage combinatoire qui associe et recoupe les mots de la recherche pour optimiser l'obtention de résultats.

Trois types de termes composent un thésaurus :

- les descripteurs : utilisés pour indexer un document, il s'agit de l'ensemble des mots autorisés pour indexer ;

- les non-descripteurs : par convention ne peuvent pas être employés pour indexer un document, qui servent à renvoyer au descripteur à utiliser. Ils sont utilisés lors de la recherche.

- les mots outils : ce sont des descripteurs qui ne peuvent être utilisés seuls, vu qu'ils sont coordonnés avec au moins un autre descripteur.

Les relations entre concepts sont de trois types :

- relation hiérarchique *stricto sensu*, base de la hiérarchie du thésaurus. Elles sont représentées par les sigles TG (terme générique) et TS (termes spécifiques), ou, en anglais, BT (broader term) et NT (narrower term).

Cette hiérarchie permet de régler la précision de l'indexation ou de l'interrogation. L'indexation s'appuiera autant que possible sur l'identification des termes spécifiques (donc

du niveau le plus bas possible), alors que la recherche selon les cas pourra faire appel aux termes génériques pour augmenter le nombre de réponses.

Cette relation s'appuie sur : des concepts génériques repérés par le sigle TG. Ils désignent les entités ou concepts généraux en référence aux autres concepts et au domaine considéré et des concepts spécifiques repérés par le sigle TS, qui précisent et identifient les entités ou concepts plus précis à l'intérieur du champ sémantique d'un terme générique donné.

- relation d'association enrichissant le réseau de relations hiérarchiques selon d'autres axes de type sujets connexes. Ces relations peuvent être de nature très variée : causalité, localisation, relations de nature temporelle, composition, etc.

Les relations d'association sont représentées par le sigle TA (terme associé), ou en anglais Related Term (RT). Ces relations entre concepts permettent à l'utilisateur de modifier progressivement son interrogation ou de l'élargir sur d'autres bases que la relation hiérarchique.

- appartenance à un groupe de concepts : Il est courant de sélectionner et regrouper des concepts selon un critère spécifique, tels que leur pertinence à un domaine particulier. Ces regroupements de concepts sont appelés suivant les contextes : thèmes, domaines, champs sémantiques, microthésaurus.

Un thésaurus s'élabore, soit manuellement par la voie d'une ou plusieurs personnes, grâce à une intelligence humaine, soit de manière automatique, par le biais de l'intelligence artificielle, grâce à des logiciels de construction automatique de thésaurus, soit par un mélange de l'approche humaine et automatique. Des systèmes de traitement automatique de textes (indexation automatique) permettent l'extraction des termes les plus fréquents d'un corpus et dans une certaine mesure facilitent l'émergence de leurs relations sémantiques. Ces infos-logiciels utilisent également des outils linguistiques de reconnaissance morpholexicale et syntaxique. Cela permet, donc, envisager le remplacement des indexeurs par des machines et de l'intelligence humaine par l'intelligence artificielle, grâce, par exemple, à l'analyse de la fréquence des mots clés, selon les critères de pertinence et importance. Il s'agit d'une démarche pragmatique et continue de rationalisation des termes descriptifs, ce qui génère un vocabulaire contrôlé puisqu'il résulte d'un long processus de tri des mots, appellations et expressions utilisés dans un domaine particulier.

Il existe trois méthodes de constitution d'un thésaurus :

- analytique (a priori) : à partir des mots clefs de l'indexation ;

- synthétique (a posteriori) : à partir de listes de mots-clés préétablies à l'aide de dictionnaires, lexiques, glossaires etc.

- mixte.

En vue de la meilleure adéquation au domaine considéré, les termes sont inventoriés, comparés, mis en relation et finalement hiérarchisés pour rendre compte des traits essentiels du domaine. Cette hiérarchie s'appuie sur une typologie : chaque terme appartient à une catégorie qui le situe par rapport à tous les autres termes retenus et qui fixe de cette manière sa priorité d'emploi. La hiérarchie des termes peut être différente d'un thésaurus à un autre et même sous réserve d'incohérence dans un usage ou un autre du même thésaurus.

Finalement, en partant du niveau le plus haut et correspondant au domaine du thésaurus, on trouve d'abord les subdivisions majeures représentant les composantes du domaine - subdivisions souvent nommés microthesaurus. Un thésaurus peut aussi concerner plusieurs domaines, comme cela est le cas d'un macrothésaurus.

Il demeure toujours une dimension arbitraire dans la hiérarchie d'un thésaurus, soit dans le choix des termes, soit dans leur position hiérarchique.

À guise d'illustration, sont disponibles en ligne, le thésaurus pour l'éducation nationale MOTBIS et le thésaurus multilingue de l'UNESCO.

Comme méthodologie à suivre pour l'élaboration d'un thésaurus, Danièle Dégez [6, Dégez] propose la rédaction d'un cahier des charges, de façon à couvrir les aspects suivants :

- le type et le nombre d'éléments à collecter (livres, articles de revues, images fixes ou animées …) ;

- la/les spécialité(s) concernée(s) par le thésaurus ;

- le public et ses besoins ;

- les objectifs ;

- les logiciels à être utilisés ;

- les ressources financières et humaines disponibles ;

- le planning (même si l'on considère qu'un thésaurus n'est jamais achevé).

On peut encore distinguer les thésaurus en fonction du mode de regroupement des termes (thésaurus à facettes) ; de la variété linguistique des termes (mono ou multilingue) ; des domaines de connaissances couverts (thésaurus spécialisé ou sectoriel, thésaurus encyclopédique).

L'actuel norme internationale pour l'élaboration des thésaurus est la norme ISO 25964-1:2011: Thésaurus et interopérabilité avec d'autres vocabulaires[49], qui remplace les normes ISO 2788:1986 (Principes directeurs pour l'établissement et le développement des thésaurus monolingues) et ISO 5964:1985 (Principes directeurs pour l'établissement et le développement des thésaurus multilingues) et au niveau français les normes NF Z47-100:1981 (thésaurus monolingue) et NF Z47-101:1990 (thésaurus multilingue).

La nouvelle norme ISO est en grande partie en consonance avec les spécifications SKOS (spécifications en langage RDF développé par le W3C, pour la publication et l'utilisation des systèmes d'organisation de concepts, tel comme les thésaurus, dans le cadre du Web sémantique.

Les relations d'équivalence entre termes représentant un même concept permettent de lutter contre la polysémie. La nouvelle norme ISO 25964-1:2011 désigne parmi l'ensemble des termes pouvant représenter un même concept : un terme préférentiel (descripteur) et des termes non-préférentiels (non-descripteurs), base de l'univocité du concept. Cette relation est représentée par le sigle EP (abréviation d'"Employé Pour"). La relation inverse des termes non-préférentiels vers le terme préférentiel est représentée par le sigle EM (abréviation d'"Employer").

Ce sont des variantes des termes spécifiques (synonymie ou quasi-synonymie) considéré comme "équivalent" dans le langage courant, ou des termes représentant des concepts assez proches pour être considérés comme "équivalent" pour les dispositifs d'accès à l'information.

1.2.3.3 Le thésaurus dans l'univers des langages documentaires

Avec l'élaboration de normes et d'applications informatiques spécialisées, comme dans le domaine voisin des ontologies, une convergence des problématiques, telles comme ressources, hiérarchie, réutilisation, etc. a rapproché les thésaurus des ontologies.

Comme bien observe dans son mémoire INTD Béatrice Pierre, [21, Pierre] malgré l'abondance des définitions données par les uns et les autres, le terme thésaurus demeure

[49] À ce propos, voir la présentation de Hélène Rabault et Hélène Zysman :http://www.bnf.fr/documents/afnor2011_norme_thesaurus.pdf

assez vague. Les frontières entre les différents types de langages documentaires sont flues et certains professionnels des sciences de l'information ont parfois tendance à faire l'amalgame entre plusieurs notions :

Thésaurus et dictionnaire : le thésaurus est souvent perçu par les non-spécialistes comme un dictionnaire. La différence est assez simple, le thésaurus n'a pas pour vocation d'être aussi exhaustif et le dictionnaire ne comporte pas de relations entre les termes excepté la synonymie ou l'antonymie. Dans un dictionnaire, le mot est le point de départ et on en dégage plusieurs concepts, dans le cas du thésaurus, c'est l'inverse, le concept est décrit par plusieurs termes.

Thésaurus et glossaire : le but d'un glossaire est d'apporter des définitions à des termes dans un domaine donné. Même si un thésaurus est lui aussi un ensemble de vocables rattaché à une spécialité, il n'y a pas d'obligation de spécifier le sens de certains mots par une définition, même si les auteurs sont formellement invités à le faire.

Thésaurus et liste de vedettes-matière : la distinction principale entre ces deux notions réside dans le fait qu'une liste de vedettes-matière a été conçue pour cataloguer une collection documentaire et qu'un thésaurus a pour objectif d'indexer un corpus de documents. L'unité de base de la liste est la vedette, soit un sujet qui regroupe plusieurs concepts tandis que celle du thésaurus est le descripteur qui ne renvoie qu'à un seul concept. En outre, la liste de vedettes-matière est essentiellement destinée au bibliothécaire tandis que le thésaurus offre un double usage documentaliste/usager.

Thésaurus et taxonomie : la différence est bien plus ténue. L'on peut dire que le degré de précision sera l'élément clé pour distinguer les deux langages documentaires. Un thésaurus est supposé être bien plus riche qu'une taxonomie qui n'est à la base qu'une classification.

Thésaurus et ontologie : La littérature grise traite souvent les similarités entre les deux objets, ce qui est tout à fait censé, mais rarement des différences. Les ontologies sont d'abord présentées comme un prolongement du thésaurus et une manière de le rendre viable et fonctionnel sur le web. Dès lors, une ontologie serait une version de thésaurus bien plus sophistiquée, avec une vision pluridimensionnelle des termes.

Finalement, à y regarder de plus près, il n'y a qu'une notion de granularité entre chaque langage documentaire.

À fin d'illustrer les nuances entre ces langages documentaires, Béatrice Pierre propose encore la reprise du schéma de Fred Leise [15, Leise, p. 124], en l'ajoutant les deux dernières lignes :

Cercle de synonymes

+termes choisis

= liste d'autorité

+ termes génériques/termes spécifiques

= taxonomie

+ termes reliés

= thésaurus

+ relations sémantiques entre les concepts (méronymie)

= ontologie

A l'aide à ce schéma, on peut visualiser pourquoi, même entre quelques professionnels de sciences de l'information les langages documentaire peuvent se mêler, compte tenu de la complexité de toutes les relations qui les composent.

1.3 L'indexation – notion et principes généraux

L'indexation est la première opération de description du contenu du document, c'est aussi la plus allégée puisqu'elle se borne à rendre compte des idées ou thèmes d'un document à l'aide de quelques mots soigneusement choisis.

C'est l'opération intellectuelle qui vise à repérer les idées émises ou les thèmes ou sujets traités dans un document et à les décrire à l'aide de mots ou expressions. On identifie ces thèmes ou idées à l'occasion de l'analyse documentaire et on étiquette ceux-ci par des mots pivot, appelés mots-clés, d'une manière générale.

Selon la définition établie par l'ADBS[50], l'indexation est le processus destiné à représenter, au moyen des termes ou indices d'un langage documentaire ou au moyen des éléments d'un langage libre, les notions caractéristiques du contenu d'un document

[50] http://www.adbs.fr/indexation-1--17361.htm?RH=OUTILS_VOC

(ressource, collection) ou d'une question, en vue d'en faciliter la recherche, après les avoir identifiées par l'analyse.

L'indexation au sens documentaire peut être :

libre : indexation en langage naturel à l'aide de notions et concepts choisis librement par l'indexeur sur la base d'un seul document.

- contrôlée : indexation en langage contrôlé et normalisé par sélection des termes d'indexation dans un langage documentaire de référence, tel comme un lexique, un thésaurus ou une encyclopédie, qui sont ainsi construites sur l'indexation de textes antérieures ; le référentiel est donc pris ou élaboré au préalable.

L'indexation libre a l'avantage de proposer un vocabulaire assez large et plus proche des utilisateurs, toutefois elle peut apporter une quantité assez considérable de synonymie. De plus, un même mot clé peut être utilisé par différents indexeurs et recouvrir des réalités différentes. Le vocabulaire contrôlé, même s'il peut s'avérer trop spécialisé pour certains publics, a l'avantage d'être cohérent sur le plan terminologique. Une solution qui semble efficace est combiner les deux techniques afin d'indexer au plus précisément, même si dans la pratique le processus reste assez compliqué à mettre en place.

Au sens informatique, l'indexation est la création d'un fichier inversé ou index d'interrogation, liste ordonnée des termes interrogeables assortis des références permettant de retrouver l'information.

L'indexation est donc à la fois l'opération intellectuelle de description du contenu d'un document à l'aide de mots-clés ou descripteurs, choisis lors du traitement documentaire et aussi l'opération automatique de création ou de mise à jour d'une base de données pour accélérer le processus de recherche d'information.

C'est seulement dans la moitié du XXème siècle que le terme indexation apparaît, vis-à-vis la croissance exponentielle du nombre de documents crées par l'homme. Avec l'avènement de l'internet et la possibilité d'y trouver des documents en texte intégral, l'indexation est devenue une question centrale.

D'après Marie Didier [7, Didier] elle doit tenir compte des moyens humains et techniques disponibles, afin d'éviter le travail inutile ou les ambitions impossibles à réaliser (point de vue des moyens), et tenir compte non moins impérieusement des fins visées et des usages attendus de la base documentaire constituée (point de vue des usages).

L'indexation peut être aussi manuelle contrôlée ou automatique. L'indexation manuelle contrôlée décrit le contenu d'un document en choisissant les descripteurs dans le langage documentaire de référence. Le résultat apporte beaucoup de pertinence, cependant le volume de documents indexés par jour reste faible et le langage contrôlé ne correspond pas obligatoirement au vocabulaire de l'utilisateur.

À son tour, l'indexation automatique, fait exclusivement par des moyens informatiques via logiciels spécialisés, est capable d'extraire automatiquement tous les mots contenus dans l'ensemble du corpus des documents, avec ou sans l'outil de contrôle du vocabulaire. Cette méthode garantit l'homogénéité du processus d'indexation, ce qui rendre possible la ré-indexation des documents. Néanmoins, l'indexation automatique, non enrichie par des traitements linguistiques, est peu pertinente, vu que tous les mots du document sont indexés.

Selon Muriel Amar [1, Amar, p. 8], la discordance entre les deux notions n'est pas réellement fondée : l'indexation automatique tentant de reproduire l'indexation manuelle, elles évoluent en parallèle mais sans se faire profiter mutuellement de leurs avancées.

La littérature spécialisée reconnaît plusieurs typologies d'indexation, qui peut prendre plusieurs formes, marquées notamment par des oppositions :

- indexation en texte intégral : index constitué de tous les mots du document. La machine doit analyser, décortiquer mot à mot pour en extraire le thème grâce aux liens sémantiques existant entre les termes ;

- indexation assistée par ordinateur : repérage des concepts significatifs du document pour le caractériser avec des descripteurs validés ou non par un professionnel ;

- indexation par assignation : sélection des termes significatifs sur la base d'un langage documentaire.

L'utilisation d'un lexique pour indexer un document a été unanimement rejetée par les professionnels des sciences de l'information et de la linguistique. La seule valeur admise est celle des langages documentaires dont l'unité de base est le descripteur. Le descripteur renvoie à un concept et donne accès au document ou à un corpus documentaire.

L'indexation a donc deux fonctions reconnues :

-indiquer de manière concise le contenu d'un document, c'est-à-dire, le descripteur qui doit exprimer un concept ;

-permettre de rechercher efficacement l'information contenue dans un document, c'est-à-dire, le descripteur est le point d'accès au document.

Finalement, il ne faut pas confondre les notions de mot-clé et descripteur. Mot-clé est le mot choisi dans le titre ou le texte d'un document, sans référence à un lexique ou à un thésaurus, caractérisant son contenu et permettant la recherche de ce document.

Descripteur est le mot ou groupe de mots retenus dans un thésaurus ou lexique de référence et choisi parmi un ensemble de termes équivalents pour représenter sans ambiguïté une notion contenue dans un document ou dans une demande de recherche documentaire.

Les nouveaux types d'indexation tels que les folksonomies et les tags, seront abordés dans le chapitre suivant puisqu'ils sont liés à l'essor du Web 2.0 dit collaboratif et sont l'objet de structuration dans le cadre du Web sémantique.

2 Les langages documentaires face au Web sémantique

2.1 Qu'est-ce que le Web Sémantique ?

Le Web sémantique[51] désigne un ensemble de technologies visant à rendre le contenu des ressources du World Wide Web accessibles et utilisables par les programmes et logiciels, grâce à un système de métadonnées formelles, utilisant notamment la famille de langages développés par l'organisme de standardisation W3C (World Wide Web Consortium), chargé de promouvoir la compatibilité des technologies employés dans le World Wide Web.

Ici, à propos du web sémantique, est indispensable la citation au billet extrêmement didactique d'Alexandre Monnin, « Qu'est-ce que le Web Sémantique ? », publié sur le C/blog du Ministère de la Culture et de la Communication.

L'expression web sémantique a été écrite pour la première fois par Tim Berners-Lee, aussi fondateur du consortium W3C, dans le célèbre article[52] publié dans la revue Scientific American en 2001. La notion de métadonnées utilisables par les machines avait déjà été proposée dès 1994[53] par le même Tim Berners-Lee.

Le web sémantique vise à dépasser la simple recherche d'informations par comparaison syntaxique avec des mots clés, en complétant cette définition syntaxique des informations par une caractérisation sémantique plus formelle de leur contenu via des schémas de métadonnées (« données sur la donnée »)[54].

Dans son article fondateur, Berners-Lee souligne que "le Web sémantique n'est pas un Web séparé, mais une extension du Web actuel dans lequel l'information est munie d'une signification bien définie permettant aux ordinateurs et aux personnes de mieux travailler en coopération."

---

[51] http://fr.wikipedia.org/wiki/Web_s%C3%A9mantique

[52] http://www.scientificamerican.com/article.cfm?id=the-semantic-web Pour la traduction française : http://www.urfist.cict.fr/archive/lettres/lettre28/lettre28-22.html

[53] http://www.w3.org/Talks/WWW94Tim/

[54] http://hal.archives-ouvertes.fr/docs/00/08/32/84/PDF/AFIS06_version_3.4.pdf , page 5.

Le web sémantique ne remplace pas le Web actuel, il joue plutôt le rôle d'une brique supplémentaire permettant de **structurer l'information** contenue sur le Web, rendant exploitable aux utilisateurs des milliards de documents disponibles sur le web. Son but est lié à la recherche de l'information: c'est de faire en sorte que l'internaute trouve plus facilement et plus rapidement l'information qu'il recherche.

D'après son idéalisateur, le Web sémantique est entièrement fondé sur le Web tel comme l'on le connait à l'heure actuelle et ne remet pas en cause ce dernier. Le Web sémantique s'appuie donc sur la fonction primaire du Web : un moyen de publier et consulter des documents. Mais les documents traités par le Web sémantique contiennent non pas des textes en langage naturel, dans plusieurs langues, mais des informations formalisées pour être traitées automatiquement. Ces documents sont générés, traités, échangés par des logiciels. Ces logiciels sont alors capables de :

générer des données sémantiques à partir de la saisie d'information par les utilisateurs ;

agréger des données sémantiques afin d'être publiées ou traitées ;

publier des données sémantiques avec une mise en forme personnalisée ou spécialisée ;

échanger automatiquement des données en fonction de leurs relations sémantiques ;

générer des données sémantiques automatiquement, sans saisie humaine, à partir de règles d'inférences.

Même dans les années 1990, les informaticiens pensaient déjà à mettre en place une relation sémantique entre les contenus publiés sur le web, mais les technologies n'étaient pas encore suffisamment développées pour la concrétiser. Asunción Valderrama et Arlette Boulogne [3, Boulogne ; Valderrama, p. 20] soutien même qu'il s'agit « du vin vieux dans des outres neuves ».

On peut ainsi situer l'évolution de l'internet :

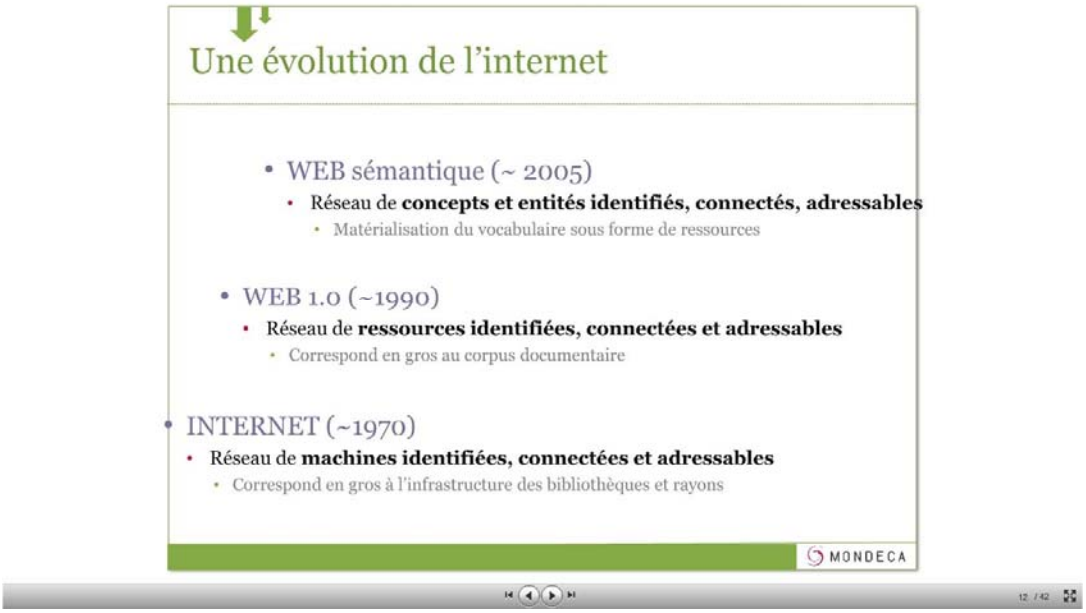

Figure 3 : Évolution de l'Internet.[55]

Le Web sémantique a donc pour objectif le partage de connaissances contenues dans des silos d'informations, appelés aussi bases de données. Les données contenues dans une base de données classique sont dites non structurées vis-à-vis d'autres bases de données s'il n'existe aucune grammaire commune entre elles.

Comme c'est le cas pour le langage humain, la structuration syntaxique et grammaticale permet la création de phrases, éléments complexes d'où peut émerger un sens compréhensible par d'autres personnes. Sans grammaire, il ne peut y avoir de dialogues entre les différentes bases de données, et sans dialogues, aucune connaissance pérenne due à des synergies cognitives ou des partages ne peut émerger.

L'existence d'une grammaire commune entre bases de données est la condition de structuration des données, donc du dialogue et de la confrontation productive des données.

En pratique, une sémantique définie strictement permettra aux machines d'effectuer des raisonnements automatisés sur les inférences et conclusions sur ses nouveaux silos de connaissance.

Le Web sémantique peut donc utiliser une myriade de nouvelles technologies pour structurer, partager et échanger les différentes connaissances qui s'y trouvent.

---

[55] Source :http://www.slideshare.net/secret/9JvClpGRzduuqx

En ce qui concerne l'emploi du terme « sémantique », Tim Berners-Lee a reconnu qu'il mettait essentiellement l'accent sur les langages de représentation des connaissances et l'automatisation des procédures de recherche, issus de l'intelligence artificielle :

« Le terme sémantique prête un peu à confusion car la sémantique s'intéresse au sens du langage pour en déduire des constructions logiques. Du coup, certains ont pensé qu'il s'agissait d'un Web qui permettrait par exemple d'effectuer des recherches sur Internet en posant des questions sous forme de phrases, en langage naturel. Or ce n'est pas son but. En fait, nous aurions dû l'appeler dès le départ « Web de données ».[56]

Le web actuellement est encore plutôt syntaxique. Syntaxe s'oppose à la sémantique, vu que la syntaxe est l'ensemble des règles précisant la manière d'écrire et/ou de disposer les informations[57]. Elle se lie à la forme, tandis que la sémantique[58] est relative à la signification d'un mot ou d'une structure linguistique, c'est-à-dire, la sémantique est liée au sens du mot. La syntaxe se rattache à la forme, et donc s'oppose au sens, la sémantique.

Le web tel que l'on l'utilise aujourd'hui est interprétable par les humains mais de façon très faible par les machines. Le web sémantique vise donc un traitement automatique des données. Le système mis en oeuvre pour rendre effectif ce web sémantique s'appuie néanmoins sur le fonctionnement cérébral humain. Les moteurs de recherche rencontrent encore des difficultés à **désambiguïser** un terme et à appréhender les notions de polysémie, homonymie, homographie... Les métadonnées des documents, qui pourtant pourraient justement aider à limiter le bruit et l'extraction de documents non pertinents, quand ils en ont, ne sont pas toujours fiables et surtout prises en compte par les moteurs de recherche.

Le web sémantique a pour objectif de modéliser les métadonnées des documents et de les relier entre elles pour créer du sens, tout en permettant l'intégration de différents contenus, applications et systèmes.

La structuration de l'information ne suffit pas à la rendre intelligible par un utilisateur ou une machine. Les regroupements d'information doivent être complétés par du sens pour être pleinement exploitables.

---

[56] Extrait d'une interview parue dans la Recherche intitulée « Le Web va changer de dimension » http://www.larecherche.fr/content/recherche/article?id=6566

[57] http://www.cnrtl.fr/definition/syntaxe

[58] http://www.cnrtl.fr/definition/s%C3%A9mantique

L'idée sous-jacente au web sémantique est donc d'ajouter dans une ressource donnée, comme par exemple, une page web des informations formelles (grâce à des balises invisibles à l'utilisateur) afin que les logiciels de recherche puissent collecter cette ressource et l'exploiter.

Le véritable apport du Web Sémantique réside avant tout dans son adaptation au Web. Il s'agit bien d'un web sémantique qui tente d'exploiter au mieux cet espace fait de ressources qu'est le Web. Les machines sont à l'aide pour organiser, filtrer, relier, présenter, contextualiser les **contenus** et pour modéliser, contrôler, publier, sélectionner des **connaissances**.

Toutefois, il est important de souligner que les technologies du web sémantique sont d'abord adoptées par les secteurs qui ont une longue tradition de contenus organisés selon des référentiels, dans un souci de valorisation du capital informationnel et qui les lie aux fondements du knowledge management (gestion de connaissances).

Le web sémantique est fondé sur les protocoles et langages standards du Web suivants:

- le protocole HTTP ;

- les Uniform Resource Identifiers (URI) ;

- le langage XML (dans le cas, majoritaire, où RDF est sérialisé en XML).

À ces standards s'ajoutent ceux qui sont propres au web sémantique :

- RDF : modèle conceptuel permettant de décrire toute donnée ;

- RDF Schema : langage permettant de créer des vocabulaires, ensembles de termes utilisés pour décrire des choses ;

- OWL : langage permettant de créer des ontologies, vocabulaires plus complexes servant de support aux traitements logiques (inférences, classification automatique…) ;

- SPARQL : langage de requêtes pour obtenir des informations à partir de graphes RDF.

2.2 L'adaptation des langages documentaires aux standards du Web sémantique et interopérabilité

L'adaptation de l'ensemble des langages documentaires aux standards du web sémantique nécessite une gamme de protocoles, modèles et autres langages qui ont été

déjà établis Ces dispositifs dépendent eux-aussi de protocoles établis par le W3C pour rendre les liens sémantiques possibles. Cependant, la multiplicité des langages qui constituent les fondations du web sémantique tend à complexifier cette infrastructure. Une démonstration assez brève des standards du W3C qui composent le Web Sémantique seront abordés synthétiquement ci-dessous.

On peut visualiser les différentes briques technologiques composant le Web sémantique d'après la vision de Tim Berners-Lee en 2000 dans la formalisation graphique suivante :

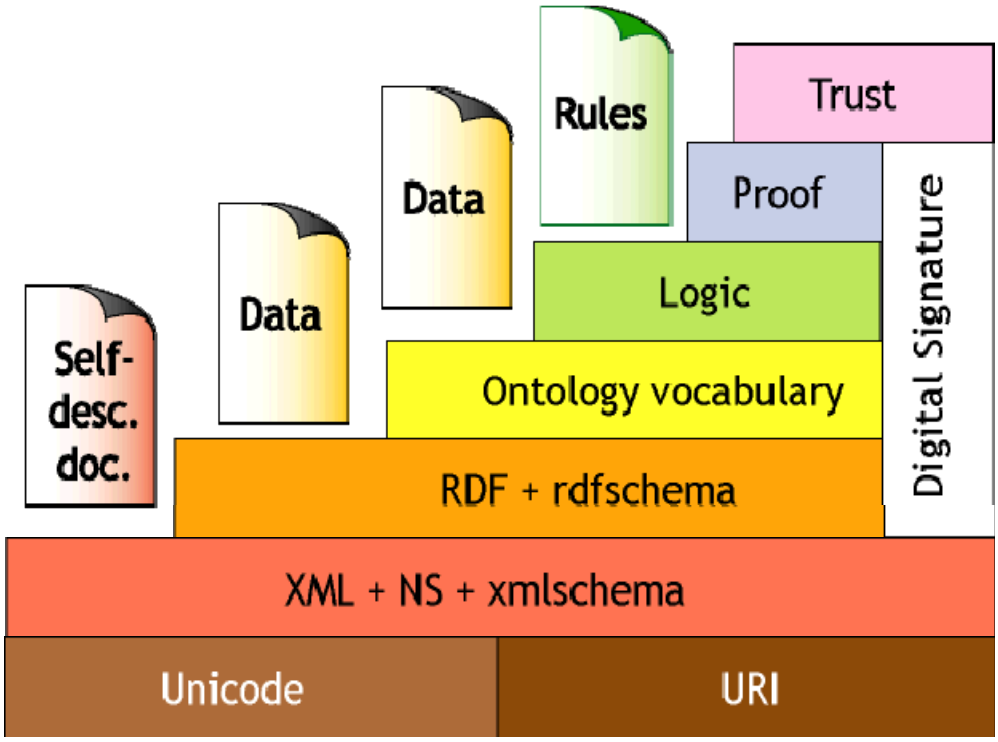

Figure 4 : Briques du Web Sémantique.[59]

À noter, dans ce « layer cake » ou « Web semantic stack », comme l'observe Alexandre Monnin [18, Monnin], que la brique la plus élevée n'est pas la vérité, « Truth », mais la confiance, « Trust »[60].

[59] Source : http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html

[60] http://cblog.culture.fr/2011/09/07/web-semantique-iri-opendat Pour savoir plus sur la brique sommet « trust », voir http://www.w3.org/2005/Incubator/webid/spec/ qui élabore le spécification WebID, conçu pour alléger les soucis liés au souvenir de logins différents, des mots de passe et des

2.2.1 HTTP : HyperText Transfer Protocol

Le « protocole de transfert hypertexte »[61] est un protocole de communication client-serveur développé pour le système web qui a été conçu par Tim Berners-Lee avec les adresses web et le langage HTML[62].

Les clients HTTP les plus connus sont les navigateurs web qui permettent à un utilisateur d'accéder à un serveur contenant les données. Il existe aussi des systèmes pour récupérer automatiquement le contenu d'un site tel que les aspirateurs de site web ou les robots d'indexation. Ces clients, à son tour, se connectent à des serveurs HTTP[63] ou serveurs web, logiciels qui servent à répondre des requêtes.

La liaison entre le client et le serveur n'est pas toujours directe. Il existe des machines intermédiaires servant de relais, tel comme le proxy (ou serveur mandataire), qui modifie les réponses et requêtes qu'il reçoit et peut gérer un cache des ressources demandées ; la passerelle (ou gateway) est un intermédiaire modifiant le protocole utilisé et le tunnel, qui transmet les requêtes et les réponses sans aucune modification, ni mise en cache.

2.2.2 URI : Uniform Ressource Identifier

L'identifiant uniforme de ressource[64] est une courte chaîne de caractères identifiant une ressource sur un réseau physique ou abstrait, et dont la syntaxe respecte les normes du W3C.

Les URIs sont la technologie de base du World Wide Web car tous les hyperliens du Web sont exprimés sous forme d'URI. Ils permettent donc l'identification des ressources sur Internet mais également de créer du lien entre elles.

---

paramètres pour les sites web. Il est également conçu pour fournir un mécanisme universel et extensible pour exprimer des informations publiques et privées concernant les utilisateurs du Web.

[61] http://fr.wikipedia.org/wiki/HTTP

[62] Hypertext Markup Language

[63] http://fr.wikipedia.org/wiki/Serveur_HTTP

[64] http://fr.wikipedia.org/wiki/Uniform_Resource_Identifier

Un URI doit permettre d'identifier une ressource de manière permanente, même si la ressource est déplacée ou supprimée. Bien que les URI soient très largement utilisés dans le monde informatique, on en retrouve d'autres applications, comme dans la bibliothéconomie : ainsi le code ISBN (International Standard Book Number), qui est l'identifiant unique d'un livre, et permet de retrouver celui-ci depuis n'importe quelle librairie ou bibliothèque, dans le monde entier. On peut considérer également les codes-barres comme une métaphore d'URI, dans le monde physique : un code-barre ne localise pas un produit mais l'identifie, comme l'ensemble des exemplaires d'un produit et pas chaque exemplaire individuellement.

Un URI peut être de type « locator » ou « name ».Un Uniform Resource Locator (URL) est un URI qui, outre le fait qu'il identifie une ressource sur un réseau, fournit les moyens d'agir sur une ressource ou d'obtenir une représentation de la ressource en décrivant son mode d'accès primaire ou son « emplacement » dans le réseau. Un Uniform Resource Name (URN) est un URI qui identifie une ressource par son nom dans un espace de noms. Un URN peut être employé pour parler d'une ressource sans que cela préjuge de son emplacement ou de la manière de la référencer. Par exemple, l'URN urn:isbn:978-85-86368-21-9 est un URI qui, étant un numéro de l'International Standard Book Number (ISBN), permet de faire référence à un livre, mais ne suggère ni où, ni comment en obtenir une copie réelle.

Le point de vue actuel du groupe de travail qui supervise les URI est que les termes URL et URN sont des aspects dépendants du contexte des URI, et que l'on a rarement besoin de faire la distinction entre les deux. Dans les publications techniques du W3C, le terme URL n'a pas été reconnu pendant longtemps, parce qu'il était rarement nécessaire de faire une distinction entre les URL et les URI. Cependant, dans des contextes non techniques, le terme URL reste omniprésent. De plus, le terme adresse web, qui n'a pas de définition formelle, est souvent employé dans des publications non techniques comme synonyme d'URL ou URI, bien que généralement il ne se réfère qu'aux protocoles HTTP.

D'après Alexandre Monnin [18 Monnin], l' « U » qui compose la sigle URI est plutôt « **universel** « que « **uniforme** ». « Universel », ici, signifie qu'**une ressource est dotée d'un certain sens**, qui demeure peu ou prou constant quelles que soient les représentations transmises par un serveur. Celles-ci sont susceptibles d'évoluer dans le temps ou de façon ponctuelle, selon les spécificités de la requête posée.

Par exemple :

« Ainsi, la page d'accueil du journal Le Monde sera-t-elle une ressource. Mais nul ne peut accéder « à la page d'accueil du Monde » conçue dans toute sa généricité et son

universalité. On y accède à un moment donné, les représentations de cette ressource variant au fil du temps, au rythme de la succession des titres de l'actualité à court terme, de celui des changements de chartes graphiques, à plus long terme. »

Et il ajoute encore : « En d'autre terme, le Web a toujours été un Web de ressources inaccessibles dont, en définitives, seules les représentations s'échangent entre clients et serveurs. Conséquence immédiate, dès la fin des années 90 on ne parla plus d'URLs mais d'URIs car le principe d'identification des ressources (et non l'adressage des documents) était désormais acquis (le glissement du document vers la ressource coïncidant avec l'adoption définitive des URIs en lieu et place des URLs). Dès lors qu'une ressource n'est de toute façon jamais accessible, fut-elle une « page » comme celle du Monde ou une personne, on ne peut plus distinguer les URIs qui identifient de soi-disant documents de celle qui identifient des choses. Dans les deux cas, elles identifient des ressources, quelles qu'elles soient, et donnent accès à leurs représentations. »

Les interactions entre la ressource, sa représentation et son URI peuvent être mieux visualisés dans le schéma suivant :

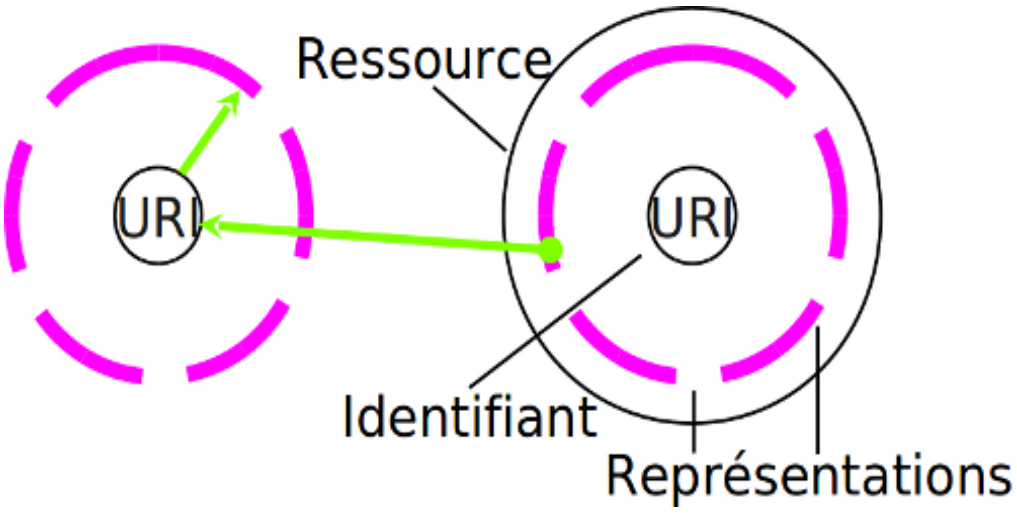

Figure 5 : Schéma URI.[65]

2.2.3 XML : Extensible Markup Language

Le langage de balisage extensible[66] est un langage informatique de balisage générique. Cette syntaxe est dite extensible car elle permet de définir différents espaces de

[65] Source : http://www.archivesdefrance.culture.gouv.fr/thesaurus/technos-web-semantique.html

noms, c'est-à-dire des langages avec chacun leur vocabulaire et leur grammaire. Cette syntaxe est reconnaissable par son usage des chevrons « < > » encadrant les balises. L'objectif initial est de faciliter l'échange automatisé de contenus complexes entre systèmes d'informations hétérogènes, toujours de façon à **promouvoir l'interopérabilité**.

Avec ses outils et langages associés une application XML respecte généralement certains deux grands principes :

- la structure d'un document XML est définie et validable par un schéma ;

- un document XML est entièrement transformable dans un autre document XML.

Dans le cours du temps, le langage XML a été largement utilisé, constatant et favorisant l'interopérabilité. La disponibilité d'une syntaxe standard et d'outils de manipulation réduit significativement le coût du cycle de développement, permettant à des programmes de modifier et de valider, sans connaissances préalables, des documents écrits dans ce langage. En effet, avant l'avènement du populaire langage généraliste de description de données qu'est XML, les concepteurs de logiciels avaient pour habitude de définir leurs propres formats de fichiers ou leurs propres langages pour partager les données entre programmes. Ceci nécessitait de concevoir et de programmer des analyseurs syntaxiques dédiés, ces tâches et bien d'autres s'effectuent désormais avec de outils XML standardisés.

Ses deux caractéristiques essentielles – être générique et extensible - permettent de structurer une grande variété de contenus, car le langage employé- sa grammaire et son vocabulaire - peut être redéfini.

Enfin, XML est une syntaxe très générique de balisage, propre à de nombreux usages et qui sert à stocker et à transférer des données structurées entre systèmes d'informations hétérogènes.

2.2.4 RDF : Ressources Description Framework

Le cadre de description des ressources[67] est un modèle de graphe destiné à décrire de façon formelle les ressources Web et leurs métadonnées, à fin de permettre le traitement automatique de telles descriptions. RDF est le langage de base du Web sémantique.

---

[66] http://fr.wikipedia.org/wiki/Extensible_Markup_Language

En annotant des documents non structurés et en servant d'interface pour des applications et des documents structurés, comme les bases de données et les répertoires de ressources, le modèle RDF **permet l'interopérabilité** entre des applications, échangeant de l'information non formalisée et non structurée sur le Web.

Un document structuré en RDF est un ensemble de triplets.

À son tour, un triplet RDF est une association de sujet, prédicat et, objet :

- le sujet représente la ressource à décrire ;

- le prédicat représente un type de propriété applicable à cette ressource ;

- l'objet représente une donnée ou une autre ressource.

Le sujet, et l'objet dans le cas où c'est une ressource, peuvent être identifiés par un URI ou être des nœuds anonymes. Le prédicat est nécessairement identifié par un URI.

Selon la leçon d'Emmanuelle Bermès[68], le sujet est toute "chose" sur laquelle on veut faire des assertions (sujet). Il a toujours une ressource URI. Ces "choses" ont toujours un type : une ***classe***.

Le prédicat permet d'exprimer les ***propriétés*** de ces "choses", ou les relations des "choses" entre elles. Il a toujours un URI.

L'objet peut être un texte (littéral) ou un URI.

Les ***classes*** et les ***propriétés*** sont déclarées dans des vocabulaires, *v.g.* RDF Schema, OWL80 et FOAF81 pour être réutilisées. RDF Schema et OWL sont donc des standards RDF permettant de déclarer des classes, des propriétés, et leur comportement.

Pour exemplifier : l'affirmation « São Paulo est localisé au Brésil » a « São Paulo » comme sujet, « est localisé » comme prédicat et « au Brésil » comme objet.

Le modèle RDF intègre les informations d'une manière formelle pour qu'une machine puisse les comprendre, rendant possible leur traitement automatique. L'objectif de RDF est de fournir un **encodage** et le mécanisme de l'interprétation pour représenter les ressources, de façon à permettre les logiciels d'y accéder, en utilisant des informations qui

---

[67] http://fr.wikipedia.org/wiki/Resource_Description_Framework

[68] http://www.bnf.fr/documents/afnor2011_bermes.pdf

autrement ne pourraient pas être utilisées. RDF est ainsi, tout simplement, un modèle pour encoder les données structurées.

Les ressources, représentées par ses respectives URIs, sont donc uniques, ce qui permet d'identifier exactement les ressources décrites.

Les documents RDF peuvent être écrits en différentes syntaxes, y compris en XML. Ainsi, il est possible d'avoir recours à d'autres syntaxes pour exprimer les triplets. RDF est simplement une structure de données constituée de nœuds et organisée en graphe.

Un document en format RDF correspond à un multigraphe : chaque triplet correspond alors à un arc orienté dont le label est le prédicat, le nœud source est le sujet et le nœud cible est l'objet.

Autrement dit, la structure RDF est extrêmement générique et sert de base à un certain nombre de schémas ou vocabulaires, conçus par le W3C, dédiés à des applications spécifiques, comme RDF Schema et OWL pour les ontologies, le langage SKOS pour la représentation des thésaurus et autres vocabulaires structurés et le FOAF pour la description de personnes.

Finalement, à propos de la distinction entre le langage XML et le modèle RDF, Alexandre Monnin [18, Monnin] affirme:

« On gagnerait davantage en compréhension en soulignant le contraste entre une famille de langages comme XML et RDF. XML sert essentiellement à structurer des contenus documentaires là où **RDF** est censé permettre de **décrire**… à peu près tout et n'importe quoi : des concepts, des fictions, des personnes, des « pages Web », des services… en un mot comme en cent : des ressources de toutes natures. »

2.2.5 RDF Schema : Ressources Description Framework Schema

Le modèle RDF Schema[69] ou RDFS est un langage extensible de représentation des connaissances qui fournit des éléments de base pour la définition d'ontologies ou vocabulaires destinés à structurer des ressources RDF. Les composants principaux de RDFS sont intégrés dans un langage d'ontologie plus expressif, OWL.

RDF Schema précise la notion de propriété définie par RDF en permettant de donner un type ou une classe au sujet et à l'objet des triplets.

---

[69] http://fr.wikipedia.org/wiki/RDFS

2.2.6 OWL : Web Ontology Language

Le langage d'ontologie du Web[70] est un langage de représentation des connaissances construit sur le modèle de données de RDF. Il fournit les moyens pour définir et modéliser des ontologies web structurées.

En pratique, le langage OWL est conçu comme une extension de Resource Description Framework (RDF) et RDF Schema (RDFS). OWL est donc destiné à la description de classes et de types de propriétés. De ce fait, il est plus expressif que RDF et RDFS, auxquels certains reprochent une insuffisance d'expressivité due à la seule définition des relations entre objets par des assertions. OWL apporte aussi une meilleure intégration, une évolution, un partage et une inférence plus facile des ontologies.

Aux concepts de classe, de ressource, de littéral et de propriétés des sous-classes, de sous-propriétés, de champs de valeurs et de domaines d'application déjà présents dans RDFS, OWL ajoute les concepts de classes équivalentes, de propriétés équivalentes, d'égalité de deux ressources, de leurs différences, du contraire, de symétrie et de cardinalité.

OWL est donc un moyen normalisé de décrire un vocabulaire pour les machines.

2.2.7 SPARQL : SPARQL Protocol and RDF Query Language

SPARQL[71] est un langage de requête et un protocole qui permet de rechercher, d'ajouter, de modifier ou de supprimer des données RDF disponible sur le Web.

On pourrait oser dire que SPARQL est l'équivalent du langage SQL (Structured Query Language)[72], le langage informatique normalisé qui sert à effectuer des opérations sur des bases de données. En utilisant SQL, on accède aux données d'une base de données via ce langage de requête, alors qu'avec SPARQL, on accède aux données structurées du Web des données. Cela signifie qu'en théorie, on pourrait accéder à toutes les données du Web avec ce standard, grâce à l'interopérabilité.

À côté du modèle RDF, SPARQL est une des technologies clés du Web sémantique.

SPARQL est adapté à la structure spécifique des graphes RDF, et s'appuie sur les triplets qui les constituent. En cela, il est différent du classique SQL (langage de requête qui

---

[70] http://fr.wikipedia.org/wiki/Web_Ontology_Language

[71] http://fr.wikipedia.org/wiki/SPARQL

[72] http://fr.wikipedia.org/wiki/SQL

est adapté aux bases de données de type relationnelles), mais s'en inspire clairement dans sa syntaxe et ses fonctionnalités.

SPARQL permet d'exprimer des requêtes interrogatives ou constructives :

- une requête *select*, de type interrogative, permet d'extraire du graphe RDF un sous-graphe correspondant à un ensemble de ressources ;

- une requête *construct*, de type constructive, engendre un nouveau graphe qui complète le graphe interrogé.

Par exemple sur un graphe RDF contenant des informations généalogiques, on pourra par une requête *select* trouver les parents ou grands-parents d'une personne donnée, et par des requêtes *construct* ajouter des relations frère-sœur, cousin-cousine, oncle-neveu, qui ne seraient pas explicitement déclarées dans le graphe initial.

L'adoption du langage de requête SPARQL permettra dans un exemple hypothétique, afficher la description de chaque ingrédient durant la diffusion d'une émission des recettes culinaires. Les téléspectateurs pourront associer ces ingrédients-là aux fiches qu'ils trouveraient dans le lexique (hébergé dans une base de données) du Programme National Nutrition Santé – Manger Bouger – et aux articles de Wikipédia.

Enfin, de façon très didactique, Emmanuelle Bermès ainsi illustre les langages utilisés dans le web sémantique :

Figure 6 : Le Web sémantique : un langage pour les machines.[73]

2.3 Référentiels, interopérabilité et le rôle du multilinguisme

Le mot « interopérabilité » n'est pas encore présente dans le portail lexical du Centre National de Ressources Textuelles et Lexicales – CNRTL.[74]

On peut dire que l'interopérabilité[75] est la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs, sans restriction d'accès ou de mise en œuvre.

Patrice Landry [13, Landry, p. 121] ainsi la définit : « On définit l'interopérabilité comme la capacité de plusieurs systèmes à communiquer entre eux **sans ambiguïté** et à échanger de l'information sans difficulté. »

Il est important faire la distinction entre les notions de « compatibilité » et d' « interopérabilité » : la compatibilité est une notion verticale qui fait qu'un outil peut fonctionner dans un environnement donné en respectant toutes les caractéristiques et l'interopérabilité est une notion transversale qui permet à divers outils de pouvoir communiquer, en permettant qu'ils fonctionnent ensemble.

La relation de compatibilité est binaire et concerne un ensemble fini de systèmes. A et B ou C et D sont compatibles, ou pas, si leurs constructions respectives leur permettent, ou pas, de communiquer et travailler ensemble. A et B seront dit interopérables grâce au respect de normes externes. L'interopérabilité est générale et ne concerne pas *a priori* des éléments ou systèmes particuliers. Elle existe au travers de normes et formats respectés par tout élément ou système qui souhaite intégrer un **réseau interopérable**, où les éléments se communiquent entre eux de façon fluide et normée. L'interopérabilité résulte donc d'un accord explicite entre les différents constructeurs de tels éléments.

Voici une échelle entre les trois degrés d'opérabilité :

---

[73] Source : http://www.bnf.fr/documents/afnor2011_bermes.pdf

[74] http://www.cnrtl.fr/definition/interop%C3%A9rabilit%C3%A9

[75] http://fr.wikipedia.org/wiki/Interop%C3%A9rabilit%C3%A9

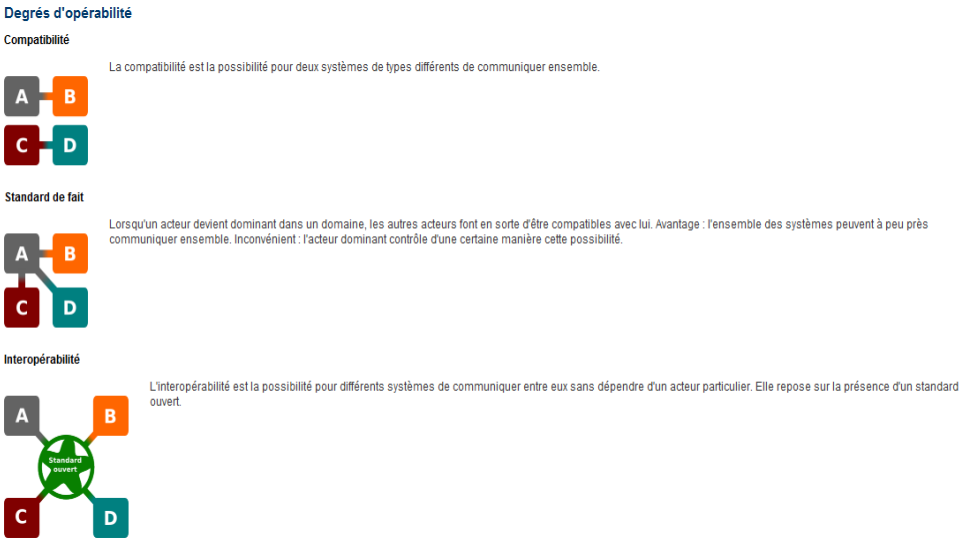

Figure 7 : Degrés d'interopérabilité[76]

L'interopérabilité est considérée comme très importante dans de nombreux domaines, spécialement dans l'informatique, vu que les différents systèmes et standards utilisés doivent pouvoir interagir sans conflits.

Compte tenu du fait que les systèmes d'information sont conçus et développés par des constructeurs divers, avec des méthodes variées, ayant pour objectif de répondre à des besoins spécifiques, ces systèmes-là doivent et suivre un ensemble de normes, que chaque système doit « greffer » dans son propre fonctionnement.

Ces normes jouent un double rôle : elles sont d'abord un indicateur de la façon dont le dialogue entre les différents systèmes doit s'opérer et elles sont aussi une **passerelle de communication**, qui va pouvoir éventuellement s'adapter aux besoins changeants des systèmes.

Deux systèmes peuvent ainsi parfaitement inter-opérer sans pour autant être conçus de la même manière. L'interopérabilité ne concerne que le comportement externe de chaque système, et non ses mécanismes internes. De plus, respecter une norme ne signifie pas ne pas avoir le droit d'en respecter d'autres ou de créer un réseau plus large de systèmes interopérables. Cela ne signifie pas non plus fermer la porte aux innovations : de simples ajouts peuvent rejoindre une norme existante, et les innovations de plus grande échelle peuvent susciter la mise en place d'une nouvelle norme, qui peut stimuler l'adoption de l'innovation et de ses applications.

---

[76] Source : http://aful.org/gdt/interop

Le problème récurrent de conflit entre normes peut avoir comme solution l'adoption des normes reposant sur des formats ouverts, et par là rapidement évolutifs. Le domaine de l'informatique illustre plus particulièrement ce point.

L'interopérabilité n'est pas par elle-même un élément concret ou un critère défini. On peut déterminer dans quelle mesure des systèmes sont interopérables en jugeant de leur respect de la norme qui a donné lieu à une interopérabilité. On comprend alors qu'on puisse parler d'interopérabilité partielle : si un logiciel, par exemple, ne respecte qu'une partie d'une norme, il ne pourra peut-être pas dialoguer correctement avec un autre programme, voire pas du tout. Dans l'absolu, seul le respect strict d'une norme donnée conduit à une interopérabilité réelle, avec l'aide de la standardisation.

Pour les systèmes informatiques, on peut citer la norme de codage Unicode, qui a permis de réduire les problèmes d'interopérabilité dus à la croissance des échanges mondiaux de fichiers de textes écrits sur le Web. Le codage de texte écrit en donnant à tout caractère de n'importe quel système d'écriture un nom et un identifiant numérique est un **exemple d'interopérabilité multilingue**.

En France, le Référentiel Général d'Interopérabilité - RGI est le document qui décrit un ensemble de normes et bonnes pratiques communes aux administrations publiques dans le domaine informatique.

2.4 Linking Open Data

En faisant la transposition du concept d'interopérabilité vers le Web sémantique, il est indispensable citer initiative Linked Open Data – Données liées ouvertes.

Lancée en 2006, le Linked Open Data propose une approche simplifiée du Web Sémantique afin d'en favoriser son développement. Cette initiative met en avant des **principes** [18, Monnin] régissant la **publication des données en ligne** :

1. Publier en RDF, c'est- à dire, donner l'accès aux données en utilisant les standards SPARQL et RDF et utiliser des URIs en guise de noms pour des choses ;

2. Utiliser des URIs accessibles via HTTP de façon à ce que les gens puissent les déréférencer (en d'autres termes, générer une représentation HTTP à partir de ces URIs) ;

3. Quand quelqu'un déréférence une URI, lui fournir des informations utiles au moyen de standards (en RDF pour répondre à un programme, en HTML pour répondre à une personne) ;

Inclure des liens vers d'autres URIs de manière à permettre aux gens d'effectuer de nouvelles découvertes (exprimer l'URI des objets liés).

Ainsi, l'initiative Linked Open Data utilise une méthode de publication de données structurées, reliées, accessibles via le web et d'abord destinées aux machines. Il s'agit de l'ensemble des données structurées accessibles sur le web, en respectant les standards HTTP, URI et RDF.

L'objectif primordial de n'est autre que la **mise à disposition, sur le Web et dans des formats standardisées et interopérables, des informations enfermées dans les bases de données, comme des silos** ; d'où son mot d'ordre : « libérez vos données ». Les technologies du Web Sémantique doivent permettre d'établir des liens entre des corpus autrefois isolés les uns des autres, voire d'autoriser de la sorte des découvertes inattendues.

Christian Bizer, Tom Heath et Tim Berners-Lee [2, Berners-Lee ; Bizer ; Heath] synthétisent l'esprit du Linked Open Data :

« Donc, tandis que le Web sémantique, ou web de données, est le but ou le résultat final de ce processus [mettre les données sur le Web de tel façon que les machines puissent les comprendre], Linked Open Data fournit les moyens d'atteindre cet objectif. »

Alexandre Monnin [18, Monnin] signale encore, comme « résultat visible de ce mouvement, le développement de l'Open Data – la publication de données administratives dans des formats interopérables, d'abord initiées dans les pays anglo-saxons (États-Unis, Angleterre…), la France rattrapant à grands-pas son retard en la matière notamment sous l'impulsion d'Etalab[77] et du projet ANR Datalift, ouvert aux problématiques des institutions culturelles à la différence de ses équivalents anglo-américains. »

À la dernière version du Linked Open Data (LOD) de septembre 2011 ont étés ajoutées plus 92 « ensemble de données », ce qui montre une croissance de approximativement 50% par rapport à la version antérieure de 2010 :

[77] Mission sous l'autorité du Premier Ministre chargée de l'ouverture de données publiques et du développement de la plateforme française Open Data « Données ouvertes ».

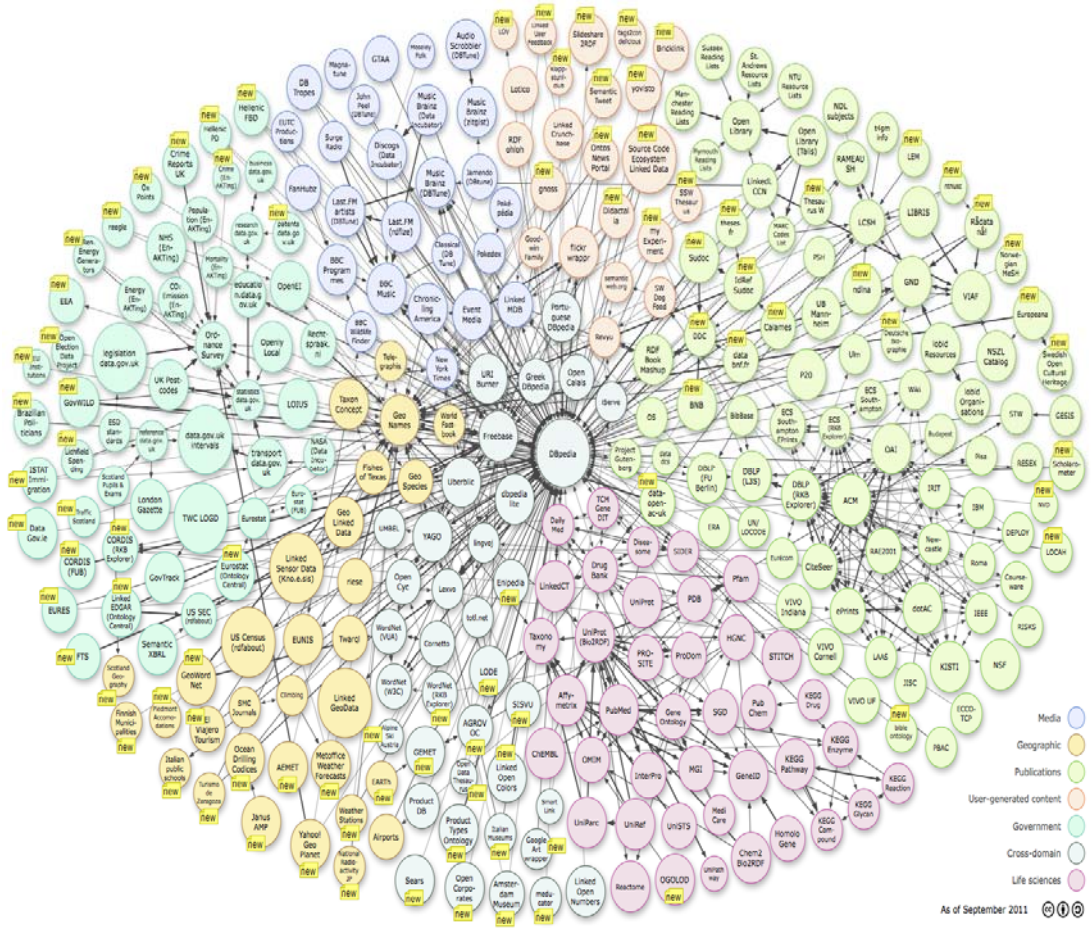

Figure 8 : Diagramme de représentation du «nuage » des données liées actuellement en ligne.[78]

À noter impérativement la place centrale occupée par DBpedia, vers où convergent tous les autres nœuds de données !

Gautier Poupeau[79] résume l'application des technologies du Web Sémantique de façon à permettre le lien entre les plusieurs référentiels et la publication de données ouvertes en affirmant que le modèle RDF permet de décloisonner les silos de données et les référentiels permettent alors de les relier, en constituant des hubs de données. Cela pourrait être aussi une bonne légende pour le diagramme ci-dessus...

[78] Source : http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.html

[79] http://www.bnf.fr/documents/afnor2011_poupeau.pdf

Finalement, le Linked Open Data est ainsi l'**espace global d'information**, comme soutiennent Tom Heath et Christian Bizer.[80]

2.5 Le rôle du multilinguisme

Il est inévitable de parler en interopérabilité sans que l'on parle aussi de multilinguisme. Face au but original du Web qui est promouvoir l'échange rapide des savoirs entre individus distants et celui du web sémantique, l'accès « universel » à l'information[81], les besoins des internautes est de retrouver l'information adéquate à un moment donné, en s'expriment dans leur langue maternelle. Ainsi, le besoin de l'internaute est trouver l'information précise, soit en français, en anglais, en portugais, en espagnol... Il réalise une requête dans une langue et trouve des résultants qui sont des ressources écrites en d'autres langues. Cependant, la profusion d'informations conduit à la détérioration de l'information utile, avec une faible exploitation.

Les langues sont en constante transformation à l'ère du numérique : on peut s'exprimer autrement. On peut simplement écrire « mdr[82] » en langage SMS ou argot Internet pour dire « C'est très drôle ».

Jamais dans l'Histoire les langues étaient tellement en contact direct. C'est remarquable la quantité de langues qui se croisent, travers de interwikis[83], dans un corpus vivant comme Wikipédia où 282 langues sont en constante relation !

La fonction du Web Sémantique est alors de permettre que l'information soit à la fois « human-readable » et « machine-readable ». L'information est donc accessible et réutilisable !

---

[80] Tom Heath and Christian Bizer (2011) Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, available at : http://linkeddatabook.com/editions/1.0/

[81] À ce propos, voir la pertinente observation d'Alexandre Monnin en ce qui concerne la signification de la lettre « U » dans l'acronyme URI : « L'universel, c'est ce que l'on retrouve dans le « U » d'URI (avant qu'il ne soit remplacé par Uniform [...]) »

[82] Mort de rire, équivalent français de « laughing out loud « (« je ris à gorge déployée ») ou « lots of laughs » (« beaucoup de rires »).

[83] Raccourci qui permet la création de liens entre des pages de site web, au lieu de la re-saisie d'une URL.

En ce qui concerne le monde de la documentation et des sciences de l'information, il faut souligner l'initiative pionnière du thésaurus multilingue de l'UNESCO, dont les principes directeurs ont été établis dans les années 1970, avant la popularité de l'Internet.

Par ailleurs, en 2003 l'UNESCO a adopté la Recommandation sur la promotion et l'usage du multilinguisme et l'accès universel au cyberespace, qui préconise le multilinguisme et aussi un juste équilibre entre les intérêts des titulaires de droits d'auteur et l'intérêt général.

Le répertoire de vedettes-matière de la Bibliothèque du Congrès – LCSH a été aussi converti au multilinguisme. Et la version virtuelle de la classification Dewey – WebDewey – est déjà dans sa deuxième version. TermSciences propose la consultation de terminologies scientifiques multilingues (lexiques, dictionnaires, thésaurus, classifications).

Wikipedia, elle-même, favorise nativement le multilinguisme, à travers des interwikis.

Le Web de données propose un cadre d'interopérabilité pour mettre à disposition, lier et échanger des données structurées en vue d'un traitement simplifié par les machines. Il s'agit de décloisonner les silos de données afin de libérer les usages faits de ces données. Il constitue donc une première étape indispensable pour envisager ensuite des traitements plus complexes sur cette masse de données.

Les langages documentaires sont, donc, en train de s'adapter à l'ère du Web Sémantique. Ils sont des outils primordiaux pour la structuration et modélisation des données, où les référentiels jouent un rôle essentiel dans l'interopérabilité et l'harmonisation entre les modèles, normes et standards adoptés dans la sphère du Web sémantique.

2.6 Folksonomies et tagging comme nouveaux procédés d'indexation

2.6.1 Folksonomie : l'indexation par les utilisateurs

Folksonomie[84], ou indexation personnelle, est un système d'indexation collaborative décentralisée et spontanée, effectuée par des non-spécialistes.

La folksonomie constitue une des fonctionnalités phare du Web 2.0, aussi dit Web Collaboratif. Son principe est très simple : permettre aux utilisateurs de décrire des ressources (billet de blog, page Web, photos, vidéos...) par des mots-clés choisis librement.

Les folksonomies constituent la possibilité pour l'utilisateur d'indexer des documents afin qu'il puisse plus aisément les retrouver grâce à un système de mots-clés. Le concept est complètement lié à l'accroissement et à l'accélération de la production d'informations.

Face à cette croissance de la masse d'information, l'indexation automatique et l'humaine, réalisé par les documentalistes, ne suffisent pas pour indexer l'ensemble des ressources disponibles sur le Web. Les internautes eux-mêmes cherchent, à leur façon, implémenter des outils et astuces pour créer des points de repère dans le Web, travers des liens, de manière à trouver rapidement les ressources souhaitées. Les mots-clés proposés dans les folksonomies correspondent aux besoins des utilisateurs, qui peuvent ainsi arriver plus facilement aux ressources dans lesquelles sont « greffés » un certain nombre de tags[85].

L'internaute indexe et se sent utile à lui-même et aux autres. On parle vraiment ici de Web Collaboratif.

Les utilisateurs jouent, de cette façon, le rôle auparavant réservé aux documentalistes, dont leur rôle donc s'élargit, dans le but de faire l'adaptation des langages documentaires et de l'indexation aux nouvelles technologies du Web.

Le terme folksonomie est une adaptation française de l'anglais « folksonomy », mot-valise combinant les mots « folk » (le peuple, les gens, les utilisateurs) et « taxonomy » (la taxinomie). Le terme a été créé par l'architecte de l'information Thomas Vander Wal[86], qui ainsi la définit :

« Folksonomie est le résultat du tagging individuel et libre d'information et d'objets (n'importe quelle ressource avec une URL) pour sa propre recherche. Le tagging est réalisé dans un environnement social (généralement partagé et ouvert aux autres). La folksonomie est créée à partir de l'acte du tagging par la personne qui consomme l'information. »[87]

---

[84] http://fr.wikipedia.org/wiki/Folksonomie

[85] Ici, en traitant des folksonomies, on se contentera en dire que tag équivaut à mot-clé. Les tags seront traités dans la section suivante.

[86] http://vanderwal.net/folksonomy.html

[87] Traduction française de « Folksonomy is the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval. The tagging is done in a social environment (usually shared and open to others). Folksonomy is created from the act of tagging by the person consuming the information ».

Malgré l'émergence de la folksonomie dans l'ambiant du Web Collaboratif, Alexandre Monnin [19, Monnin] note, d'après la définition ci-dessus, que « l'aspect collaboratif n'y est pas explicitement posé comme une condition nécessaire à la constitution d'une folksonomie. »

Les expressions potonomie, peuplonomie, taxinomie populaire, taxinomie sociale ou encore « social bookmarking » sont aussi employés pour traiter de la folksonomie. Cependant, l'expression française recommandée par la Commission générale de terminologie et de néologie, travers le répertoire FranceTerm, est « indexation personnelle »[88], qui a la définition suivante : « Classification de contenus de l'internet par l'attribution de mots-clés librement choisis par un utilisateur. »

Les internautes, dans l'usage de la folksonomie ne sont pas contraints à une terminologie prédéfinie mais peuvent adopter les termes qu'ils souhaitent pour décrire leurs ressources. Ces termes sont souvent appelés mots-clés ou tags.

La folksonomie, aussi appelée « social bookmarking » est alors l'action de partager des liens entre les différents utilisateurs d'Internet. C'est un répertoire comprenant des signets qui permettent le partage des favoris avec les autres internautes. En se créant son propre dossier, l'usager peut alors enregistrer, sur le réseau, les informations qu'il veut divulguer, y compris ses liens favoris, quelques descriptions, des notes...

2.6.1.1 L'intérêt et usage des folksonomies

L'intérêt des folksonomies est lié à l'effet communautaire : pour une ressource donnée sa description est l'union de l'ensemble des descriptions de cette même ressource qui ont été faites par différents internautes. Ainsi, partant d'une ressource, et suivant de proche en proche les termes (tags) choisis par des autres contributeurs, il est possible d'explorer et de découvrir des ressources connexes et liées.

L'ensemble des mots-clés d'une personne peut être visualisé par des nuages de mots clés. Ce concept permet un survol de l'ensemble des centres d'intérêts d'une personne ou même d'un groupe. Les nuages de tags « tagcloud » sont une sorte de **condensé sémantique** d'un document dans lequel les concepts clés sont dotés d'une unité de taille (dans le sens du poids de la typographie utilisée) permettant de faire ressortir leur importance : les plus gros sont les plus utilisés par la personne/communauté.

---

88 http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000021530619&dateTexte=&categorieLien=id

Figure 9 : Nuage de mots-clés en anglais liés au Web 2.0.[89]

Ces tags fonctionnent comme des marqueurs sémantiques qui balisent cette masse d'information et construisent ainsi des réseaux sémantiques décomposables en réseaux plus restreints comme par exemple celui d'un utilisateur et des relations qu'il entretient avec d'autres personnes.[90]

La folksonomie se présente alors comme un véritable moyen de communication universel, puisque toutes les personnes connectées à l'Internet peuvent entrer en relation et avoir accès aux ressources en ligne.

Sont exemples de site web qui utilisent les folksonomies : del.icio.us et Connotea, comme sites de partage de signets/URLs favoris; Flickr, comme site de partage de photos ; Box, comme site de stockage en ligne et Wikipedia comme encyclopédie collaborative.

---

[89]Source : http://fr.wikipedia.org/wiki/Fichier:Web_2.0_Map.svg

[90] http://www.ieml.org/spip.php?article154

Thomas Vander Wal[91] distingue deux types de folksonomies, les « étroites » (narrow folksonomies) et les « générales » (broad folksonomies). Les folksonomies étroites sont surtout utilisées dans un objectif individuel tandis que les générales privilégient l'aspect collectif et collaboratif du partage d'information. Ainsi les sites de partages de favoris, comme del.icio.us ou Connotea, sont plutôt des folksonomies générales puisqu'un même site peut être partagé par plusieurs utilisateurs et recevoir le même tag. Ce type de folksonomie s'appuyant sur des réseaux sociaux ne fait pas que classer de l'information et la partager. Il met en relation des utilisateurs qui partagent les mêmes centres d'intérêts. **L'utilisateur-indexeur devient** à son tour **un peu indexé** et mis en relation à la fois avec d'autres mots-clés, d'autres sites et d'autres usagers.

Olivier Le Deuff [14, Le Deuff] explicite la différence entre les systèmes de folksonomies et les systèmes de classification :

« Le système diffère donc nettement en cela des systèmes classificatoires classiques, comme la classification décimale universelle ou la classification Dewey, qui s'inscrivent dans des processus plus longs et dont le but est d'obtenir un classement cohérent de documents physiques dont le contenu est inscrit dans la durée. Les folksonomies, au contraire, ne reposent sur aucun thésaurus, ce qui confère à l'utilisateur une liberté totale quant au choix des mots-clés. Les folksonomies sont donc initialement centrées sur l'usager. Ce dernier les utilise dans un but personnel, afin d'organiser son propre système d'information. »

D'après Elie Francis et Odile Quesnel [9, Francis ; Quesnel], on peut encore distinguer deux niveaux d'indexation collaborative, selon la personne de l'indexeur. Le premier est l'indexation par l'auteur. Quand un particulier ou une organisation met en ligne un site web, il choisit lui-même les descripteurs qui caractérisent le contenu de son site pour le référencer sur les sites Web et moteurs de recherche. À leur tour, les internautes qui accèdent ce site peuvent le référencer avec leurs propres termes, qui ne seront probablement pas les mêmes que ceux qui ont étés choisis par l'auteur. Les individus peuvent à l'aide de sites tel comme del.icio.us « classer » leurs sites préférés et les faire partager parmi tous les autres internautes grâce aux folksonomies. Elles se présentent sous forme de nuages de mots, lesquels apparaissent de taille variable : les plus gros étant les plus usités par la communauté. Chacun peut décider d'utiliser les mots proposés par d'autres ou d'en incorporer de nouveaux. Il y a une logique de partage, de consensus.

[91] http://vanderwal.net/random/entrysel.php?blog=1635

Pour Gautier Poupeau[92], la folksonomie présente les avantages suivantes :

- Améliorer la recherche d'information dans sa collection de ressources personnelles ;

- Constituer un vecteur de sérendipité[93] ;

- Donner aux autres utilisateurs une idée du contenu de sa collection de ressources ;

- Faire apparaître des réseaux sociaux implicites par l'utilisation commune de tags entre différents utilisateurs ;

Il faut mettre l'accent sur le plan informatique, vu que toute ressource décrite par les internautes, représente moins de travail en terme d'indexation automatique.

La force des folksonomies réside aussi dans le fait qu'elles ne nécessitent aucun consensus. On ne parle pas alors de vocabulaire contrôlé ou thésaurus !

Nonobstant ces avantages, qui permettent d'accroître et de rendre plus rapide la diffusion de l'information, vu que les utilisateurs sont libres pour choisir leurs propres mots-clés, et aussi d'observer les habitudes des internautes lorsque ceux-ci employent leurs propres termes/tags, la folksonomie révèle aussi des limitations que Gautier Poupeau ainsi liste :

- Impossibilité d'organiser les tags entre eux sous la forme d'une taxinomie ou d'un thésaurus ;

- Deux utilisateurs peuvent partager un tag identique, mais en avoir une conception différente ;

- la folksonomie n'est pas multilingue, c'est à dire qu'il n'existe aujourd'hui pas de moyens pour relier le même tag exprimé en anglais et en français ;

- Deux tags différents peuvent être synonymes voire identique ou presque (penser aux fautes d'orthographe, à l'utilisation du singulier ou du pluriel), sans qu'on puisse les relier.

---

[92] http://www.lespetitescases.net/moat-donner-du-sens-a-vos-tags

[93] De l'anglais « serendipity » : « Pure luck in discovering things you were not looking for », en français l'art ou la faculté de trouver ce que l'on ne cherche pas.

On ajouterait aux limitations des folksonomies la génération de bruit, dans le cas par exemple de tags composés de mots homonymes, lors de la réalisation des requêtes, ce qui ne contribue absolument pas à l'accès à l'information.

À côté de la question du multilinguisme, l'ambiguïté se présente aussi, car un concept est différemment appréhendé selon la culture, la mentalité et le niveau de connaissance de celui qui a taggué une ressource et celui qui va chercher cette même ressource.

En effet, la folksonomie est un **outil de recherche** qui comprend les signets des autres, qui font en sorte que les internautes échangent des informations entre eux. C'est une manière de créer, de partager gratuitement et de rendre public une masse immensurable de ressources.

Il s'agit aussi d'un système qui **change les méthodes de navigation et de recherche**, puisqu'il permet d'accéder à des ressources pertinentes qu'un moteur de recherche aurait pu ignorer.

Il est alors aux documentalistes, à l'aide des technologies du Web sémantique qui sont en train de se mettre en place de faire **interagir** les différentes approches l'indexation, soit-elle professionnelle, industrielle (moteur de recherche) ou collaborative, de manière complémentaire, ayant pour objectif produire une indexation de qualité qui peut attendre les besoins de l'ensemble des utilisateurs.

2.6.2 Le tagging : Qu'est-ce qu'un tag ?

Comme vu précédemment, le « terme-clé », qui revient le plus souvent dans les folksonomies est celui de tag[94]. Le tag peut désigner en fait un mot-clé, une catégorie ou une métadonnée. Le mot anglais tag signifie en français : étiquette de balisage, étiquetage, fléchage, marquage, voire traçage [14, Le Deuff].

Il s'agit d'un mot-clé ou terme associé à une ressource, qui sert à décrire telle ressource. Le tag se présent, donc, comme une métadonnée.

Les tags sont habituellement attribués de façon informelle et personnelle par les utilisateurs/internautes à leurs propres ressources. Le système de tagging ne fait pas partie d'un schéma de contrôle formellement défini. **Ici, il est indispensable prend le tag comme une simple chaîne de caractères, exprimés en langage naturel.**

---

[94] http://fr.wikipedia.org/wiki/Marqueur_(m%C3%A9tadonn%C3%A9e)

Les tags sont typiquement utilisés sur le web dans des taxonomies dynamiques, flexibles, générées automatiquement pour des ressources en ligne comme les fichiers informatiques, les pages web, les images numériques, et des sites de partage de signets.

Chaque tag est présenté sous la forme d'un lien hypertexte[95] pointant vers une page d'index répertoriant toutes les pages qui l'utilisent. Cela permet à un lecteur de trouver rapidement toutes les pages associées avec l'expression « rue Montorgueil » par exemple. En employant la recherche par tag, l'internaute trouvera toutes les pages utilisant un ensemble de tags, tel que « rue Montorgueil » et « Monet », ce qui peut l'amener à la ressource La rue Montorgueil, à Paris. Fête du 30 juin 1878, tableau de Claude Monet qui illustre la fête du 30 juin 1878 dans la rue Montorgueil à Paris.

Voici la définition qu'en donnent Guy Marieke et Emma Tonkin[96] :

« Alors, que sont vraiment les tags ? Une définition simple serait de dire que les tags sont des mots-clés, des catégories de noms, ou des métadonnées. Essentiellement, un tag est simplement un ensemble de mots-clés textuels librement choisi. Cependant, du fait que les tags ne sont pas créés par des spécialistes de l'information, ils ne suivent à l'heure actuelle aucune directive formelle omniprésente. Cela signifie que les ressources peuvent être catégorisées avec n'importe quel mot définissant une relation entre la ressource en ligne et un concept issu de l'esprit de l'utilisateur. Le nombre de mots qui peut être choisi est indéfini, dont certains sont des représentations évidentes et d'autres ont peu de **sens** en dehors du contexte de l'auteur du tag. »[97]

Le tag peut alors prendre toutes les formes – et les sens - possibles, selon le désir de l'internaute et surtout selon sa culture et sa maîtrise de la langue. Comme le système de tagging ne s'appuie sur aucun thésaurus, des mots absents du dictionnaire ou des néologismes peuvent devenir des tags.

---

[95] Référence dans un système hypertexte permettant de passer automatiquement d'une ressource en consultation à une autre ressource liée

[96] http://www.dlib.org/dlib/january06/guy/01guy.html

[97] Traduction de « So what exactly are tags? A simple definition would be to say that tags are keywords, category names, or metadata. In essence, a tag is simply a freely chosen set of textual keywords. However, because tags are not created by information specialists, they do not at present follow any ubiquitous formal guidelines. This means that items can be categorised with any word that defines a relationship between the online resource and a concept in the user's mind. Any number of words might be chosen, some of which are obvious representations, others making less sense outside the tag author's context. »

Alexandre Monnin [19, Monnin] préfère ne pas définir un tag : « […] il ne faut pas chercher de définition canonique précise du tag ». Il conçoit le tag comme une opération qui se repose sur un ensemble de trois axes :

- le tag lui-même ;

- la ressource ;

- l'utilisateur.

Le philosophe affirme encore qu'un tag est une entité bipartite, formée à la fois d'un libellé et d'une étiquette :

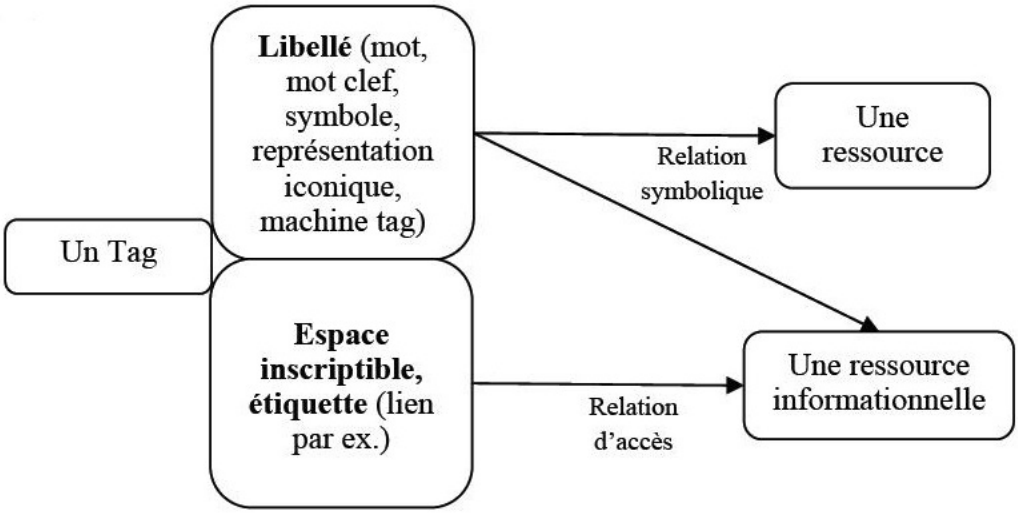

Figure 10 : La bipartition du tag, d'après Alexandre Monnin[98]

« Le libellé n'est rien d'autre, quant à lui, que la suite de caractères inscrite à même le tag, lui-même conçu à la manière d'une étiquette. En ménageant un accès à la ressource (informationnelle ou non), cette étiquette permet à l'utilisateur d'associer à celle-ci le texte qu'il désire. Il devient dès lors loisible d'indexer, d'évaluer, de partager ou encore de retrouver des objets qui échappaient jusqu'alors à ces possibilités d'annotations. Précisément ce que permettait depuis longtemps, dans l'univers analogique, le traditionnel post-it : produire une surface matérielle accueillant du texte là où celle-ci faisait défaut. »

Le libellé d'un tag est donc un espace vide inscriptible susceptible d'accueillir les entités symboliques voire informatiques de son choix. « […] Les libellés inscrits sur les tags sont susceptibles d'accueillir des entités contrastées, linguistiques ou non, interdisant *de*

[98] Source :http://ic2009.inria.fr/docs/papers/Monnin_IC2009_41.pdf

facto une intelligence globale de la sémantique sous-jacente à leur usage. » [19, Monnin, p. 10].

Si l'on considère que les libellés sont des espaces vides, « il convient de tenir les libellés des tags non pour des termes, des sujets, voire de simples mots (clefs ou non) mais tout cela à la fois – parmi une kyrielle d'autres choses encore. » [19, Monnin, p. 9]

Alexandre Monnin propose encore une distinction entre les concepts d'autres chaînes de caractères tels comme mot-clé, descripteur et vedette-matière, qui se mêlent lors que l'on parle tag, et qui peut induire à réduire le rôle d'un tag à un simple libellé/ chaîne de caractères.

Mot-clé[99] est le terme choisi dans le titre ou le texte d'une ressource, sans référence à un lexique ou à un thésaurus, pour caractériser son contenu et permettre la recherche de cette ressource. Bien qu'issu du langage naturel, le mot-clé se conçoit la plupart du temps comme extrait directement d'un document analysé. La différence entre mot-clé et tag s'aperçoit nette : « Tagger c'est aussi ajouter un contenu absent d'un document (ou d'une ressource) ; en d'autres termes, lui adjoindre un contenu extrinsèque. »

Descripteur[100] est le terme ou groupe de termes retenus dans un thésaurus ou lexique de référence et choisi parmi un ensemble de termes équivalents pour représenter sans ambiguïté une notion contenue dans un document ou dans une demande de recherche documentaire. Il s'agit donc d'un terme normalisé. « Un descripteur n'existe par conséquent, qu'imbriqué dans un vocabulaire contrôlé définissant avec précision la place revenant à chaque terme. » [19, Monnin]

Vedette-matière, traduction de l'anglais « subject heading » est intrinsèquement lié au sujet d'une ressource donnée. « Un sujet n'est pas un simple mot. Entité lexicale complexe extraite d'un langage documentaire artificiel, et non du langage naturel, il s'agit avant tout d'un syntagme résultant de la coordination de plusieurs descripteurs » [19, Monnin]. Elle est le point d'accès qui représente synthétiquement le sujet ou l'un des sujets d'une ressource.

---

[99] http://www.adbs.fr/mot-cle-17878.htm?RH=OUTILS_VOC

[100] http://www.adbs.fr/descripteur-16756.htm?RH=OUTILS_VOC

2.6.2.1 Les faiblesses du tagging tout court

En dépit de l'utilisation de tags en tant que système posséder des avantages, tels comme la liberté donnée l'utilisateur d'indexer leur propres ressources, ce système relève une absence de distinction sémantique, car, comment déjà vu, un même tag peut avoir différents sens ou significations, selon l'approche faite par l'utilisateur. L'attribution des tags est ainsi très subjective, ce qui peut entraîner conflits lors de la mise en relation entre ressources qui ont été taggés/décrites de façon différente par les utilisateurs. Cette des balises peut conduire à des connexions inappropriées entre les éléments.

Une des solutions pour améliorer le système serait de former les utilisateurs à l'indexation par tag. Ce c'est qui propose Olivier Le Deuff [14, Le Deuff]: à l'indexation ou la « tag literacy ».

Pour cela, il faudrait aussi que les sites permettent l'usage de plusieurs tags pour définir un une ressource. Il faudrait supprimer les mauvais tags, c'est-à-dire, les tags mal indexés, ainsi qu'établir une liste de tags inefficaces : les tags mal orthographiés, les tags mal conçus, notamment les groupes de tags collés ensemble, les tags personnels n'ayant aucun intérêt collectif ou encore un tag unique qui n'apparaît qu'une seule fois dans un répertoire donné.

On peut citer certaines règles de bonne indexation par tags :

- l'utilisateur doit penser collectivement : les tags sont certes personnels mais peuvent également être utilisés par d'autres ;

- employer le pluriel pour définir des catégories. Le pluriel est plus approprié car la catégorie peut contenir différentes variations ;

- ne pas employer de majuscules, à moins que le mot ne puisse être compris sans son emploi;

- inclure des synonymes afin d'éviter les confusions ;

- observer et utiliser les conventions d'indexation des sites et des réseaux sociaux utilisés ;

- contribuer à ce que les efforts d'indexation soient efficaces en collaborant et en ajoutant des tags à d'autres ressources.

Certes, cet ensemble de règles est utile. Toutefois, contribuer à une indexation efficace implique non seulement des compétences par les utilisateurs, mais aussi du temps

et de l'argent. D'après Ian Davis[101], l'indexation par les utilisateurs semble moins coûteuse en temps, mais c'est le temps passé à retrouver l'information qui s'accroît, contrairement aux systèmes d'informations hiérarchisées, où le coût d'indexation par les professionnels est plus élevé, mais la recherche d'information reste facilitée pour l'utilisateur.

En plus, comme souligne Alexandre Monnin, seul l'utilisateur peut établir avec certitude à quoi renvoient les libellés de chacun de ses tags, ce qui s'aggrave par le fait de qu'un ensemble de tags peut servir à décrire des entités très différentes : une page, un lien, une partie d'un texte, une action, un jugement, une relation, une personne, un lieu, un événement, etc. Cette indétermination cadre tout à fait avec l'absence de sémantique propre à la partie libellé des tags. « Elle n'en exige pas moins un attachement scrupuleux au contexte, seul à même de rendre compte de l'utilisation singulière d'un tag car, si des relations sont ipso facto établies, elles n'en sont pas explicitées pour autant. »

2.7 Rendre exploitable le potentiel des tags découlant des folksonomies : le tagging sémantique !

S'il était possible de traduire les tags exprimés en langage naturel sous forme de coordonnées uniques, les machines pourraient s'affranchir des problèmes liés au langage naturel et nous fournir automatiquement des informations comme par exemple l'étendue d'un réseau sémantique ou tout simplement vérifier par qui et selon quelles nuances l'interprétation d'un document est partagée ou pas. L'enjeu est d'organiser la pratique du tagging pour lui donner une dimension sémantique plus calculable et interopérable.

Certes, les folksonomies ont permis d'élargir le périmètre occupé par les ressources pertinentes auxquelles on peut accéder. Elles ont également permis la croissance des échanges d'information entre les utilisateurs. Cependant, ces ressources-là sont en vrac, dispersées sur le Web. Le défi qui s'impose est de permettre d'organiser collectivement les ressources et de les partager, dans un environnement multilingue où elles ne sont pas facilement interopérables.

Alexandre Monnin [19, Monnin, p. 12], en insistant sur la variabilité des inscriptions (libellés) qu'un tag (étiquette) est susceptible de recevoir, a montré que les libellés n'étaient pas pourvus d'une sémantique fixe. « Aussi n'est-il pas possible de partir du postulat qu'un tag réfère toujours à un concept en vertu de sa sémantique dénotationnelle. »

C'est donc pour cela, que le philosophe fait appel aux apports des technologies du Web sémantique.

---

[101] http://tagsonomy.com/index.php/ian-davis-on-why-tagging-is-expensive/

Un tag, partie essentielle et composante des folksonomies, « n'est finalement qu'une chaîne de caractères dont le sens exact est connu du seul « taggueur » qu'un autre utilisateur peut éventuellement appréhender, mais en aucun cas une machine qui se repose uniquement sur la morphologie du tag pour l'exploiter. »[102]

Gautier Poupeau est donc d'accord avec Alexandre Monnin en affirmant que la solution à ces problèmes réside en partie dans les technologies du Web sémantique qui vise précisément à donner du sens aux tags.

Sémantiser un tag est d'abord choisir un référentiel à partir duquel va être associée à ce tag un URI. Une description de la notion du tag encodée en RDF est par suite associée à cet URI. Ainsi, le tag n'est plus seulement une chaîne de caractères, mais possède un véritable sens donnée par l'ensemble des triplets RDF qui décrit la notion utilisée dans le tag.

Par exemple, en choisissant comme référentiel DBpedia (Wikipedia structuré en format RDF) dans sa version anglaise, pour le tag « Topinambour », sera associé à l'URI « http://dbpedia.org/page/Jerusalem_artichoke » qui correspond à la description de la notion de Web sémantique dans DBpedia. Sont ainsi résolus, entre autres, les problèmes liés à l'orthographe de la notion et à la langue, puisque les tags « Helianthus tuberosus », « tupinambo », ou « patata de Judea » sont reliés à la même notion via l'URI adopté.

C'est cela qui explique Bernard Vatan[103] : « Au lieu de considérer le marquage [tagging] comme la pose d'un mot-clé sur une ressource, avec tous les problèmes connus que cela comporte (homonymie, ambiguïté, synonymie, langues multiples), on considère que derrière le tag (le libellé en langage naturel, une chaîne de caractères) se cache un concept qui lui peut être défini de façon formelle dans l'univers RDF, typiquement défini par une URI décrite dans un vocabulaire contrôlé. Un tag sémantique sera donc un objet de type relation (de classe Tag avec majuscule) à trois pattes dont l'une est le « tag minuscule », c'est-à-dire le libellé, la deuxième la ressource marquée, et la troisième l'URI du concept. »

Gautier Poupeau propose le schéma suivant pour illustrer le processus de tagging sémantique, en utilisant comme référentiel un thésaurus :

---

[102] http://www.lespetitescases.net/moat-donner-du-sens-a-vos-tags

[103] http://mondeca.wordpress.com/2009/06/12/common-tag-standard-les-tags-la-semantique/

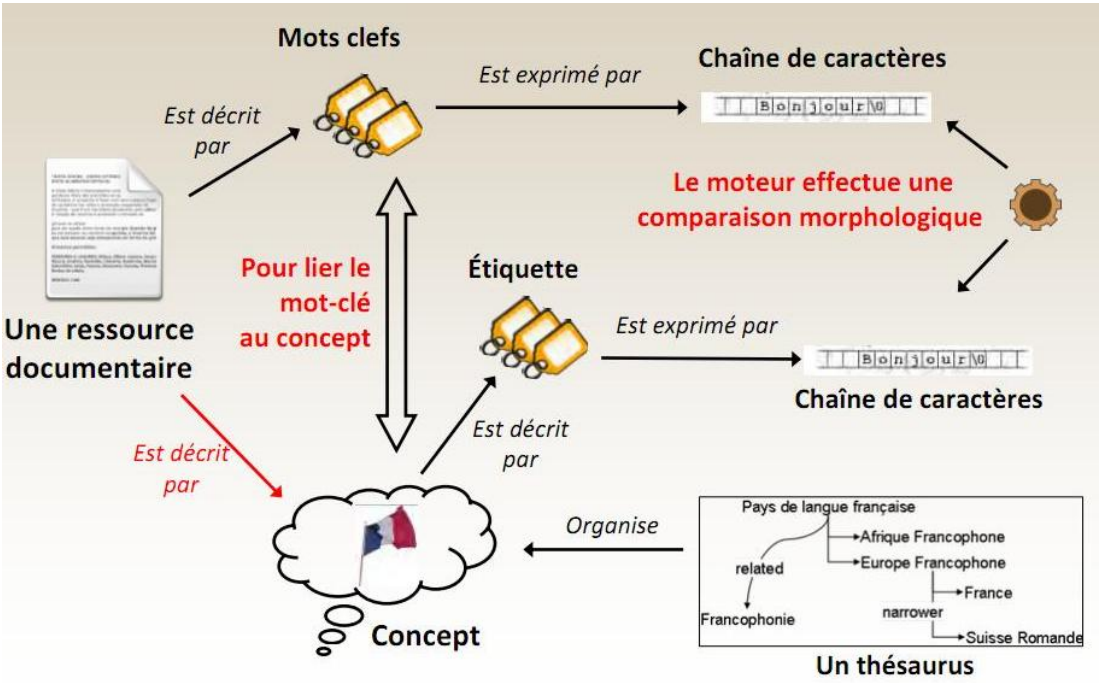

Figure 11 : Le processus de tagging sémantique.[104]

Alexandre Passant [20, Passant, p. 124-125] énumère des outils qui combinent le système de tags et les technologies du Web Sémantique :

Annotea : une des premières applications sociales de partage de contenus basée sur les technologies du Web Sémantique, cet outil propose dès 2001 un système d'annotations et de partage de ressources Web ouvert et reposant sur les technologies du Web Sémantique. Cet outil permet à chaque communauté de disposer de son propre serveur d'annotations, les différentes annotations produites étant ensuite disponibles en RDF utilisant un modèle particulier d'annotations combiné à un vocabulaire « bookmark » ;

Revyu : site web entièrement basé sur les standards du Web Sémantique, vu que l'ensemble des annotations produites au sein de cet outil est en outre disponible en RDF. Les utilisateurs peuvent évaluer et noter tout ce qui se peut nommer. Cet outil repose sur des heuristiques permettant de lier automatiquement les revues à des ressources déjà existantes, par exemple des livres en vente sur Amazon.com auxquels un URI propre a été assigné.

- Faviki : site web de partage de tags propose un service de gestion de favoris où les tags sont des identifiants DBpedia. **Les concepts de Wikipédia deviennent donc des**

[104]Source : http://www.bnf.fr/documents/afnor2011_poupeau.pdf

concepts sémantisés. Il prend ainsi en compte la notion de multilinguisme associée aux tags, puisqu'un même URI peut être associé à plusieurs termes.

Encore selon Passant [20, Passant, p. 125], « d'autres outils sont axés plus spécifiquement sur la gestion des tags, et plus particulièrement sur la manière de les organiser pour pallier à leurs limites. Ainsi, les outils de « bookmarking » SemanticScuttle, Gnizr et Semanlink permettent de définir des relations hiérarchiques entre tags, le second offrant un export RDF des contenus annotés en utilisant certaines des ontologies [...] (notamment la angTag Ontology, SIOC et SKOS), le dernier étant basé sur son propre modèle de représentation des tags reposant sur [l'ontologie] SKOS. Dans une approche différente, GroupMe propose aux utilisateurs de regrouper les tags par catégories pour faciliter la recherche d'information, représentant le tout avec sa propre ontologie. Sweetwiki permet également l'organisation de tags (et utilise son propre modèle de représentation), cette fois-ci au sein d'un wiki. L'approche reste fidèle à la philosophie wiki en permettant à tous les utilisateurs du système de gérer cette organisation commune de l'ensemble des tags du wiki de manière ouverte et collaborative.

Une fois de plus, Passant cite d'autres exemples d'outils de traitement de tags. Il ajoute davantage :

« S'ils n'utilisent pas explicitement les technologies du Web Sémantique, d'autres outils permettent manuellement de structurer ou d'enrichir les systèmes à base de tags et de bénéficier de ces enrichissements au moment de la recherche d'information. Ainsi, toujours dans une approche permettant de dériver des relations taxonomiques à partir de folksonomies, Bibsonomy (outil collaboratif de gestion de références bibliographiques […]) propose aux utilisateurs de définir eux-mêmes des relations hiérarchiques entre tags. » Avec Extreme Tagging, les utilisateurs ont la possibilité de typer les tags et les relations entre ces tags. Cet outil permet donc de structurer les folksonomies. « Par exemple, on va pouvoir taguer le tag « apple » par fruit dans une action de tagging, et par « mac » dans une autre. Si l'idée est intéressante, l'utilisation de simples tags pour définir ces types nous semble conduire aux mêmes problèmes que ceux qu'elle souhaite résoudre. » Et on revient encore au but initial : comment arriver de façon univoque/rapide à la ressource précise/souhaitée ?

À propos des utilisations multiples que l'on peut faire à partir de la sémantisation des tags, Bernard Vatant[105], en considérant « que les trois pattes du Tag [le tag lui-même, l'utilisateur et la ressource, d'après Monnin] peuvent être utilisées deux par deux et dans tous les sens », énumère les suivantes :

[105] http://mondeca.wordpress.com/2009/06/12/common-tag-standard-les-tags-la-semantique/

1- Attacher un tag (libellé) à une ressource comme dans le tagging ordinaire ;

2- Rechercher les ressources marquées par un libellé ;

3- **Attacher un concept contrôlé à une ressource (indexation sémantique)** ;

4- Retrouver toutes les ressources attachées à un concept (recherche par métadonnées contrôlées) ;

5- Retrouver tous les libellés utilisés pour un même concept, dans toutes les langues ou dans une langue donnée ;

6- Retrouver tous les concepts attachés à un libellé ;

Pour Vatant, « les utilisations 5 et 6 sont des effets collatéraux du standard qui ne sont pas des moins intéressants. Si le marquage sémantique se fait dans le contexte de la ressource marquée, il est de fait aussi un enrichissement du concept par des libellés. On marque donc le concept en même temps que la ressource, et tout le monde y gagne. »

Finalement, il ajout que « pour l'instant les outils d'aide au marquage comme Zemanta[106] ne fonctionnent qu'en anglais, mais le standard fournit bel et bien l'infrastructure pour l'intégration du marquage multilingue. »

2.8 Du tagging sémantique à l'indexation sémantique !

Passer d'un *modus operandi* classique de tagging à une indexation sémantique qui s'appuie dans un référentiel n'est pas forcément une procédure simple.

L'indexation sémantique permet donc d'indexer les ressources non plus par de simples termes, exprimés via les tags (simples chaînes de caractères), mais par des concepts, formatés en RDF et exprimés via URIs. On passe ainsi d'une indexation par mot-clé à une indexation par un concept issu d'un référentiel préalablement choisi.

Elle est capable aussi de résoudre les différents problèmes posés par les systèmes à base de tags, comme vu précédemment (renvoi), en représentant sans ambiguïté et de manière interprétable par les machines les significations associées aux tags alors

[106] Outil qui propose aux bloggers et autres créateurs de contenus des images, des tags ainsi que des liens vers des articles ayant un rapport avec le sujet des textes qu'ils écrivent. La compréhension du contenu par Zemanta peut être intégré à Common Tag pour des tags sémantiques, de façon à relier et normaliser les tags.

sémantisés. Ce nouveau processus ouvre donc de nouvelles perspectives pour la recherche d'information.

Ainsi, **l'indexation sémantique consiste, lors de l'analyse de la ressource, à rattacher chaque ressource à un concept sous-jacent, issu d'un référentiel**. Ainsi, par exemple, pour le mot cougar, il faut déterminer s'il s'agit du félin, de la voiture américaine ou de la femme mûre fréquentant des hommes plus jeunes qu'elle.[107] Ce processus n'est pas du tout simple, vu qu'il faut lister tous les concepts (les plusieurs sens du mot) qui s'expriment à travers un mot donné, de façon à livrer le concept le plus adéquat au mot, en le transformant en un véritable concept qui peut être appréhendé de manière univoque par les utilisateurs. On passe alors par la modélisation des concepts, en touchant toutes les rapports entre les sens des mots, tels comme les relations d'homonymie, de synonymie, d'antonymie, de polysémie, d'hyperonymie, d'hyponymie, de paronymie etc, tout en passant par les soucis liés à la lemmatisation, l'ambiguïté, aux faux-amis... Ici la sémantique[108] prend son sens premier, comme la branche de la linguistique qui étudie les signifiés des mots, au contraire du sens attribué à sémantique dans l'expression web sémantique, où il faut expliciter manuellement le sens des informations afin que les machines puissent les exploiter de façon automatique, sans ambiguïté et à grande échelle.

C'est ce processus que nous essaierons d'expliciter, par le biais d'une approche pratique, dans la deuxième partie de ce mémoire, consacrée à la réalisation d'une synthèse sur l'opération de reprise de l'existant de tags, dans le but de leur sémantisation, compris dans le corpus de ressources du portail Histoire des Arts.

---

[107] http://fr.wikipedia.org/wiki/Cougar_(femme)

[108] Étude d'une langue ou des langues considérées du point de vue de la signification; théorie tentant de rendre compte des structures et des phénomènes de la signification dans une langue ou dans le langage. http://www.cnrtl.fr/definition/s%C3%A9mantique

Deuxième partie

Construire un corpus sémantisé : le cas d'histoiredesarts.culture.fr

3 Contexte du projet, le portail Histoire des Arts

Depuis 2009, le programme histoire des arts est un enseignement obligatoire à l'école, au collège et au lycée, conforme l'arrêté du 28 août 2008[109] qui trace les directives pour l'organisation de l'enseignement de l'histoire des arts à tous les niveaux de la scolarité.

Cet enseignement, inscrit dans les différents programmes disciplinaires, a pour ambition de transmettre à chaque élève une culture artistique commune fondée sur des références précises, diversifiées et inscrites dans leur dimension historique.[110]

Toutes les disciplines contribuent à l'acquisition des compétences et des connaissances du socle commun mises en œuvre dans l'enseignement de l'histoire des arts, dont notamment :

- le développement de la capacité à analyser une œuvre d'art ;

- la construction d'une culture personnelle ;

- la maîtrise de l'expression orale ;

- l'épanouissement de la curiosité et de la créativité artistiques des élèves ;

- la découverte des métiers et des formations liés à ces pratiques artistiques et culturelles.[111]

« L'enseignement de l'histoire des arts est un enseignement de culture artistique partagée. Il concerne tous les élèves. Il est porté par tous les enseignants. Il convoque tous les arts. Son objectif est de donner à chacun une conscience commune : celle d'appartenir à l'histoire des cultures et des civilisations, à l'histoire du monde. Cette histoire du monde s'inscrit dans des traces indiscutables : les oeuvres d'art de l'humanité. L'enseignement de l'histoire des arts est là pour en donner les clés, en révéler le sens, la beauté, la diversité et l'universalité. »

---

[109] http://www.education.gouv.fr/cid22078/mene0817383a.html

[110] http://www.education.gouv.fr/pid25535/bulletin_officiel.html?cid_bo=58238

[111] Idem.

Le programme histoire des arts est donc un enseignement fondé sur une approche **pluridisciplinaire** et transversale des oeuvres d'art, dont la transmission du savoir implique la conjonction de plusieurs champs de connaissances.

C'est dans le but de faciliter le travail de l'ensemble des enseignants qu'a été conçu le portail Histoire des Arts[112], hébergé, produit et coordonné par le Ministère de la Culture et de la Communication, à travers le Département des programmes numériques - DPN et qui contribue au nouveau Portail interministériel pour l'éducation artistique et culturelle.

Le portail Histoire des Arts, en ligne depuis septembre 2010, est destiné à la communauté éducative : enseignants, documentalistes, animateurs... Pour tous les domaines artistiques et culturels, il signale des documents en ligne présentant une oeuvre ou un groupe d'oeuvres. Cet annuaire recense des milliers d'oeuvres d'art en ligne, classées selon le programme d'enseignement d'Histoire des arts de l'Éducation Nationale. Ces ressources Web sont élaborées par les services compétents des établissements publics culturels nationaux, sous tutelle du Ministère de la Culture et de la Communication, ou de ses partenaires institutionnels qui les conservent et les mettent en valeur, diffusant des analyses d'œuvres, des dossiers thématiques, des expositions virtuelles, des ressources iconographiques...

Cet annuaire s'organise ainsi en fonction du nouvel enseignement de l'histoire des arts : grands domaines artistiques, périodes historiques, thématiques. Il est ainsi conçu pour fournir un ensemble de données pour chacune des rubriques du programme de l'enseignement.

Histoire des Arts permet d'accéder directement à la ressource dont l'internaute a besoin sur le site de l'institution productrice de celle-ci. C'est un gain de temps précieux, en même temps qu'une indication sur les richesses à explorer éventuellement sur ce site.

Il contient :

- Plus de 4500 ressources commentées (tableaux, sculptures, films) accessibles en ligne ;

- Une recherche multicritère (mots-clés, périodes historiques, domaines artistiques, thématiques, villes, institutions, format informatique) ;

---

[112] Pour une présentation textuelle, voir le billet écrit par le chef de projet utilisateur Bertrand Sajus :http://cblog.culture.fr/2011/01/31/histoiredesarts-culture-fr. Pour une présentation visuelle, partagée avec Marion Martin Laprade, voir la vidéo : http://www.dailymotion.com/video/xe9i6l_annuaire-histoire-des-arts_news#rel-page-4

- Une recherche par carte cliquable (régions et départements) ;

- Des repères chronologiques par domaines artistiques (peinture, musique, architecture, cinéma) ;

- Des rubriques d'actualités ;

- Des accès[113] par widgets (Netvibes, iGoogle, Wordpress), flux RSS, OpenSearch, qui permet d'ajouter d'un clic l'annuaire Histoire des Arts dans la barre de recherche du navigateur de l'utilisateur.

Les ressources référencées doivent, donc, d'une part, être disponibles en ligne, scientifiquement validées, intelligibles et pertinentes au regard de l'organisation de l'enseignement de l'histoire des arts (selon l'arrêté du 28 août 2008), et d'autre part, comporter une valeur pédagogique pour tout enseignant souhaitant enrichir sa documentation.

La sélection et la description des ressources s'effectuent de concert entre les sites éditeurs notamment leurs services éducatifs et l'équipe éditoriale de l'annuaire d'histoire des arts, qui est aussi chargée de mettre à jour régulièrement les liens inactifs du site internet et enrichit ses contenus au fil du temps.

Le corpus documentaire d'Histoire des Arts, destiné à un large public éducatif, est au cœur d'un réseau collaboratif de plusieurs centaines d'institutions culturelles et constitue, ainsi, un échantillon représentatif, à l'échelle nationale, des ressources culturelles numériques en ligne.

[113] http://www.histoiredesarts.culture.fr/outils

4 Présentation du projet HDABO

4.1 Présentation générale du projet HDABO

L'objectif du projet HDABO est, premièrement de doter histoiredesarts.culture.fr d'un back office ouvert aux contributeurs institutionnels du site.

La sémantisation des tags contenus/utilisés dans le corpus de ressources du portail Histoire des Arts, réalisé dans un premier module, est inscrite dans un projet de création d'un « back-office » Histoire des Arts – HDABO, conçu par Bertrand Sajus, chef de projet utilisateur de Histoire des Arts.

Il s'agit d'une plate-forme de gestion des notices à distance qui vise à substituer l'utilisation du logiciel de gestion de bases de données FileMakerPro. Ce back-office est ouvert approximadament à 400 institutions contributrices qui peuvent créer, mettre à jour et supprimer leurs propres notices via une interface Web et garantit la robustesse informatique de l'outil de gestion. Son utilisation évitera aussi que l'équipe éditoriale de Histoire des Arts - HDA soit obligée de ressaisir toutes les notices issues des institutions partenaires, ayant aussi comme but l'augmentation de la réactivité et la pérennité de l'annuaire, de manière à renforcer les liens de collaboration entre l'administration centrale, représentée par le Département des Programmes Numériques du Ministère de la Culture et de la Communication - DPN-MCC et les institutions partenaires.

Compris dans HDABO, on trouve une fonctionnalité/outil informatique pour l'indexation des ressources, dans le cadre du Web Sémantique, qui se montre beaucoup plus élaborée que l'indexation libre, avec l'extraction, après l'analyse documentaire de la ressource, de mots-clés librement choisis par l'indexeur, utilisée auparavant, tout en enrichissant les modalités de recherche.

Bertrand Sajus et l'équipe de chercheurs de l'Institut de Recherche et Innovation du Centre Pompidou - IRI ont constaté **l'importance et l'opportunité que représente le réseau sémantique de plus d'un million de termes définis dans la version française de Wikipédia pour décrire des ressources culturelles.** Des tags sémantisés, liés aux entrées de Wikipédia, seront ainsi générés via la plate-forme HDABO. Les acteurs extérieurs, c'est-à-dire, les institutions partenaires pourront, en utilisant une interface de programmation - API, accéder aux métadonnées et montrer les potentialités de navigation et de découverte qu'offre l'approche web sémantique. Cette API sera faite, à un deuxième moment, à partir des données de HDA, sur un site indépendant, qui sera objet d'une preuve de concept à être réalisé dans le cours de 2012.

4.1.1 Le choix du référentiel Wikipédia/DBpedia

4.1.1.1 Présentation de Wikipédia

Wikipédia est une encyclopédie collective établie sur Internet, universelle, multilingue et fonctionnant sur le principe du wiki. Wikipédia a pour objectif d'offrir un contenu libre, objectif et vérifiable que chacun peut modifier et améliorer.

Le nom Wikipédia est un mot-valise formé à partir de « wiki », terme désignant un type de site web, et « encyclopedia », mot anglais pour encyclopédie. Un wiki, de l'hawaiien « wiki wiki » signifiant « rapide », est un site web dont les pages peuvent être modifiées facilement et sans connaissances techniques préalables à l'aide d'un navigateur web.

Fondée le 23 mars 2001, Wikipédia en français est l'une des nombreuses déclinaisons internationales de Wikipédia. En novembre 2011 elle contient 1.178.000 articles et compte plus de 5.000 contributeurs actifs.

Contrairement à une encyclopédie traditionnelle, Wikipédia est librement distribuable : elle est non seulement disponible gratuitement sur Internet, mais peut aussi être copiée et utilisée librement, avec mention de la source et des auteurs. La réutilisation, même commerciale, du contenu est possible grâce à la licence Creative Commons.

D'après Laure Endrizzi [8, Endrizzi, p. 8], « la coordination entre les langues s'avère plus ou moins structurée selon les versions [...] et vise à promouvoir l'établissement de renvois multilingues entre les articles et la traduction d'articles de qualité. Par exemple, à partir de l'article New York de la Wikipedia française, on peut accéder à l'article éponyme dans une vingtaine de langues. Les liens entre les articles sont appelés 'interlanguage links' ou 'liens interlangues' et répondent à une syntaxe propre à la gestion des liens 'InterWiki' utilisés entre les différents projets de la fondation Wikimedia. »

Ceci dit, les données et métadonnées de Wikipédia constituent un terrain déjà fertilisé pour l'application de toutes les technologies du Web Sémantique, spécialement celles liées au multilinguisme.

4.1.1.2 Wikipédia comme référentiel pour Histoire des Arts

Étant donné son fort contenu culturel, le choix d'adopter Wikipédia comme référentiel pour le tagging sémantique semble évident.

Wikipédia a déjà un rôle inédit dans l'histoire du savoir: la constitution d'une encyclopédie où la liberté, le partage, la collaboration sont indispensables, tout en rendant l'information de plus en plus accessible.

La représentation du site du Ministère de la Culture et de la Communication est particulièrement forte : il est en quatrième position dans le classement des 1000 sites les plus cités sur la version française de Wikipédia[114], avant même le site de l'INSEE[115], des sites de la presse et Google !

La Wikipédia française contenait 64.046 liens vers le site du Ministère de la Culture en mai 2011. D'où le constat que Wikipédia est un gros utilisateur des bases de données du Ministère, comme Mérimée, Palissy, Joconde... Et prochainement Histoire des Arts ?

Ce n'était pas hyperbolique l'ouverture du discours du neuroscientiste hongrois Szilveszter E. Vizi, lors de la conférence de Dix ans de Wikipédia en début 2011 :

« Wikipédia est la mémoire partagée de l'humanité. »[116]

De plus, le modèle d'affichage de résultats créé par Google met très souvent en tête les articles de Wikipédia. Si l'on fait une recherche, en utilisant le moteur de recherche Google, sur Pablo Picasso, le premier résultat qui apparaît dans la liste de résultats est l'article sur Pablo Picasso dans Wikipédia. Ce premier résultat, d'ailleurs montre déjà quelques informations qui sont structurées, tel le début du texte de l'article, les principaux thèmes (liste d'oeuvres, cubisme,…). Il s'agit donc d'un affichage encyclopédique, qui renvoie ces thèmes à d'autres articles de Wikipédia.

4.1.1.3 Le miroir sémantique de Wikipédia[117] : DBpedia

DBpedia est un projet d'extraction de données de Wikipédia dans sa version anglaise pour en proposer une version web sémantique structurée. Ce projet est mené par l'Université de Leipzig, l'Université Libre de Berlin et l'entreprise OpenLink Software.

---

[114] http://fr.wikipedia.org/wiki/Utilisateur:Emijrp/External_Links_Ranking

[115] Institut national de la statistique et des études économiques

[116] En anglais : "Wikipedia is the common memory of mankind." http://meta.wikimedia.org/wiki/Quotes

[117] « In a nutshell, DBpedia is the Semantic Web mirror of Wikipédia » http://blog.dbpedia.org/2011/07/09/official-dbpedia-live-release/

DBpedia est donc une initiative tripartite pour extraire des informations structurées de Wikipedia et rendre cette information disponible sur le Web. DBpedia permet de poser des requêtes complexes à partir de Wikipédia et de lier d'autres jeux de données disponibles sur le Web aux données de Wikipédia.

Vu la quantité incroyable d'informations dans Wikipédia qui peuvent être utilisées de façons nouvelles et intéressantes, le projet DBpedia permet aussi d'inspirer des nouveaux mécanismes pour la navigation, des liens et d'améliorer l'encyclopédie elle-même.

Les metadonnées de DBpedia sont inscrites dans le projet Linked Open Data. Dbpedia génère des URIs pour des millions de concepts. D'autres référentiels se mettent en contact et établissent des liens RDF à partir des jeux de données de DBpedia, faisant de DBpedia le « hub » central d'interconnexion[118] du Web de Données.

En septembre 2011, la dernière version de DBpedia, DBpedia 3.7, possède 15 éditions, c'est-à-dire, 15 langues distinctes, qui contiennent des données de tous les pages de Wikipédia dans une langue spécifique. Les URIs et tous les jeux de données appartenants à chaque édition sont attribués directement à partir du titre non- anglais de l'article de Wikipédia, ce qui signifie, par exemple, qu'il existe 16 différents URIs pour se reporter à la ville de Paris. Les liens inter-langues de Wikipédia sont aussi extraits à partir de différentes versions de Wkipédia.

Cependant, si un article de Wikipédia ne possède pas une version anglaise, les données ne peuvent pas être extraites. Ce sont les cas des articles sur Les Frères Jacques, Yvette Horner, Louis Mazetier, le burin... D'où l'importance d'une version française de DBpedia, destinée à extraire les données à partir de la version française de Wikipédia et non de sa version anglaise, de façon à garder l'ensemble des articles, sans aucune perte de données.[119]

C'est d'après la démonstration du potentiel de l'utilisation des données de Wikipédia par le projet HDABO qu'une version française de DBpedia sera élaborée en 2012.[120]

---

[118] « central interlinking-hub »

[119] Pour savoir plus sur d'autres faiblesses de DBpedia : http://www.slideshare.net/AntidotNet/change-et-interoprabilit-des-donnes-structures-sur-le-web?from=ss_embed à partir du slide 13.

[120] http://www.culturecommunication.gouv.fr/Espace-Presse/Dossiers-de-presse/Budget-2012-du-ministere-de-la-Culture-et-de-la-Communication

4.1.2 Acteurs de HDABO

Les acteurs du projet de module de tagging sémantique sont :

- le Département de Programmes Numériques – DPN du Ministère de la Culture et de la Communication – MCC : maîtrise d'ouvrage et éditeur de Histoire des Arts ;

- IRI[121] (Institut de Recherche et d'Innovation du Centre Pompidou) : développeur de fonctionnalités de tagging sémantique mises à disposition du projet HDABO.

4.1.3 Les fonctionnalités de la plate-forme HDABO

La description des fonctionnalités, prévues en septembre 2011, reprend le cahier des charges conçu et rédigé par Bertrand Sajus.

La plate-forme de gestion des notices de l'annuaire Histoire des Arts permet :

- Gestion des accès aux plusieurs centaines de contributeurs institutionnels, y compris les établissements publics culturels nationaux – EPs -, sous tutelle du Ministère de la Culture, et les collectivités territoriales, associations…

- Edition des notices : les notices peuvent être créées, modifiées, supprimées par les correspondants institutionnels. L'équipe éditoriale HDA garde le contrôle du processus d'édition.

- Annuaire intégré des institutions partenaires

- Détection automatique des liens morts et alerte

- Mise à jour du front office automatique via « batch processing »/traitement par lots[122]

---

[121] Institut crée en 2006, au sein du Centre Pompidou, sous l'impulsion du philosophe Bernard Stiegler, pour anticiper, accompagner, et analyser les mutations de l'offre et de la consommation des pratiques culturelles permises par les nouvelles technologies numériques, et pour contribuer parfois à les faire émerger. L'IRI a pour ambition de participer ainsi à l'élaboration de nouvelles formes d'adresse au public et de contribution, de dispositifs critiques collaboratifs et de technologies éditoriales et relationnelles dans les domaine de la culture et des savoirs, à la fois en théorisant et en formalisant les technologies concernées et les pratiques sociales qu'elles suscitent, et en développant les applications culturelles et scientifiques, notamment dans le domaine muséal et plus généralement comme technologies destinées aux amateurs.

[122] http://fr.wikipedia.org/wiki/Traitement_par_lots

4.1.4 Les fonctionnalités du tagging sémantique

Le projet HDABO implique la « sémantisation » des tags existants. Cette opération ne peut être réalisée sans une intervention humaine assez longue, mais certains traitements peuvent être exécutés en mode semi-automatique. En amont du projet, l'IRI proposera à l'équipe HDA un module de reprise d'existant. Ce module n'aura pas de vocation pérenne, il sera indépendant du module fonctionnel intégré à HDABO. Son utilisation sera limitée à la reprise d'existant et ne s'appliquera qu'au champ tag.

Les entrées de Wikipédia de langue française seront utilisées comme référentiel de tagging par défaut, via une interface de type liste de complétion. Si le mot-clé recherché existe dans Wikipédia, 3 métadonnées sont générées par importation : libellé du tag, URL Wikipédia, URI DBpedia. Si le mot clé n'existe pas dans Wikipédia, seul un libellé est produit. Dans ce dernier cas, le tag n'est pas sémantisé : il ne sert qu'au front office de HDA.

Le classement humain par ordre d'importance décroissante des tags fait l'objet d'une fonctionnalité spécifique. Grâce à cette fonctionnalité, le nombre ordinal du tag (dans la série des tags appliqués à une ressource) vient enrichir les autres métadonnées d'indexation. L'indice de pertinence est destiné à optimiser les fonctionnalités heuristiques de restitution des données ainsi que la pertinence des listes de résultats.

Tous les tags sémantisés sont repris dans une table contenant les 4 métadonnées suivantes : libellé, URL Wikipédia, URI DBpedia, indice de pertinence. L'index des libellés de tags (sémantisés et non sémantisés) est réutilisé, sous forme de simples chaînes de caractères, par le front office de HDA.

4.1.5 Les avantages du projet

La plate-forme de gestion permet:

- Robustesse informatique de l'outil de gestion (substitution à un logiciel bureautique) ;

- Gain de productivité de l'équipe HDA dans la création des notices (saisie directe des données de base par les contributeurs) ;

- Simplification des processus éditoriaux entre l'équipe HDA et les contributeurs (réduction des risques d'erreur) ;

- Garantie pour la mise à jour des données (plus de rapidité, détection automatique des liens morts…) ;

- Edition automatique des nouvelles données sur le front office en traitement par lots (substitution à une opération manuelle) ;

- Renforcer le réseau des personnels opérationnels (400 personnes) contribuant à l'enrichissement de HDA.

Les données sémantisées permettront:

- Stimuler la créativité technologique des partenaires institutionnels du MCC. Parmi les centaines de contributeurs institutionnels de HDA, certains sont à même de lancer des projets orientés web sémantique. Ces porteurs de projet pourront s'appuyer sur ce retour d'expérience et convaincre d'autant plus facilement leurs institutions que HDABO s'applique à leurs données ;

- Sensibiliser les institutions partenaires du MCC aux enjeux de l'indexation sémantique ;

- Émulation les acteurs du web sémantique via un appel à projet national ;

- Soutenir la R&D en offrant à la communauté du web sémantique un terrain d'expérimentation via le corpus sémantisé de HDA ;

- Valoriser les contenus de l'annuaire, et à travers eux, les institutions contributrices ;

- Expérimenter un nouveau mode de réutilisation et de chaîne de valeur appliqués aux données numériques du MCC ;

- Expérimenter un mode de collaboration entre le MCC et la fondation Wikimédia France.

4.2 Présentation du module HDABO de reprise d'existant des tags

Le module 1 de HDABO sert uniquement à traiter le champ tag du corpus HDA, il n'a aucune autre fonction. Il est accessible via HTTP à l'équipe éditoriale de HDA, qui utilise le navigateur Mozilla Firefox dans sa version 3.6.16.

La première contrainte liée à la reprise de l'existant des tags est sa volumétrie. En septembre 2011, HDA comptait environ :

- 5.000 notices ;

- 18.000 tags différents ;

- 14 tags en moyenne par notice ;

- ce qui résulte en 70.000 tags attribués aux notices.

Tout d'abord, il ne faut pas oublier que ces 70.000 tags devront être re-vérifiés, voire modifiés dans les notices, manuellement. D'où l'extrême importance de l'ergonomie, et bien sûr de la robustesse et de la fiabilité de l'outil à être utilisé. Un simple calcul permet de comprendre cet enjeu : un traitement manuel de 30 secondes par tag représente 5 mois d'équivalent temps plein. Si, par exemple, le traitement informatique de la validation d'un tag augmente ne serait-ce que de 10 secondes le processus, cela signifie que le temps de traitement manuel sera rallongé d'au moins un mois !

La finalité du module 1 est de reprendre l'ensemble du corpus HDA existant pour en produire des métadonnées sémantisées, résultat, uniquement des tags qui sont sémantisés. Ces métadonnées sont au nombre de 6 :

- Le libellé du tag ;

- L'URL Wikipédia du tag ;

- Le permalien[123] du tag, c'est-à-dire, la version du tag dans Wikipédia au moment de l'acte du tagging ;

- L'URI DBpedia du tag ;

- L'indice de pertinence (indique la pertinence de chaque tag pour chaque notice, c'est-à-dire, le nombre de fiches qui contient tel tag) ;

- La facette (les tags sont typés selon les catégories : Datation, Localisation, Créateur, École/Mouvement, Discipline artistique). Cependant, un tag peu être dépourvu de facette, c'est-à-dire, pas tous les tags rentrent dans les catégories précedentes.

- L'alias HDA du tag : uniquement pour certains tags.

---

[123] Type d'URL conçu pour référer un élément d'information et pour rester inchangé de façon permanente, ou du moins, pour une certaine période de temps. Le permalien de Wikipédia dans HDABO permet de prendre connaissance de toutes les modifications qui ont été faites après un moment M de consultation de l'article M et de comprendre pourquoi et de quelle manière les informations comprises dans cet article ont été utilisées dans ce moment M.

L'ensemble des libellés de tags, qui comprend les sémantisés et non sémantisés, seront réutilisés par le front office de HDA et traités comme des simples chaînes de caractères.

Ainsi, l'ensemble des ces 6 métadonnées peut être visualisé dans le tableau exemplaire suivant :

libellé	URL Wikipédia	Facette/ catégorie	Permalien Wikipédia	URI DBpedia	Indice
Corps humain	http://fr.wikipedia.org/wiki/Corps_humain	[aucune]	http://fr.wikipedia.org/w/index.php?title=Corps_humain&oldid=64485588	http://fr.wikipedia.org/w/index.php?title=Sculpture&oldid=64798450	1
Sculpture	http://fr.wikipedia.org/wiki/Sculpture	Discipline artistique	http://fr.wikipedia.org/w/index.php?title=Sculpture&oldid=64798450	http://dbpedia.org/page/Sculpture	5

4.2.1 La typologie des liaisons entre HDA et Wikipédia

Sont prévus 4 types de liaison entre le tag utilisé dans le corps de HDA et les entrées de Wikipédia, ce qui entraîne à chaque type de liaison un traitement particulier :

- Appariement : le tag est associé à l'URL d'un article de Wikipédia ;

- Redirection : le tag est associé à l'URL de redirection établie par Wikipédia vers un article. À savoir que, dans ce cas, il existe deux ou plusieurs URLs de Wikipedia pour le même article. Ce cas peut être illustré par l'exemple du tag « Louis XIV ». Ce tag donne comme réponse une URL redirigé : http://fr.wikipedia.org/wiki/Louis_xiv, qui originaire de son URL « canonique » : http://fr.wikipedia.org/wiki/Louis_XIV_de_France. A ces deux URLs s'ajoute une troisième, qui est la page d'homonymie : http://fr.wikipedia.org/wiki/Louis_XIV_(homonymie) ;

- Homonymie : le tag est associé à l'URL d'une page d'homonymie ;

- Résultat nul : le tag est associé à l'URL d'une page de résultat nul dans Wikipédia. Cette page peut suggérer au tagueur des entrées pertinentes, grâce à la fonction de recherche en texte intégral de Wikipédia.

Seuls les tags dont la liaison Wikipédia est de type « Appariement » ou « Redirection » feront l'objet d'un traitement sémantique, qui sera capable de générer des métadonnées (libellé, URL Wikipédia, URI DBpedia, indice de pertinence). Les tags qui n'auront pu être reliés à Wikipédia pourront être conservés (il ne s'agira, donc, dans ce cas, que d'une simple chaîne de caractères), mais ne seront pas stockés dans la table des tags sémantisés.

À noter que tous les tags n'auront pas d'URI DBpedia, car la version anglaise de Wikipédia ne couvre évidemment pas la totalité du contenu de la version française.

4.2.2 En préalable : Le traitement par lot du corpus de tags

Un premier traitement en mode « batch » de l'ensemble de tags sera réalisé en amont du traitement manuel. Son objectif est de « pré-traiter » la totalité de l'index des tags, ce qui fera gagner du temps à l'équipe HDA, avant d'initialiser le traitement manuel.

Le traitement consiste à associer automatiquement, par appariement, un lien à chaque tag, hors du contexte des notices. Il s'agit d'un appariement simple de chaînes de caractères comparant les libellés des tags avec les entrées de Wikipédia. Ce traitement doit également identifier le type du lien et lui attribuer un code couleur dont la fonction est d'avertir le tagueur de la nature du lien, selon la typologie suivante :

- Appariement : lien en couleur bleu **gras** ;

- Redirection : lien en couleur bleu **gras** et en ***italique*** ;

- Homonymie : lien en couleur noir et en **gras** ;

- Résultat nul : lien en couleur noir.

4.2.3 Les fonctionnalités d'édition des tags

L'index des tags, déjà traité par lot, sera restitué à travers une interface d'édition qui permettra à l'équipe HDA d'optimiser manuellement cet index général des tags, hors du contexte des notices. Cette interface est dotée des fonctionnalités suivantes :

Fonction de classement de l'index général des tags :

Avant le retraitement manuel, l'ndex est automatiquement classé selon un ordre à triple clés :

- clé 1 : fréquence d'utilisation des tags par les visiteurs du site HDA : classement décroissant

- clé 2 : occurrence des tags dans le corpus HDA : classement décroissant

- clé 3 : ordre alphabétique : classement croissant

Fonction d'archivage du tag initial :

Le tag d'origine doit être conservé tel quel dans une table, non visible par l'utilisateur. Cet archivage sera utile pour l'optimisation de la fonction de suggestion de tags à développer dans un module 2, qui permettra aux contributeurs institutionnels du site HDA de produire eux-mêmes les tags nécessaires à la description de leurs ressources.

Cette fonctionnalité peut être visualisée dans le tableau suivant, dont la première colonne correspond à l'index des tags d'origine que le module 1 doit archiver, et la colonne 2 correspond au champ tag « label » de l'interface utilisateur :

Tag d'origine	Tag retraité
Centre Pompidou	Centre national d'art et de culture Georges-Pompidou
restauration	Restauration

Fonction « nature du lien » :

L'apparence du libellé du tag est différent selon la nature du lien avec Wikipédia et Histoire des Arts, comme vu précédemment:

- Appariement : **Bleu gras**

- Redirection : ***Bleu gras italique***

- Homonymie : **Noir gras**

- Résultat nul : Noir

- Fonction « Voir l'article Wikipédia » :

Un pictogramme (une flèche bleue) cliquable permet d'accéder à une page Wikipédia correspondant au tag. Le lien, qui s'ouvre dans un nouvel onglet du navigateur Firefox, peut renvoyer à un article, une page d'homonymie ou une page de résultat nul et sera ou non modifié lors de l'analyse du tag.

Le lien vers DBpedia est également visible et accessible via le clique du pictogramme flèche verte.

Fonction « Modifier un tag » :

La zone dans laquelle est situé le tag – le champ tag - est cliquable. Il s'agit d'une cellule dans une interface de type tableau :

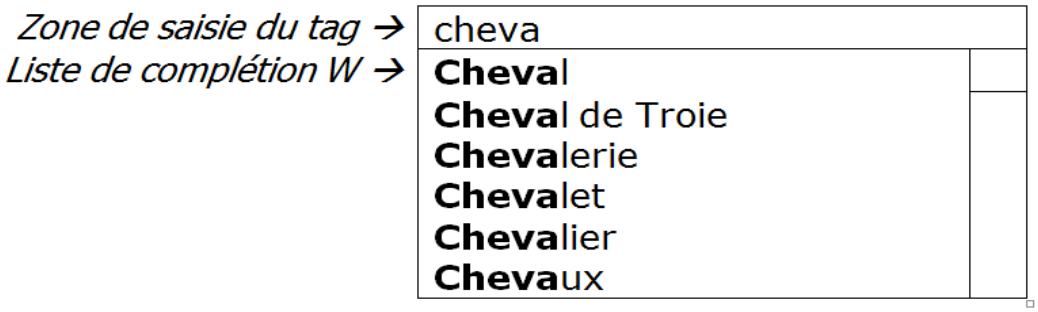

Figure 13 : Le champ tag schématisé par Bertrand Sajus.

Lorsque l'utilisateur clique dans cette zone, le tag est présélectionné et devient modifiable. La liste de complétion Wikipédia se déploie immédiatement, avec un ascenseur pour faciliter la navigation. L'utilisateur peut également saisir une autre chaîne de caractères.

Si l'utilisateur clique sur un mot de la liste de complétion, le tag est validé et le lien immédiatement est généré avec l'article de Wikipédia.

Si l'utilisateur tape un mot qui ne figure pas dans la liste de complétion et tape « Entrée », le mot est validé. Dans ce cas, le lien Wikipédia renverra vers la page de résultat nul.

Fonction « Désactiver/réactiver un lien vers W » :

Une case à cocher permet de désactiver un lien vers un article de Wikipédia tout en conservant le tag.

Dans ce cas, le tag n'est plus sémantisé. Cette fonction est utile lorsqu'un tag est homonyme du titre d'un article de Wikipédia. Lorsque le lien est déconnecté de Wikipédia (« dé-sémantisé »), le tag change d'apparence en passant du bleu au noir.

Par exemple, le tag « harmonie », au sens commun du terme[124], ne fait l'objet d'aucun article dans Wikipédia. Cependant, Wikipédia contient un article sur la notion musicale d'harmonie. Dans ce cas, si le tagueur veut conserver la notion commune d'harmonie, il doit désactiver le lien vers Wikipédia.

Fonction d'attribution de facettes :

Le tag est « qualifié » à l'aide d'un jeu de « catégories » qui jouent le rôle de facettes.

Ces facettes seront choisies en nombre réduit cinq et forment une liste fermée :

- Créateur (Victor Hugo, Claude Monet, Christian Dior, Catherine Deneuve...) ;

- Discipline artistique (peinture, cinéma, littérature...) ;

- Datation (XIVe siècle, 1920, Âge du bronze…) ;

- Localisation (Chine, Toscane, Paris, Montmartre…)

- École/Mouvement (romantisme, art gothique, fauvisme…)

Cette fonctionnalité d'attribution de facette, limitée aux tags liés à Wikipédia, permet, dans la mesure où un tag peu être classé a priori et par défaut dans l'une des catégories listées ci-dessus, un classement en amont pourra être établi au moment de la reprise générale de l'index des tags, avant l'étape de re-taggage notice par notice.

L'interface de classement par facettes se présente comme un menu déroulant accessible en vis-à-vis de chaque tag et est aussi modifiable.

Fonction édition d'un alias :

Dans certains cas, le libellé de Wikipédia n'est pas adapté au contexte éditorial de HDA. Un exemple notoire concerne le tag « Louis XIV » dans HDA et qui dans Wikipédia

[124] Combinaison spécifique formant un ensemble dont les éléments divers et séparés se trouvent reliés dans un rapport de convenance, lequel apporte à la fois satisfaction et agrément. : http://www.cnrtl.fr/definition/harmonie

devient « Louis XIV de France ». « Centre Pompidou » et « Centre national d'art et de culture Georges-Pompidou », respectivement, est un autre exemple.

Cette fonction n'est pas laissée à la liberté de chaque tagueur, du au risque d'incohérence par rapport à l'ensemble des tags, mais est centralisée. Par défaut, tout choix de tag dans la liste de complétion Wikipédia intégrée à l'interface de tagage importe le libellé Wikipédia, non modifiable. C'est dans l'interface de gestion de l'index général que le responsable de l'ndex peut introduire un libellé alternatif local pour certains tags. Cette fonctionnalité n'est donc pas accessible dans l'interface de re-taggage des notices.

Fonction « Supprimer un tag » :

Un clic sur le pictogramme de suppression supprime le tag.

Fonction « Annuler la dernière action » :

En cliquant sur le numéro identifiant du tag - « id » -, l'utilisateur-taggeur annule la dernière action et récupère le libellé du tag.

L'ensemble de ces fonctionnalités est disponible sous forme tabulaire, et peut ainsi être aperçu :

id	label	original_label	Lien W	Lien D	Catégorie	Supprimer le lien W	Alias	Nb de fiches	Popularité
1416	Statue colossale	statue colossale	W					6	22100
1322	Lati	latin						18	22073
958	Latin	nouvelle						9	21937
4033	Latitude	granite						3	21880
9063	Latium	luthier						5	21825
378	Latino	Bretagne			Localisation			60	21808
284	Image	image						89	21807
43	Centre national d'art et de culture Georges-Pompidou	Centre Pompidou					Centre Pompidou	89	21795
433	Restauration	restauration						62	20964
59	Littérature	littérature			Discipline artistique			236	20855
892	Parc	parc						18	20743
627	Opéra	opéra						104	20618
5640	Antiquité gallo-romaine	Antiquité gallo-romaine	W					39	20552
2421	Homme	homme						30	20481
8843	Cordes	cordes						0	20152

Figure 14 : Interface utilisateur de HDABO après le traitement par lot.

Suite à l'opération de traitement général de l'index des tags décrite ci-dessus, une deuxième opération consiste à calculer en « batch », notice par notice, l'indice de pertinence

de chaque tag. Un traitement en « batch » préalable permet d'attribuer un indice de pertinence à chaque tag, pour chaque notice.

Suite à ce traitement, l'équipe HDA pourra re-traiter les tags notice par notice. Cela implique une autre interface que celle utilisée pour la reprise générale de l'index des tags. Cette interface donne accès aux tags dans le contexte de chaque notice. Ses fonctionnalités sont disponibles via une interface de type tabulaire comparable à l'interface de reprise générale des tags. Mais elle donne une vue sur l'ensemble de la notice.

Ce processus ne sera pas abordé dans ce mémoire.

5 D'un tag ambigu à un tag sémantisé : la méthode appliqué en employant le module 1 de HDABO. Les démarches, les exemples

Tout d'abord, il faut mentionner que, du fait de ne pas être francophone d'origine, pourrait rendre le travail encore plus complexe, surtout, si l'on considère que corpus dont les tags font partie est entièrement dédié à l'histoire, culture et langue françaises. Cependant, cela m'a permis d'analyser les tags sans aucun préjugé, sans aucune préconception 100 % certaine, sans vices d'usage de la langue. Cela a donnée une neutralité et, en moindre mesure, une liberté en termes de traitement du vocabulaire, de la sémantique.

Le Portail Lexical du Centre National de Ressources Textuelles et Lexicales a été outil utilisé pour aider à désambiguïsation, surtout en ce qui concerne l'homonymie et l'homographie des mots.

Les tags ont été analysés selon leur popularité : les mots-clés plus recherchés par les utilisateurs, en ordre décroissant d'importance. Ensuite, dans le module 1 de HDABO, les tags sont classés en ordre décroissant par rapport à son incidence dans le corpus d'Histoire des arts, c'est-à-dire, les tags qui ont été utilisés en une plus grande quantité de fiches sont affichés en premier lieu. Comme dernier critère, les tags sont classés en ordre alphabétique quand le nombre de fiches se maintient inaltéré.

Comme critère de désambiguïsation, on été seulement considérés les tags qui ont été utilisés dans 5 fiches, maximum. Les tags qui dont les libellés ont un sens très large comme « valeur », n'ont pas été traités.

Nonobstant, même après la lecture des notices et l'analyse du rapport entre les tags qui ont été attribués à une fiche qui contient le tag en traitement, il faut aller à la ressource/document pour comprendre « dans quel sens » ce tag-là a été employé.

Lors de la réalisation du travail de sémantisation des tags du corps d'Histoire des Arts, la méthode employée a été la suivante :

Appariement : même si les tags qui avaient déjà étés automatiquement sémantisés, à savoir, les tags qui avaient déjà un article univoque dans Wikipédia, il était prudent de vérifier si le contenu de l'article abordait exactement sujet qui appartenait au corpus d'Histoire des Arts. Il fallait ouvrir le lien vers Wikipédia et analyser la page de redirection ou d'homonymie de Wikipédia, pour savoir si l'article avait vraiment rapport avec le corpus HDA.

Tel est le cas du tag « octant ». Wikipédia envoie vers l'article qui concerne la constellation. Néanmoins, d'après l'analyse des fiches relatives à ce tag, on peut constater qu'octant se rapport à l'ancien instrument de navigation !

Une fois que le bon article de Wikipédia a été trouvé, pour conserver la graphie correcte et les caractères spéciaux du titre de l'article, son libellé a été copié-colé dans le champ tag : « Octant (instrument) ». Ou encore « Jiří Kylián », « Leoš Janáček », « Kōbe »

Cet exemple est la preuve empirique d'un travail que les machines ne sont pas encore capables de faire à l'heure actuelle.

Redirection : Comme exemple de page de redirection qui ne posent aucun problème, on peut citer le tag qui avait pour libelle « abbé Suger », que Wikipedia envoie vers l'article qui a comme titre « Suger de Saint-Denis », nouveau libellé attribué au tag, désormais sémantisé. Le même processus peut-être observé pour le tag « gouvernement de Vichy » qui devient « Régime de Vichy »

Parfois, le lien vers lequel Wikipédia envoie n'a aucun rapport avec les notices qui contiennent le tag en traitement. C'est le cas de « jeu collectif », qui Wikipédia redirige vers « Jeu sportif » : les notices qui ont le tag « jeu collectif » n'ont aucun rapport avec le sport. Dans le corpus Histoire des Arts, « jeu colletif » concerne les musiciens qui jouent ensemble !

Et encore : pour le tag « doryphore », Wikipédia pointe vers l'insecte, mais le tag concerne une notice qui a pour objet la statue de Polyclète. Le tag acquiert donc le libellé « Doryphore (Polyclète) ».

De même pour le tag « baigneur ». Wikipédia fait le renvoi vers « poupée » , mais la notice traite du tableau « Baigneurs » de Paul Cézanne, qui retraite des personnes qui se baignent et pas le « poupon de celluloïd ou de plastique (que l'on peut baigner) servant de jouet ».[125]

Homonymie :

Quant à l'homonymie, un exemple très parlant se réfère au tag qui avait originalement le libellé « cheville ». Le lien vers Wikipédia envoyait à une page d'homonymie, qui liste plusieurs significations pour le mot cheville :

- Cheville, l'articulation qui relie la jambe et le pied ;

- Cheville ouvrière, une cheville dans une construction de direction ;

- Cheville, une pièce, souvent en plastique que l'on enfonce dans un mur avant de fixer une vis ;

- De la même origine, la cheville utilisée en menuiserie est une petite pièce en bois ou en métal enfoncée à force pour réaliser l'assemblage mécanique de deux autres pièces. Son utilisation est comparable à celle d'un clou ;

- Cheville, un mot inutile au vers que l'on rajoute pour obtenir le bon nombre de syllabes ;

- Cheville, une pièce mobile permettant de régler la tension des cordes d'un instrument de musique ;

- Cheville, activité bouchère d'abattage et de vente en gros de viande en carcasse, la cheville étant en l'occurrence le crochet auquel le chevillard attachait la viande.

On pourrait supposer, n'oubliant jamais que les ressources auxquelles les tags sont liés font référence à l'histoire d'art, que le tag cheville peut avoir tous sens cités ci-dessus. Il faut donc analyser les notices liés à ce tag-là – travers le bouton "nb de fiches" - pour vérifier que les fiches se réfèrent au métier du luth et à la guitare classique, ce qui permet donc de choisir l'article qui a prend le mot cheville comme la « pièce mobile permettant de régler la tension des cordes d'un instrument de musique ». Le libellé du tag devient alors « Cheville (musique) », terme univoque, lors du moment du tagging sémantique dans le corpus d'Histoire des Arts.

Résultat nul :

Comme le corpus des tags a été constitué de façon folksonomique, il fallait analyser l'orthographie du tag pour savoir si d'abord le tag n'avait pas été écrit d'une façon erronée, induite par une faute de frappe lors de la saisie dans les respectives notices. Comme exemples, « actor studio » qui devient « Actors Studio » et «Ray HarryHausen qui devient « Ray Harryhausen » ou avec l'inclusion de caractères spéciaux : « Edith Piaf » devient « Édith Piaf » ou de tirés : « Saint Jacques de Compostelle » devient « Saint-Jacques-de-Compostelle ».

---

[125] http://www.cnrtl.fr/definition/poup%C3%A9e

Les fautes d'orthographe, qui n'empêchent pas de sémantiser les tags exemplifiés ci-dessus, rendent impossible la sémantisation de « vitre ». On n'est pas capable de savoir s'il agit de vitre, en tant que plaque de matériau transparent ou la commune de Vitré, qui sémantisé deviendrait « Vitré (Ille-et-Vilaine) ».

Les erreurs de saisie comme « XIXèmelittérature » n'ont pas été considérés, vu qu'il s'agit de deux tags : « XIXe siècle » et « Littérature »

Parfois, il était possible de sémantiser un tag vers lequel Wikipédia n'envoyait à aucun article. C'est le cas, par exemple du tag « abbé Hilduin », qui n'a pas d'entrée dans Wikipédia. D'après l'analyse des notices, on observe qu'il s'agit de Hilduin de Saint-Denis, qui peut ainsi être sémantisé, en devant le tag « Hilduin de Saint-Denis ».

À la fin, encadrer le tag dans les catégories, si possible : « Vienne », même si l'on ne peut pas désambiguïser, vu qu'il s'agit d'une ville en France et de la capitale de l'Autriche, on peut quand même catégoriser. Ou encore, même pour les tags qui ne portent pas d'entrée dans Wikipédia, c'est possible de les catégoriser, comme dans l'exemple de l'architecte David Grenne, qui l'on peut mettre dans la catégorie Créateur.

6 Bilan du processus de sémantisation dans HDABO

À la fin du traitement de l'index général de tags et avant le travail d'analyse fiche à fiche, notice par notice, qui évaluera la pertinence du tag désormais sémantisé contenu dans les fiches, en excluant les mots-clés qui ont été vidés de leur substance (les tags qui contenaient des fautes de frappe ou qui assemblaient deux tags différents, v.g. « XXèmepoésie »), environ 15.000 tags envoyaient au moins une fiche. De ces 15.000 tags, environ 12.000 tags ont été liés à des entrées de Wikipédia, c'est-à-dire, ont été sémantisés, ce qui permet affirmer que 80% de l'index général de tags a été sémantisé.

D'une part, les tags qui n'ont pas pu être sémantisés sont des mots polysémiques, comme « composition », « espace », « exposition », « forme », « rupture », qui en conséquence renvoient à des pages d'homonymie, qui même tenant en compte l'orientation du corpus de Histoire des Arts, il est pas possible d'effectuer leur désambiguïsation et qui répercutent dans une quantité considérable de fiches/notices, à savoir : beaucoup de ressources possèdent ces tags dans leurs notices.

D'autre part, des tags qui ne sont pas encore sujet d'un article dans Wikipédia, comme dans les cas de tags qui ont comme libellé le nom d'artistes contemporains.

Quant à relation temps/coût du travail, en admettant qui serait possible de traiter manuellement environ 400 tags par jour, à temps plein, dans un corpus constitué de 15.000 tags, la sémantisation de l'index général de tags pourrait être réalisé en 8 semaines.

L'absence d'articles dans Wikipédia, qui pourrait être considéré comme une faiblesse dans l'option pour l'adoption de cette encyclopédie comme référentiel, peut avoir comme justificative le fait de qu'il n'y a pas une quantité important d'utilisateurs qui recherchent sur ces sujets, dont la communauté de contributeurs de Wikipédia n'a pas encore rédigé des articles correspondants.

L'inclusion de catégories aux tags, lors de l'exécution du processus de sémantisation des tags, permet aussi à Wikipédia d'identifier les sujets/personnes qui ne font pas l'objet d'un article.

Cela dit, l'absence d'items dans un référentiel, en l'occurrence, des articles dans Wikipédia, permet une autre approche pour la recherche et restitution de l'information qui sera basée sur l'usage des items appartenants à un référentiel donné.

7 Perspectives du tagging et indexation sémantique

Les perspectives du tagging et de l'indexation sémantique sont liées surtout à de nouvelles possibilités de recherche et restitution de l'information.

Parmi les perspectives du tagging et indexation sémantique, on pourrait citer :

- l'intégration dans des applications d'indexation et de recherche ;

- aisance de la navigation ;

- affiche de résultats équivalents ;

- recherche multilingue, bien aussi choisir ou forcer la langue d'affichage ;

- auto-complétion ou choix des concepts correspondant aux caractères saisis ;

- naviguer par des relations hiérarchiques et associatives

- déployer ou fermer une arborescence ;

- passer d'une présentation à une autre (alphabéthique, hiérarchique) ou avoir les 2 en parallèle

- catégorisation pour affiner la recherche.

La recherche multilingue permet de rechercher sur un contenu qui n'est pas dans la langue d'interrogation de l'utilisateur. Elle permet donc à l'utilisateur de trouver l'information souhaitée sans avoir besoin de faire la traduction de son équation de recherche. Ainsi, un utilisateur qui pose une requête dans un moteur de recherche sur « coing », n'aura pas besoin de faire la traduction ou trouver des équivalents linguistiques vers l'anglais, « quince », le portugais « marmelo » ou l'espagnol « membrillo ». Le moteur de recherche est donc capable d'afficher les résultats de ressources qui contiennent non seulement le mot « coing », mais aussi tous les autres mots dans les idiomes choisis.

On peut aussi envisager la construction d'interfaces innovantes, employant les potentialités du tagging sémantique, comme une carte heuristique (la navigation qui se fait du général vers le spécifique), une ligne de navigation temporelle (ligne de temps), carte géolocalisée, outil de recherche « cross-language », seuls ou ensemble !

C'est grâce à l'association de données structurées à un contenu qu'une navigation par facettes, avec fonctionnalités comme celles cités ci-dessus, devient possible.

Le tagging sémantique permet aussi la création des ontologies des tags, en utilisant la puissance des ontologies pour les relier. C'est ce qui proposent MOAT – Meaning Of A Tag d'Alexandre Passant et NiceTag d'Alexandre Monnin, Fabien Gandon, Freddy Limpens et David Laniado.

L'utilisation des catégories permet aussi de diviser les catégories par lot, en créant des index catégoriques, ce qui habilite l'utilisateur à trier et affiner les résultats d'une recherche, ce qui favorise aussi la réutilisation des ressources, puisque l'utilisateur va pouvoir naviguer dans l'arborescence des catégories.

En ce qui concerne l'utilisation de Wikipédia comme référentiel, l'exploitation des métadonnées contenues dans les infobox[126] des articles a un énorme potentiel. On peut, par exemple, créer un index par lieu de naissance de toutes les chanteuses brésiliennes, en affichant leurs images respectives ou la localisation des lieux sur une carte, grâce aux métadonnées contenues dans les infobox, sans que ces métadonnées-là soient stockées dans la base de données qui contient les ressources elles-mêmes. On parlera alors de « meta-meta-données »[127], ce qui représente une importante économie de travail. Tout cela est possible grâce au contenu librement réutilisable de Wikipédia sous la licence Creative Commons.

L'utilisation de l'arborescence des catégories présente dans Wikipédia permettra aussi d'harmoniser les thésaurus, sans aucun travail conceptuel. Il fallait reprendre l'organisation des catégories déjà faite par Wikipédia. Le thésaurus deviendra dynamique, sans besoin de mises à jour, vu qu'il sera attaché (« greffé ») à la politique éditoriale de Wikipédia.

On pourrait encore envisager une « indexation croisée »[128], entre plusieurs référentiels, comme Wikipédia et Wiktionnaire, pour citer deux projets de la Fondation Wikimedia. Ou encore entre les 295 référentiels qui font partie du projet Linking Open Data, en septembre 2011, comme, par exemple, MusicBrainz et GeoNames.

---

[126] Une infobox sur Wikipedia est une table préformatée de données dynamiques qui présente sommairement des informations importantes sur un sujet dans un cartouche ou encadré placé en général à droite de l'article.

[127] D'après Bertrand Sajus.

[128] Idem.

Conclusion

Et alors ? Pourquoi faire du tagging et de l'indexation sémantique ?

Le tagging sémantique du corpus d'histoiredesarts.culture.fr permet la structuration sémantique de son contenu, et par là même il en augmente considérablement la sérendipité. Au-delà des avantages immédiats pour l'utilisateur final, cette expérience contribue à faire prendre conscience aux institutions culturelles de l'importance de grands référentiels en ligne répondant aux normes du web de données. Elle met également en évidence l'importance d'un accès (SPARQL endpoint) à la version française de Wikipédia. Sans cet accès, la sémantisation est dépendante de la version anglaise de Wikipédia.

Si l'on croyait que le développement de l'informatique documentaire à partir des années 1970 et des technologies du Web à partir des années 1990 pourraient mettre fin à des techniques documentaires telles que l'indexation, surtout avec la vulgarisation de la folksonomie parmi les internautes dans les années 2000, son absence de structuration, a rendu le repérage de l'information une tâche parfois herculéenne.

Les référentiels, qui ont toujours été le fondement de la gestion et de l'organisation de l'information, ont aussi été mis en cause avec la propagation du Web 2.0, surtout dans un contexte où sont devenues possible des recherches en texte intégral.

Toutefois, comme vu antérieurement, la folksonomie, qu'Alexandre Passant caractérise comme « un amas de tags chaotiques et non organisés » [20, Passant], n'a pas pu montrer son efficacité dans l'optimisation de la recherche. L'escapade de la folksonomie quant à création de relations explicitement définies entre les tags ne permet pas que ces relations soient prises en compte au niveau de la recherche et restitution de l'information.

Le Web de Données réhabilite ainsi le rôle des référentiels. C'est « la revanche des référentiels face à la folksonomie », comme atteste Gautier Poupeau[129].

Il a été possible de signaler l'importance d'attacher un concept contrôlé, issu d'un référentiel, lors de l'indexation des ressources, ce qui apporte aux moteurs de recherche des données que la folksonomie tout simplement n'est pas capable d'offrir.

Dans l'esprit de complémentarité entre le Web 2.0 le Web 3.0, auquel Tim Berners-Lee se réfère quand il nous montre sa vision du Web de Données, la folksonomie, pourtant, ne doit pas être vue comme ennemie des référentiels. Bien au contraire ! L'annotation libre des ressources sert à pousser l'enrichissement et l'évolution de référentiels vivants et multilingues tels comme Wikipédia.

L'adoption de référentiels reliés et interopérables lors de l'indexation enrichit la performance de la recherche, ce qui permet d'affiner la qualité des résultats trouvés, tout en restituant l'information de manière plus précise, facile et rapide pour l'utilisateur.

Le capital d'informations intrinsèque d'un référentiel comme Wikipédia et l'exploitation de ses métadonnées, alliés à toutes les technologies du Web Sémantique, nous démontrent que les langages documentaires et les techniques documentaires sont en constant évolution. Cela ouvre la porte à la sémantisation de thésaurus, à l'indexation croisée entre référentiels interopérables...

L'avenir du monde de l'information ne pourrait pas être rien de plus qu'attirant !

[129] http://www.bnf.fr/documents/afnor2011_poupeau.pdf

Bibliographie

Cette bibliographie analytique comporte 21 références classées par ordre alphabétique d'auteurs. Les références sont notées, dans le corps du texte, par un numéro suivi du nom de l'auteur entre parenthèses ; cette notation permet de situer la référence dans la bibliographie. Cette bibliographie, élaborée selon les normes Z44-005 (décembre 1987) et NF ISO 690-2 (février 1998), a été arrêtée au 20 novembre 2011.

Le corpus du texte contient des hyperliens vers la plupart des ressources consultées, en privilégiant le contenu de Wikipédia, ce que permet de démontrer la richesse et le potentiel d'exploitation de cette encyclopédie.

[1] AMAR, Muriel. Les fondements théoriques de l'indexation : une approche linguistique. Paris, ADBS Éditions, 2000. 355 pages.

Publication de la thèse de doctorat de la conservatrice, qui traite l'indexation comme une pratique fondatrice de la documentation.

[2] BERNERS-LEE, Tim; BIZER, Christian; HEATH, Tom; Berners-Lee, Tim. (2009). Linked Data: The story so far. International Journal on Semantic Web and Information Systems (Special Issue on Linked Data), 2009. Disponible sur : http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf [Consulté le 8 octobre 2011].

Article qui établie les fondements du projet Linked Data.

[3] BOULOGNE, Arlette ; VALDERRAMA, Asunción. Web 3.0 et web sémantique : du vin vieux, dans des outres neuves. Documentaliste-Sciences de l'information, 2009, vol. 46, n°. 3, p. 20-21.

Les auteures soutiennent que la recherche sémantique n'a rien de totalement nouveau, vu que l'information a toujours été structurée en utilisant des métadonnées.

[4] CHAUMIER Jacques. Les ontologies : Antécédents, aspects techniques et limites. Documentaliste-Sciences de l'Information, 2007/1, Vol. 44, p. 81-83. Disponible sur : http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-81.htm [Consulté le 29 septembre 2011].

Article qui trace un panorama sur les ontologies.

[5] DÉGEZ, Danièle ; MENILLET, Dominique. Thésauroglossaire des langages documentaires : un outil de contrôle sémantique. Paris : ADBS Éditions, 2001. 181 p. (Sciences de l'information. Série Recherches et documents)

Ouvrage relatif au domaine de l'analyse, de l'indexation et des langages documentaires, qui propose une liste de définitions de termes et qui met en évidence les relations sémantiques qu'ils entretiennent entre eux.

[6] DÉGEZ, Danièle. Construire un thésaurus, Archimag n°. 222, mars 2009, p. 44-45
L'auteure propose une procédure de base, à travers la rédaction d'un cahier des charges, pour concevoir un thésaurus.

[7] DIDIER, Marie. Indexation, structuration et encodage des fonds iconographiques : le fonds Léon Lefebvre de la Bibliothèque Municipale de Lille. 2005. Mémoire. ENSSIB. 111 pages. Disponible sur : http://www.enssib.fr/bibliotheque-numerique/document-809 [Consulté le 3 octobre 2011].

Mémoire qui traite de l'indexation, structuration et encodage des collections numériques.

[8] ENDRIZZI, Laure. L'édition de référence libre et collaborative : le cas de Wikipédia. Cellule de Veille Scientifique et Technologique de l'Institut National de Recherche Pédagogique. Avril 2006. 39 pages. Disponible sur : http://ife.ens-lyon.fr/vst/DS-Veille/Dossier_Wikipedia.pdf [Consulté le 17 octobre 2011].

Dossier sur l'édition de référence libre et collaborative qui analyse le cas de Wikipédia, en mettant en évidence le multilinguisme de cette encyclopédie.

[9] FRANCIS, Elie ; QUESNEL, Odile. Indexation collaborative et folksonomies. Documentaliste-Sciences de l'information, 2007, vol. 44, n°. 1, p. 59-60.

Article qui traite de l'indexation collaborative et des folksonomies, en faisant la distinction par rapport à la personne de l'indexeur.

[10] GILCHRIST Alan. Thesauri, taxonomies and ontologies – an etymological note. Journal of Documentation, Vol. 59, nº. 1, 2003, pages 7 – 18. Disponible sur : http://bibliologia.info/archivos/Thesauros%20taxonomias.pdf [Consulté le 28 septembre 2011].

Article qui aborde les différences étymologiques entre les thésaurus, les taxonomies et les ontologies et qui présente 5 acceptions différentes pour le concept de taxonomie.

[11] GRUBER, Thomas R. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Human-Computer Studies 43, p. 907-928. Substantial revision of paper presented at the International Workshop on Formal Ontology, March, 1993, Padova, Italy. Available as Technical Report KSL 93-04, Knowledge Systems Laboratory, Stanford University. Disponible sur : http://tomgruber.org/writing/onto-design.pdf [Consulté le 22 septembre 2011].

Article qui décrit le rôle des ontologies et leur façon de contribuer au partage des connaissances.

[12] HUDON, Michèle. Le passage au XXIe siècle des grandes classifications documentaires. Documentation et bibliothèques, 2006, vol. 52, n°. 2, p. 85-97.

Article qui traite de l'adaptation des grandes classifications documentaires aux nouvelles technologies du Web.

[13] LANDRY, Patrice. Multilinguisme et langages documentaires : le projet MACS en contexte européen. Documentation et Bibliothèques, ASTED, avril-juin 2006, p. 121-129.

Dans cet article, le multilinguisme est abordé comme facteur d'interopérabilité entre les langages documentaires.

[14] LE DEUFF, Olivier. Folksonomies, BBF, 2006, n°. 4, p. 66-70. Disponible sur : http://bbf.enssib.fr/consulter/bbf-2006-04-0066-002 [Consulté le 8 octobre 2011].

Article qui présente un panorama des folksonomies, en montrant les avantages et les inconvénients de cette pratique.

[15] LEISE, Fred. Controlled vocabularies: an introduction. The Indexer, vol. 26, nº. 3, september 2008, pages 121-126. Disponible sur : http://www.ingentaconnect.com/content/index/tiji/2008/00000026/00000003/art00009 [Consulté le 21 septembre 2011].

Article qui présente une introduction sur les concepts et les terminologies des vocabulaires contrôlés.

[16] MANIEZ, Jacques. Les langages documentaires et classificatoires : conception, construction et utilisation dans les systèmes documentaires. Paris, Éditions d'organisation, 1987. 291 pages.

Cet ouvrage de référence sur les langages documentaires présente définitions, méthodes, modèles et exemples sur tous les types de classification et d'autres langages documentaires.

[17] MANIEZ, Jacques. Des classifications aux thésaurus : du bon usage des facettes. Documentaliste – Sciences de l'information, 1999, vol. 36, n° 4-5, p. 249-262. Disponible sur : http://www.adbs.fr/des-classifications-aux-thesaurus-du-bon-usage-des-facettes-13338.htm?RH=REVUE [Consulté le 3 octobre 2011].

Cet article aborde les classifications et les thésaurus vis-à-vis l'utilisation des facettes.

[18] MONNIN, Alexandre. Qu'est-ce que le Web Sémantique ? Disponible sur : http://cblog.culture.fr/2011/09/07/web-semantique-iri-opendat [Consulté le 7 septembre 2011].

Billet d'Alexandre Monnin qui explique de façon très didactique ce qui est le Web sémantique.

[19] MONNIN, Alexandre. Qu'est-ce qu'un tag ? Entre accès et libellés, l'esquisse d'une caractérisation. IC 2009 : 20es Journées Francophones d'Ingénierie des Connaissances, « Connaissance et communautés en ligne », Hamamet : Tunisie, 2009. Disponible sur : http://ic2009.inria.fr/docs/papers/Monnin_IC2009_41.pdf [Consulté le 7 septembre 2011].

Manuscrit qui fait la caractérisation des tags en les distinguant rigoureusement de toutes les expressions linguistiques, ce qui culmine pour constater l'absence de sémantique propre aux tags.

[20] PASSANT, Alexandre. Technologies du Web Sémantique pour l'Entreprise 2.0. 2009. Thèse. Informatique. Université Paris IV- Sorbonne. 298 pages. Disponible sur : http://www.edf.com/fichiers/fckeditor/Commun/Innovation/theses/These_Passant_vf.pdf [Consulté le 2 octobre 2011].

Thèse qui s'inscrit dans le cadre des travaux relatifs à la complémentarité entre le Web sémantique et Web 2.0 et qui traite dans le chapitre 3 des folksnomies et du tagging sémantique.

[21] PIERRE, Béatrice. L'avenir des langages documentaires dans le cadre du Web sémantique : conception d'un thésaurus iconographique pour le Petit Palais. 2010. Mémoire. INTD-Cnam. 115 pages. Disponible sur : http://memsic.ccsd.cnrs.fr/docs/00/57/50/53/PDF/PIERRE.pdf [Consulté le 20 septembre 2011].

Ce mémoire est une étude sur les langages documentaires et leur utilité pour structurer les informations dans l'idée d'un Web sémantique.