L'usage des vocabulaires contrôlés dans les systèmes informationnels d'aujourd'hui : exploitation du thésaurus sur le site du journal Libération

Silvie Skarupova

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

Ecole Management et Société

Département Culture Information Technique et Société (CITS)

INTD

MEMOIRE pour obtenir le

Titre professionnel "Chef de projet en ingénierie documentaire" INTD

RNCP niveau I

Présenté et soutenu par

Silvie Skarupova Pecnard

25 octobre 2012

L'usage des vocabulaires contrôlés dans les systèmes informationnels d'aujourd'hui : exploitation du thésaurus sur le site du journal *Libération*

Jury : Sylvie Dalbin
Bénédicte Dumont

Promotion 42

Remerciements

Merci à Bénédicte Dumont pour son implication dans la réalisation de ma mission tout au long de mon stage et son intérêt pour l'avancement de mon mémoire, mais également pour ses encouragements, sa confiance et son soutien. Merci à son équipe qui s'est montrée très disponible pour répondre à mes questions.

Je remercie Sylvie Dalbin pour sa disponibilité et son grand professionnalisme qu'elle partage avec tant de pédagogie, de générosité et de naturel. Merci pour tous ses commentaires qui m'ont considérablement aidée à voir plus clair et à me sentir plus à l'aise dans la structuration et le développement de ce mémoire, et m'ont permis de surmonter des moments de passages à vide.

Notice

Le mémoire traite des notions liées à l'accès à l'information dans le contexte des sites d'information de presse. A travers le projet d'amélioration de la recherche sur le site web du journal *Libération*, l'auteur analyse le vocabulaire contrôlé existant, le thésaurus, et propose sa transformation en taxonomie, un nouveau référentiel exploitable lors de la mise en place d'une taxonomie de navigation. L'auteur défend l'usage des vocabulaires contrôlés dans les systèmes informationnels d'aujourd'hui et notamment dans la structuration de la base de données de la presse d'actualité pour offrir des moyens d'accès aux internautes. Enfin, cette étude aboutit à une réflexion autour des nouveaux enjeux pour les documentalistes et les professionnels de l'information, liés aux technologies actuelles, et particulièrement au web sémantique, et à un repositionnement de leur statut face à la stratégie de l'entreprise tout en évoquant l'importance d'une bonne gestion de projet.

Mots-clés : accès à l'information, site web, presse, presse en ligne, documentaliste, méthode, analyse, vocabulaire contrôlé, langage documentaire, thésaurus, taxonomie, indexation, indexation automatique, indexation contrôlée, web sémantique, web de données, technologie, technologie de l'information, usager, architecture de l'information, métier.

Table des matières

Introduction

Les sites web font aujourd'hui entièrement partie de tous les domaines de notre société. Ils sont créés dans l'objectif de faire partager des idées et des connaissances, pour des raisons éducatives, pédagogiques, scientifiques, commerciales ou de loisirs. Leur fonction première est tout simplement d'informer. De nombreuses entreprises ne pourraient plus se passer d'un site web devenu un moyen de stratégie et de réussite qui reflète l'activité et la politique de l'entreprise. Par conséquent, les moyens indispensables à investir dans sa création, son développement, sa maintenance et sa mise à jour ne sont pas à négliger.

Il en est de même pour un site d'information d'un journal de presse. Quelle que soit la réponse à la question de savoir si la presse électronique représente un nouveau type de média ou non, il est certain qu'un site d'information occupe aujourd'hui une place de plus en plus centrale pour les gestionnaires de l'information, et ce, quelle que soit la nature de l'information diffusée. En ce sens, il est important de rappeler sa valeur ajoutée par rapport au support papier.

Le présent mémoire s'interroge sur les sites de la presse dite d'actualité et particulièrement sur celui du journal *Libération,* ainsi que sur la façon dont le site liberation.fr pourrait être amélioré. La qualité de l'accès à l'information dépend d'abord de l'effort consacré à structurer l'information et de la granularité de cette structuration. La création des sites d'information est, la plupart du temps, perçue d'un point de vue informatique, graphique ou ergonomique, mais qu'en est-il du regard purement informationnel, celui des professionnels de l'information et des documentalistes ? Le service informatique et le service de la documentation se sont posé cette question. Quelle place pourra occuper le service de la documentation de *Libération* dans cette structuration et plus précisément quel rôle pourra y jouer leur outil du quotidien, qui sert depuis longtemps à indexer et à chercher les articles ? Ce vocabulaire contrôlé, appelé le « thésaurus », pourra-t-il acquérir une nouvelle fonction directement liée à la structuration du site ?

Les utilisateurs d'un site ont besoin de naviguer pour trouver les informations qui les intéressent. A la manière d'une grande partie des sites commerciaux d'aujourd'hui, les sites d'information peuvent également offrir la possibilité de trier les résultats, de chercher et naviguer par critères. Le développement de cette recherche/navigation à facettes passe par ce qu'on appelle communément la taxonomie de navigation. J'ai d'abord tâché de définir ce terme, pour en venir ensuite au propos central de ce mémoire, le rapprochement de l'existant (thésaurus) et de la taxonomie de navigation. C'est l'analyse du vocabulaire contrôlé actuel qui donne une visibilité sur la modélisation d'un nouveau langage documentaire et, en même temps, d'un nouvel outil pour structurer le site et un moyen pour les utilisateurs de s'y retrouver.

Ensuite, j'ai cherché à évaluer, de manière plus générale, dans quelle mesure aujourd'hui, les vocabulaires contrôlés étaient utiles, voire, indispensables, à la création des sites d'information. Quelle part joue l'existant ? Quelle serait la structure du vocabulaire contrôlé adapté au corpus de la base de données ? Les professionnels de l'information parlent de thésaurus, mais aussi d'ontologies, de *topics maps* (cartes topiques), de taxonomies, de vocabulaires contrôlés hybrides construits en fonction des besoins de tel ou tel site. Toutes ces modélisations exigent des technologies ou une connaissance des technologies dont le web sémantique promet un grand pas en avant pour l'activité documentaire. Quel serait le rôle des professionnels de l'information dans ces nouvelles perspectives ? Les documentalistes de *Libération*, par exemple, peuvent-ils s'attendre à un changement dans leur travail et doivent-ils même être initiateurs de ce changement ?

Tout au long de ma réflexion, j'avais recours aux ouvrages et articles dont l'essentiel est répertorié dans la bibliographie (voir p. 83). La plupart de ces références sont de sources françaises car pour la construction de ce mémoire elles se sont avérées suffisamment représentatives de la problématique. En revanche, il est indéniable que la production anglo-saxonne reste plus importante et plus riche, notamment pour les questions sur l'implication des architectes de l'information dans la structuration des sites web.

Première partie
Accès à l'information

La qualité d'accès à l'information sur un site d'information passe par la bonne gestion d'une quantité importante d'informations. Il est nécessaire de s'interroger sur la façon dont les lecteurs profitent de la richesse que peut représenter la base de données d'un tel site. On verra que le site liberation.fr offre des possibilités de recherche limitées. Les lecteurs ne sont pas accompagnés et aidés dans la recherche par sujets. En dehors des systèmes d'information et des standards d'échange développés spécialement pour le secteur de la presse, les documentalistes contribuent depuis longtemps à classer thématiquement des articles de presse. Ce savoir-faire s'appuie sur des vocabulaires contrôlés et des langages documentaires.

1 La structure des sites d'information de presse

1.1 Un site d'information de presse

Les sites d'information, et notamment ceux de la presse, ont pour objectif principal de diffuser des informations liées aux événements les plus récents. De plus en plus, ils complètent, offrent d'autres contenus que celui du journal papier, voire, remplacent le support papier traditionnel. Avec la crise de la presse, un grand débat existe depuis plusieurs années déjà autour de la question suivante : faut-il garder ou abandonner le support papier. Pratiquement tous les titres des quotidiens, en France ou à l'étranger, ont modifié leur façon de publier l'information. Certains titres ont abandonné le support papier et n'existent qu'en version électronique, certains ont réparti la publication en deux supports (papier et web) et certains ont été créés avec le web et n'existent que sur le web. Rares sont aujourd'hui les titres qui continuent leur publication uniquement en version papier (par exemple, *Le canard enchaîné*). Il va donc de soi qu'un site d'information de presse demande à être soigné et mis à jour régulièrement, et ceci non seulement au niveau du contenu lui-même (des informations publiées), mais également au niveau de l'ergonomie, du graphisme, de son référencement sur Internet et de sa structure pour faciliter l'accès.

Par structure, on entend ici moins la structure purement éditoriale, c'est-à-dire la répartition logique et rationnelle de l'ensemble des contenus en rubriques qui sont définies en fonction des thèmes et de la ligne éditoriale de tel ou tel journal, que la structure moins visible au premier abord, intérieure en quelque sorte, celle du contenu global, c'est-à-dire de l'ensemble des articles publiés sur le site d'un journal. L'intérêt d'un site d'information, par rapport au support papier, réside dans son usage documentaire, c'est-à-dire dans la possibilité de fournir au lecteur l'article qu'il cherche. Ceci est possible grâce à la base de données en texte intégral. Cet archivage du contenu permet de consulter d'autres informations que celles publiées le jour même. C'est cette structuration de la base de données que le lecteur découvre au moment où il se met à chercher une information particulière, ou quand il se met à naviguer sur le site sans vraiment chercher une information précise. Ce sont la recherche et la navigation qui permettent de se faire une idée de la façon dont un site web est structuré. La bonne structuration reflète l'architecture du site, l'organisation et la modélisation de l'information. C'est une vraie valeur ajoutée par rapport au support papier.

Je n'aborderai pas dans ce mémoire les méthodes de recherche 2.0 telles que blogs, wikis, flux RSS ou réseaux sociaux, qui sont basées sur l'interaction sociale et qui peuvent également contribuer à cette structuration, l'exploiter et la mutualiser. Les méthodes de recherche 2.0 ne représentent pas directement le contexte du projet de *Libération* qui demande à être traité surtout du point de vue documentaire.

1.2 Le site web de *Libération*

L'informatisation de *Libération* date de 1994 et la mutation numérique du journal commence en 1995 avec l'ouverture du site. En France, il fait partie des premiers sites web de presse créés avec celui du *Monde* et des *Echos*. Toute la gestion du contenu du journal fonctionne avec deux systèmes principaux : Méthode (système éditorial EidosMédia assurant le

processus de la production, *workflow*, des journalistes) et Quai (le back-office du site permettant la relation avec les internautes).

Dans un premier temps, le site web reprenait tout simplement les articles du support papier. Au fur et à mesure, le site s'est différencié du journal papier, il est devenu complémentaire tout en assurant la fonction d'archives des articles papier et web. Les abonnés du journal ont aujourd'hui accès à tous les articles publiés instantanément après la publication et les non-abonnés peuvent consulter les articles dans les 24 heures à compter de la publication. En moyenne, parmi 300 nouveaux objets quotidiens sur le site, il y a 100 articles papier, 50 articles web, 100 dépêches et 50 objets média[1]. Les articles web ne relèvent pas du même genre éditorial que ceux du papier ; il s'agit avant tout d'enquêtes rapides. Les journalistes du web ne sont pas spécialisés dans tel ou tel domaine, comme c'est le cas des journalistes papier.

Les actualités du jour sur le site sont réparties en 12 rubriques[2] principales qui correspondent à celles du support papier. Quant à la recherche d'une information précise, on s'aperçoit assez rapidement que le lecteur ne peut pas aisément naviguer et chercher l'article de la base de données en fonction des thèmes ou des sujets.

Recherche simple : tri des résultats par pertinence

L'utilisateur du site, après avoir posé sa requête, peut trier les résultats par date ou « par pertinence », mais ce choix n'est pas systématique. Le principe du classement des résultats est celui de l'occurrence, c'est-à-dire d'un critère statistique (combien de fois un mot dans la requête apparaît-il dans le document), et sans traitement linguistique (singulier/pluriel, féminin/masculin, synonymes, fautes d'orthographe).

Le tri des résultats « par pertinence » est désactivé quand les requêtes sont trop générales et remontent trop de résultats, - ce « qui coûte cher en ressources [informatiques] » (selon le responsable informatique) -, c'est-à-dire que la pondération[3] de l'article par rapport à l'ensemble des articles n'est plus effectuée. En effet, une telle opération nécessite le traitement d'une grande quantité d'articles et cela pose un problème de gestion du serveur hébergé à l'extérieur (PilotSystem).

Alors que c'est surtout à ce moment-là - ***quand le sujet est très large et les résultats nombreux - que l'utilisateur a besoin d'être aidé et que l'option par pertinence s'avère nécessaire.***

L'utilisateur est tout simplement noyé dans des centaines, voire des milliers d'articles. Il n'a pas la possibilité d'affiner sa recherche.

[1] Information recueillie auprès de la responsable du Service de la documentation de *Libération*.
[2] A la Une, Edito, Politiques, Société, Monde, Economie, Médias, Vous, Sport, Terre, Sciences, Désintox
[3] Certains champs (titres, sous-titres) ont plus de poids que le contenu de l'article. La pondération de l'article (calcul du poids d'un article reposant sur l'occurrence) est déterminée en fonction de l'algorithme statistique du moteur de recherche.

Recherche avancée

Dans le cas où l'utilisateur opte pour la recherche avancée, il est également très limité par rapport à la définition du sujet. La recherche avancée porte uniquement sur les trois critères:

- période (aujourd'hui / depuis 6 mois / depuis 1994 / à une date précise)
- source (journal papier / site web / tout)
- rubrique (25 au total) : choix des rubriques « papier » et « web » confondues

La possibilité d'affiner la recherche par sujet (ou mot-clé) n'est pas proposée sur le site : elle n'est possible ni via un formulaire de recherche, ni via la navigation par mots-clés.

La recherche par sujet peut être aujourd'hui effectuée grâce aux outils linguistiques, appelés vocabulaires contrôlés, qui permettent la structuration de l'information ainsi que la gestion de l'information.

Je rappelle rapidement dans la partie suivante les origines de la création des bases de données, parallèlement à la constitution des bases de connaissances.

2 Des classifications du savoir à la gestion des bases de connaissances

Les grandes classifications émergent au XIX^{e} siècle et ont connu un grand essor surtout dans la première moitié du XX^{e} siècle. La raison de leur développement est liée au besoin de classer des grandes bibliothèques pour organiser le savoir et retrouver les documents. Ce besoin de classer le savoir existe depuis l'Antiquité et on doit les premières classifications bibliographiques durables à Francis Bacon qui classe le savoir en trois grands domaines : l'histoire, la poésie et les arts, la philosophie et les sciences.

Les limites des grandes classifications

Les classifications d'aujourd'hui sont issues de quelques grands modèles. Avec l'accroissement et le développement des bibliothèques, les besoins des bibliothécaires aux USA aboutissent aux importantes classifications comme la classification Dewey ou la CDU.

Les grandes classifications, en répertoriant et en classant les domaines du savoir, ont certes permis de définir des sujets traités dans la représentation des connaissances de l'homme, mais elles se limitent à classer le savoir par grands domaines universels seulement. L'apport des premières classifications est donc avant tout dans cette première organisation des connaissances, dans le classement par domaines qui correspondent aux thèmes ou aux sujets d'un document. Ces classifications étaient et restent destinées avant tout à l'usage des bibliothécaires et s'avèrent insuffisantes dans l'organisation des bases de connaissances des autres secteurs.

2.1 La presse et l'IPTC

Depuis l'âge d'or des grandes classifications, les secteurs de l'activité humaine se sont considérablement multipliés. Chaque secteur a ses propres besoins et crée ses propres organisations et systèmes d'information. L'arrivée de l'information numérique a également très largement contribué aux besoins de trouver de nouveaux moyens d'organisation et de structuration de l'information.

Dublin Core[4] (ou *Dublin Core Metadata Initiative*) permet d'organiser et de structurer les ressources numériques, précisément les métadonnées. Il est construit autour de 15 éléments de description :

- formels (titre, auteur, éditeur)
- intellectuels (sujet, description, langue)
- relatifs à la propriété intellectuelle

Ces éléments ont été normalisés sous le nom d'ISO Standard 15836.

[4] Wikipedia, consulté le 4/10/2012.

Pour faciliter l'échange des actualités, le secteur de la presse a développé ses propres vocabulaires contrôlés et ses formats d'échange au sein du consortium de l'IPTC (*International Press Telecommunication Council*)[5]. L'IPTC est une organisation internationale initialement créée dans les années soixante pour promouvoir des standards d'échange à destination de la presse et réunissant les principales agences de presse ainsi que les principaux éditeurs de journaux et de fournisseurs d'information.

Les standards IPTC[6] résultent de la volonté de définir des éléments de données permettant l'identification et la description des ressources. Ils sont utilisés par les agences de presse et les journaux. Les premiers formats et langages XML étaient NIFT (News Industry Text Format) et NewsML (multi-media news).

L'IPTC recommande aujourd'hui le standard de métadonnées spécifiques (appelées également sémantiques), IPTC Core[7], représentant les principaux champs d'annotation pour la description d'un fichier (text, image, vidéo) qui sont les suivants[8] :

- Titre principal
- Titre
- Domaine
- Date de création
- Créateur
- Fournisseur
- Source
- Copyright
- Identification de tâche
- Droits et conditions d'utilisation
- Emplacement
- Ville
- Département/Province
- Pays
- Code pays
- Instructions
- Auteur de la description
- Mots-clés
- Scènes
- Codes sujet
- Description

Avec les informations de plus en plus multimédia, l'IPTC a développé l'architecture NAR G2. NAR qui définit quatre objets principaux (newsItem, packageItem, conceptItem et knowledgeItem). Une nouvelle génération de standards de l'IPTC a été élaborée pour décrire les actualités (NewsML G2), les événements (EventsML G2), les actualités sportives (SportsML G2) ou l'échange international de données (IIM).

---

[5] Wikipedia, consulté le 4/9/2012.
[6] Site web IPTC, consulté le 4/9/2012.
[7] IPTC Core est fondé sur le schéma XMP, qui s'appuie sur les 15 champs de Dublin Core et a été développé par Adobe en 2001. IPTC Core XMP ainsi succède aux métadonnées IPTC/IIM, devenues obsolètes. In *Journée d'étude audiovisuel ADBS*, Arlette Boulogne. Documentaliste, 2006/2-vol.43, p. 136-142.
[8] Source : « Définitions des annotations IPTC », consulté le 5/10/2012.

IPTC G2 Family permet depuis 2008 de maintenir les standards basés sur XML ou de passer au web de données.

Les vocabulaires contrôlés de l'IPTC sont appelés *NewsCodes*. Ces *NewsCodes* permettent de décrire des articles, des photos et autres médias. Ils sont définis en 36 vocabulaires contrôlés dont le thésaurus *SubjectCodes* qui contient à peu près 1300 termes. [35]

Les standards IPTC aident à harmoniser la description des documents de différentes natures (texte, image, vidéo) à travers l'indexation de ces documents.

2.2 Analyser et structurer le contenu via l'indexation

Les classifications bibliographiques ont été confrontées dans les années 1950 à l'explosion de l'information. Elles ont subi des mutations et ont fait naître d'autres formes de vocabulaires comme les langages d'indexation ou des langages combinatoires. Ces derniers « offrent une solution plus souple, plus simple, et moins coûteuse pour analyser le contenu documentaire »[9]. Il s'agit d'une nouvelle forme d'organisation du contenu.

Selon le vocabulaire proposé par ADBS, l'indexation est un « 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 (ressources, collection) ou d'une question, en vue d'en faciliter la recherche, après les avoir identifiées par l'analyse.[10] »

L'indexation documentaire est une sorte d'attribution, car on attribue à un document des mots-clés, des thèmes, des concepts. Elle peut être libre ou contrôlée.

Depuis quelques années, on assiste sur le web à l'usage de plus en plus fréquent de **l'indexation libre** et ceci surtout sous forme de folksonomie. Le terme de folksonomie vient de l'anglais *folksonomy* combinant les mots *folk* (peuple, gens) et *taxonomy* (taxonomie). L'indexation collaborative libre, développée par exemple sur le site del.icio.us, n'est pas considérée comme une indexation structurée bien qu'elle utilise les techniques documentaires. Elle ne repose pas sur un langage documentaire. Par conséquent, l'indexation libre sort du cadre de la problématique développée dans ce mémoire[11].

L'indexation contrôlée est basée sur un vocabulaire contrôlé ou sur un langage documentaire. C'est également le cas au service de la documentation de *Libération* qui s'appuie sur un thésaurus. L'indexation des documentalistes à *Libération* est une opération intellectuelle (l'analyse) visant à repérer les thèmes et les idées principaux pour décrire et/ou résumer un article.

Du point de vue informatique, « l'indexation est utilisée pour tout champ associé à un index[12] » et on en parle souvent en termes d'**indexation automatique** où « les mots conservés dans l'index peuvent être tous les termes contenus dans le texte, sauf les mots vides (indexation dérivée), une sélection automatique de mots ou de termes extraits du texte, une sélection automatique de mots sur la base d'un langage documentaire (on parle alors d'**indexation par assignation**).[13] »

---

[9] [6, Maniez, 1987, p. 34]
[10] ADBS, consulté le 3/10/2012.
[11] En revanche, ce point a été développé par Odile Quesnel et Elie Francis [14].
[12] ADBS, consulté le 3/102012.
[13] ADBS, consulté le 5/10/2012.

L'intérêt d'utiliser à la fois l'indexation humaine et l'indexation automatique sera développé plus loin (voir p. 71).

Les langages d'indexation sont des langages documentaires qui servent d'outil de description pour structurer des documents. Cette structuration, effectuée grâce à l'indexation, peut être soit humaine et manuelle, soit automatique et effectuée par les machines.

Les langages documentaires permettent ainsi à l'utilisateur de se retrouver dans une masse de documents et de gérer des volumétries importantes. Ils permettent de classifier le savoir et/ou les connaissances. Ce sont avant tout des outils de travail dont les deux fonctions principales sont l'indexation et la recherche. Un langage documentaire doit correspondre et être continuellement adapté au corpus qu'il est censé représenter. Il n'est pas figé et évolue en fonction des besoins du corpus de documents.

Pour mieux cerner la nécessité des langages documentaires, il est utile de rappeler le(s) rôle(s) des vocabulaires contrôlés en général.

3 Le rôle des vocabulaires contrôlés

Les vocabulaires contrôlés, par opposition au langage libre, ont, bien avant Internet et le web, prouvé leur utilité dans la recherche par sujets et de façon générale, dans la gestion de l'information. Les professionnels de l'information savent que les sites web doivent être structurés car avec des centaines, voire des milliers d'informations nouvelles tous les jours, l'approche purement informatique ou éditoriale ne suffit plus. Pour construire et entretenir une navigation claire, logique et en même temps rapide et intuitive, les documentalistes continuent d'avoir recours aux vocabulaires contrôlés ainsi qu'aux langages documentaires. Je rappelle ici, brièvement, quelques notions liées aux vocabulaires contrôlés en précisant celui du thésaurus, parfois comparé à l'ontologie, afin de pointer d'une part, les points communs qui peuvent exister entre les différents vocabulaires contrôlés, et d'autre part, leur aspect évolutif en fonction de leur usage et leur environnement direct.

3.1 Notions de base

Vocabulaire contrôlé

Il s'agit d'un ensemble de concepts représentés par des termes reconnus et validés par un groupe et utilisés pour indexer les contenus et pour rechercher l'information. [1] [6]

En ce sens, l'une des façons les plus simples d'accès à l'information est l'index. L'index est une liste ordonnée de termes comme ceux d'un livre. Les termes sont accompagnés d'une référence (le numéro de la page, par exemple) permettant de retrouver l'information.

Langage documentaire

« Le langage documentaire est un langage artificiel qui fournit une représentation formalisée et univoque des documents d'un corpus et des questions qui intéressent un groupe d'usagers, afin de permettre le repérage simple des documents du corpus qui répondent aux questions de ces usagers. »[14]

Le langage documentaire sert à établir une communication entre le documentaliste et l'utilisateur dans la mesure où « il sert d'intermédiaire entre l'auteur du document et ses lecteurs éventuels [15]». En quelque sorte, il sert, d'après Maniez, à « coder » l'information.

Un langage documentaire est une forme plus complexe de vocabulaire contrôlé.

« On distingue les langages combinatoires ou postcoordonnés dont les éléments peuvent être combinés entre eux, *a posteriori* lors de l'indexation ou de la recherche, et les langages précoordonnés contenant des combinaisons de notions établies *a priori*.[16]» Les derniers étant des classifications (déjà évoquées) ou des listes de vedettes-matières.

---

[14] Philippe Lefèvre, La recherche d'information, du texte intégral au thésaurus, in [1, Rabault]
[15] [6, Maniez, p. 238]
[16] ADBS, consulté le 5/10/2012.

3.2 Le thésaurus et les ontologies

Le thésaurus, né dans les années cinquante, est un langage combinatoire. C'est une « liste organisée de termes 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) exprimées par des signes conventionnels (…) On peut 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).[17]»

Le thésaurus est considéré comme le langage documentaire le plus perfectionné « dont l'ambition est de sélectionner un terme unique pour désigner un concept (…). L'usage d'un thésaurus autorise l'indexeur à utiliser autant de descripteurs qu'il le souhaite (en tenant compte des conventions fixées par l'entreprise) »[18].

Un autre type de vocabulaire contrôlé, souvent comparé au thésaurus, est l'ontologie. A l'origine, l'ontologie est une discipline de la philosophie étudiant, de façon globale, la nature et l'organisation de l'être. Aujourd'hui, l'ontologie est surtout liée au développement, à la conceptualisation et à la modélisation de différents domaines de l'activité humaine dans le contexte informatique et s'inscrit dans la modélisation des connaissances, notamment à travers les technologies du web sémantique. Ce n'est donc pas un langage documentaire au même niveau que le thésaurus. [24]

Pour une rapide comparaison avec le thésaurus, voici les mots de Yolla Polity, cités par Jacques Chaumier :

« Les ressemblances entre un thésaurus et une ontologie sont frappantes. Dans les deux cas, il s'agit d'un vocabulaire contrôlé, utilisé et validé par les acteurs d'un domaine. Dans les deux cas, ce vocabulaire est structuré et doté de relations sémantiques entre les termes qui le composent. Mais les ressemblances s'arrêtent là car la sémantique des objets et des relations dans une ontologie est une sémantique formelle qui n'est pas destinée à être interprétée par des êtres humains […]. Leur caractère formel les rend aptes à alimenter des traitements et des raisonnements menés par des automates.»[19]

Les ontologies doivent leur développement depuis les années quatre-vingt-dix à l'informatique et au web. Le développement des ontologies demande des moyens importants :

« Le chantier de construction d'ontologies est ouvert mais il pose de sérieux problèmes dont celui du caractère prohibitif des coûts et des délais de mise au point d'une ontologie couvrant ne serait-ce qu'un champ spécifique d'un secteur industriel, médical ou scientifique.[20] »

Contrairement à l'ontologie, le thésaurus est un langage indépendant du contexte informatique, d'Internet et du web. Sa structure répondait, bien avant l'informatique et le web, aux besoins d'un grand nombre de services de la documentation, notamment ceux des titres de presse en vue de structurer les articles publiés.

---

[17] ADBS, consulté le 5/10/2012.
[18] [32, Zaclad, p. 137]
[19] [9, Chaumier, p. 83]
[20] Citation de Yolla Polity, in [9, Chaumier, p. 83].

Pour mieux comprendre le rôle du thésaurus dans la structuration de la base de données de *Libération*, il est d'abord important de décrire les points essentiels de l'activité documentaire et particulièrement celle de l'indexation telle qu'elle est pratiquée par les documentalistes du journal.

Deuxième partie
Activité d'indexation à *Libération*

Avant toute description et analyse du vocabulaire contrôlé existant, il me semble ici important de noter que le journal a subi plusieurs plans sociaux dont le dernier, en 2006-2007, a sensiblement touché le Service de la Documentation; le nombre de documentalistes étant passé de 15 à 6 personnes (dont 1 personne à mi-temps) ! Cette réduction des effectifs explique, en partie, l'état du principal outil des documentalistes dans l'activité d'indexation, qui est le thésaurus, et en particulier l'absence de sa mise à jour depuis 2007.

1 Pratiques adoptées

J'ai mené une enquête auprès des documentalistes pour comprendre la(les) façon(s) dont ils indexent des articles du journal afin de rendre compte des points communs et des divergences dans cette activité. Les interviews ont été construits autour des questions suivantes :

- Combien d'articles en moyenne indexez-vous par jour ?
- Quelle partie (champs sémantiques/domaines) du thésaurus consultez-vous le plus et quelle partie le moins ?
- Pourriez-vous décrire brièvement votre technique d'indexation ? Indexez-vous au terme le plus précis ou plutôt avec une indexation « profonde » ? (Par exemple, pour un article parlant du Parti socialiste, utilisez-vous : politique – parti politique – Parti socialiste ou simplement Parti socialiste ?)
- Combien de descripteurs utilisez-vous au minimum et combien de descripteurs au maximum pour indexer un article ? Et en fonction de quels critères ?
- Posez-vous souvent les termes librement (absents du thésaurus) et quels sont ces termes (exemples) ? Correspondent-ils aux partis politiques, aux noms propres (personnalités + entreprises) ou aux noms communs... ?
- Quelle est l'utilisation des mots-outils ?
- Quelles sont pour vous les situations les plus complexes dans l'indexation ?
- Les journalistes vous interrogent-ils par rapport à la façon dont on peut chercher par mots-clés ? Et vous-même, effectuez-vous souvent la recherche par mots-clés ?
- Avez-vous d'autres remarques par rapport aux incohérences du thésaurus, des concepts manquants ou inutiles… ?

Avant de rentrer dans les détails, les réponses obtenues m'ont tout d'abord permis d'avoir des données primaires et d'ordre général :

Les cinq documentalistes ont tous adopté à peu près la même façon d'indexer malgré quelques écarts. Ils indexent chacun en moyenne entre 20 et 30 articles par jour et consacrent à l'indexation entre 2 et 3 heures par jour, c'est-à-dire que l'indexation demande au quotidien à peu près 10 à 15 heures de travail. Chacun indexe « sa » rubrique du journal, mais ils peuvent indexer d'autres rubriques en cas d'absence de l'un des collègues.

Ensuite, j'ai traité et regroupé leurs réponses autour de cinq points suivants :

1.1 Au terme le plus précis OU en profondeur

La plupart du temps, les documentalistes articulent les deux techniques. Le choix de la technique de l'indexation dépend du sujet de l'article, du domaine et… du point de vue de chacun.

Pourtant, dans l'ensemble, la tendance est plutôt à l'indexation au terme (concept) le plus précis. Pour exemple, ils posent directement « vodka » sans poser « boisson alcoolisée »,

bien que ce soit le terme générique de « vodka »… A noter toutefois que tous ne sont pas unanimes sur ce choix, et dans ce type de cas

Les documentalistes ont une tendance à poser des termes génériques comme « musique », « cinéma » ou « danse » dès que l'article traite de l'un de ces sujets. Ensuite, ils affinent en fonction des termes censés représenter le contenu de l'article. Mais cette indexation de haut niveau est finalement une forme de classement qui leur garantit de mieux retrouver tous les articles sur la musique, le cinéma, etc. En revanche, elle est également source de bruit. Mais en règle générale, les documentalistes ne posent jamais les descripteurs de tous les niveaux de la hiérarchisation. Par exemple, s'il est question de l'immunité diplomatique, ils ne poseront pas tous les termes des niveaux précédant le terme « immunité diplomatique » :

- diplomatie
    - relations diplomatiques
        - diplomate
            - immunité diplomatique

Le choix des autres descripteurs dépendra du sujet de l'article, bien sûr, mais par rapport à un article précis, elle résultera du point de vue de chaque documentaliste. C'est d'ailleurs essentiellement sur cette notion de « points de vue subjectifs » que repose l'hétérogénéité de l'indexation.

1.2 Nombre de descripteurs

L'hétérogénéité dans l'indexation impacte également le nombre de descripteurs qui n'est jamais identique. La plupart du temps, le nombre varie d'un article à l'autre en fonction de la complexité du sujet, mais aussi en fonction des personnes. Les documentalistes n'ont pas de règles établissant le minimum ou le maximum de descripteurs à poser. Certains ont tendance à poser entre 5 et 8 descripteurs, certains en posent rarement plus que 4. Néanmoins, ils affirment qu'ils posent au minimum 2 descripteurs et au maximum 10.

1.3 Mots-outils

Tous les documentalistes, sans exception, utilisent très souvent des mots-outils, car ils les aident à bien résumer et à compléter l'article. En revanche, l'utilisation « à la recherche » est plus rare. Les documentalistes utilisent les mots-outils comme n'importe quel autre terme du thésaurus.

1.4 Utilisation des termes « libres »

Tous les documentalistes posent des termes librement, c'est-à-dire des termes absents du thésaurus. Ils affirment tous qu'il s'agit avant tout de noms propres (surtout nom de personnalités et d'entreprises). Quand il s'agit d'une personnalité, ils la posent tous dans

l'ordre suivant : prénom + nom ; jamais dans l'ordre nom + prénom. A défaut de la gestion des candidats descripteurs (via le champ « candidats descripteurs »), aucune gestion ni aucune organisation rationnelle des termes nouveaux ne peuvent être envisagées. Ce problème est la source de plusieurs formes orthographiques du même terme, ce qui peut ensuite considérablement compliquer la recherche. Une personne affirme avoir posé également des noms communs (« installation », « performance »).

1.5 Changement du nom

Quand un nom propre change (c'est souvent le cas d'une entreprise), les documentalistes ne procèdent pas tous de la même façon. Certains utilisent pendant un temps les deux noms, d'autres posent directement et seulement le nouveau.

Exemple :

Compagnie Générale des Eaux > Vivendi.

Remarque :

Ce problème pourrait être résolu grâce au logiciel de gestion du thésaurus qui garderait l'ancien nom en tant que candidat descripteur, ce qui permettrait de retrouver l'article indexé avec l'ancien nom, ou bien que celui-ci soit identifié en tant que terme non préférentiel (relation d'équivalence/synonymie).

2 Impact et manque de fonctionnalités

2.1 Fonctionnalité de suggestion de termes (appelée « je suis curieux »)

Cette fonctionnalité automatique suggère des descripteurs pour l'indexation. Elle est basée sur l'occurrence de mots utilisés par rapport au contenu de l'article. La plupart des documentalistes s'en méfient et certains ne s'en servent jamais… mais pas tous ! Cette fonctionnalité peut également suggérer des termes absents du thésaurus, ce qui est, encore une fois, source de plusieurs formes orthographiques pour un seul descripteur. La présence de ce terme n'incite pas à chercher le concept pertinent qui peut exister dans le thésaurus.

2.2 Absence de gestion du thésaurus

Il est impossible de suggérer rapidement la création d'un nouveau terme lié à un événement important. C'est le cas, par exemple, du « 11 septembre 2001 ». Pendant assez longtemps après l'événement, les documentalistes indexaient avec « Etats-Unis », « New York », « Centre d'affaires », avant de créer le terme « 11 septembre 2001 ».

Du fait qu'aujourd'hui le thésaurus ne bénéficie plus d'aucune gestion et n'est adossé à aucun logiciel de gestion du thésaurus (aucune navigation possible dans le thésaurus; absence de relations d'équivalence et associatives; absence de gestion de candidats descripteurs; le fichier PDF déconnecté de la base de données, voir p. 44), les documentalistes ne sont tout d'abord pas aidés dans l'indexation malgré le fait qu'ils la soignent au mieux.

Les documentalistes essaient de synthétiser et de résumer les articles de la façon la plus parfaite possible. En revanche, une telle démarche, sans aucune aide non plus à la recherche (aucune suggestion de descripteurs, avec une navigation très limitée dans le fichier PDF de presque 300 pages), ne peut pas exploiter l'indexation de manière optimale. Cette indexation permet inévitablement de poser certains descripteurs librement (avec différentes formes orthographiques, fautes d'orthographe... !), elle n'interdit donc pas des termes absents du thésaurus, car on ne dispose pas de suggestions de descripteurs associés, de synonymes et de la fonction d'autopostage (déploiement des termes voisins). Le minimum de gestion informatique de cet outil principal servant à indexer et structurer la base de données du journal fait défaut.

2.3 Objectif nouveau

Une mauvaise, voire une inexistante, gestion du thésaurus a naturellement des répercussions sur l'indexation elle-même. L'indexation peut s'éloigner de son but, qui est de faciliter la recherche et précisément de permettre la recherche par sujets. La recherche par sujets est en back-office du site de *Libération* traduite, via le formulaire de recherche, par mots-clés. Les mots-clés correspondent aux descripteurs du thésaurus.

En matière de recherche, les documentalistes ne suivent pas tous la même logique. Certains disent effectuer la recherche souvent par mots-clés, mais certains n'y pensent pas immédiatement et la font tout d'abord par auteur, date, titre, voire en texte intégral même s'ils connaissent le sujet. Ce qui tend à montrer qu'ils se sentent avant tout concernés par l'indexation et ne cherchent pas toujours à en profiter au moment de la recherche. Mais cette sorte de réticence a sa raison d'être qui est étroitement liée au manque de fonctionnalité informatique (ici l'autopostage) évoqué plus haut.

Il est primordial de rappeler que les documentalistes sont aujourd'hui dans la logique de l'activité documentaire traditionnelle, liée uniquement au cadre du service de la documentation, et que cette logique ne pourra pas être la même dans l'usage d'un référentiel sur le site web. Jusqu'ici, ils indexaient pour pouvoir retrouver par eux-mêmes. La prochaine étape serait une optimisation, voire une orientation fine de l'indexation automatique (voir p. 71), pour que les internautes trouvent et naviguent sur le site ! Cette démarche diffère de l'actuelle démarche dans la mesure où l'indexation ne viserait plus uniquement la possibilité de retrouver les articles, mais elle valoriserait l'ensemble des contenus produits par les journalistes sur le site web public. Dans cette perspective, c'est la démarche même des documentalistes qui se trouverait considérablement valorisée.

Mais pour cela, les indexeurs doivent tout d'abord disposer d'un outil de gestion documentaire adapté pour que leur référentiel (thésaurus, taxonomie, ontologie) puisse être relié à la base de données du site, qui lui-même sera relié à un moteur de recherche efficace.

Troisième partie
En quête de solution appropriée

1 Analyse des questions des utilisateurs du site d'information

La liste des questions les plus fréquemment posées par les internautes a été extraite à l'aide de la plateforme de surveillance XITI. Cette liste (voir les extraits en Annexes 1, p. 92) recouvre les requêtes des internautes sur une période de deux ans[21] (du 01/07 2010 au 30/06/2012). En fonction de la source des questions, XITI permet d'offrir deux listes :

1) **Liste « Moteurs de recherche externes (MRE) » :** la liste contenant les questions les plus fréquemment posées dans les moteurs de recherche externes (principalement Google mais aussi Yahoo ou Bing) et qui permettent d'atterrir sur le site de *Libération*.

2) **Liste « liberation.fr » :** la Liste comprenant les questions les plus fréquemment posées directement sur le site (via l'onglet Recherche).

1.1 Matière et méthode de l'analyse

Les deux listes contiennent 500 lignes[22]. Les listes sont constituées de questions dans l'ordre des occurrences (combien de fois la question a été posée), dont le nombre est indiqué pour chaque ligne. Il y a effectivement un écart important entre la 1re et la dernière ligne : 17 171 visites et 364 visites pour la liste « liberation.fr », et 223 174 visites et 3 327 visites pour la liste « Moteurs de recherche externes »

Certaines questions sont identiques au niveau du sens mais diffèrent par la forme orthographique ou par la formulation utilisée, par exemple dsk/strauss kahn/dominique strauss kahn. Cela veut dire que le même nom de la même personnalité peut occuper plusieurs lignes. Les 500 lignes ne contiennent donc pas 500 questions ; elles contiennent 500 termes.

En supervisant les deux listes, on s'aperçoit que l'écrasante majorité des questions concerne les noms propres constituant trois grands groupes :

- personnalités
- pays et lieux
- entreprises et organisations

[21] Le choix de la période dépend de la variété d'événements qu'on souhaite englober dans un espace de temps suffisamment récent : le choix de la période très courte risque de réduire la liste à très peu d'événements alors que la période trop longue ne rend pas compte des intérêts actuels des lecteurs.
[22] Le choix d'extraire précisément 500 lignes résulte du conseil de l'informaticien selon lequel ce nombre donnerait une vision suffisamment exhaustive des questions des utilisateurs.

Les noms communs sont bien moins nombreux et n'associent que très peu d'occurrences par rapport aux noms propres. Les listes n'offrent pas la distinction entre majuscules et minuscules; elles ne contiennent que des termes en caractères minuscules.

En comparant les deux listes, on s'aperçoit rapidement que les requêtes de la liste « MRE », permettant d'atterrir sur le site de *Libération*, portent, en grande partie, sur les actualités liées aux personnalités plutôt qu'aux lieux. Sur les 100 premières requêtes dans les « MRE », on enregistre 35 sur des personnalités et 15 sur des lieux. Les requêtes de la liste « libération.fr » concernent davantage les actualités associées aux lieux, qu'aux personnalités. Sur les 100 premières lignes, on compte 14 personnalités et 27 lieux.

Pour aboutir à des résultats approfondis et détaillés, l'analyse devra porter sur l'ensemble des requêtes de chaque liste et éventuellement être élargie au-delà des 500 lignes extraites ici, malgré le fait que les occurrences baissent ensuite considérablement, et ceci particulièrement pour les questions posées sur le site. En revanche, les questions posées sur les moteurs de recherche externes représentent des occurrences suffisamment élevées pour permettre au gestionnaire du site de repérer les sujets ou thématiques qui ramènent les utilisateurs sur le site du journal.

Dans une analyse plus approfondie de chaque liste, je me suis limitée, au cours de ce mémoire, à recenser uniquement les 30 requêtes les plus souvent posées pour chacun des trois groupes, ce qui donne **une toute première idée liée à l'intérêt, la curiosité et le besoin en information des utilisateurs du site.**

Le chiffre entre parenthèses correspond à la position du terme dans la liste respective (avec 500 termes confondus).

1.2 Requêtes « Moteurs de recherche externes »

30 noms de personnalités les plus posés :

1	dsk (1) / strauss kahn (65) / dominique strauss kahn (66)	11	mohamed merah (44) / merah (80)	21	nicolas sarkozy (73)
2	marine le pen (3) / le pen (35) / marina (170)	12	valérie trierweiler (49)	22	federer (75)
3	mélenchon (4) / melenchon (16) / jean luc melenchon (229)	13	tristane banon (50)	23	david carradine (77)
4	hollande (11) /françois hollande (33)	14	carla bruni (54)	24	marco simoncelli (80)
5	kadhafi (12)	15	anne sinclair (67)	25	leila trabelsi (85)
6	bayrou (25)	16	tony vairelles (68)	26	celine bara (87)
7	ben laden (37)	17	astrid herrenschmidt (69)	27	vanessa paradis (105)
8	thierry rolland (40)	18	bernard giraudeau (70)	28	laurent gbagbo (106)
9	segolene royal (41) / ségolène royal (79)	19	nicolas charrier (71)	29	woerth (109)
10	ben ali (42)	20	eric zemmour (72) / zemmour (86)	30	morano (113)

30 noms de pays et lieux les plus posés :

1	tunisie (6)	11	egypte (45)	21	italie (222)
2	syrie (9)	12	grece (56) / grèce (185)	22	lybie (255)
3	iran (10)	13	israel (59)	23	congo (263)
4	japon (18) /japon nucléaire (49) / japon séisme (230) / séisme au japon (288)	14	afghanistan (81)	24	bangkok (266)
5	fukushima (22)	15	corée du nord (118)	25	algérie (282)
6	toulouse (23) / toulouse tuerie (137) / fusillade toulouse (244)	16	chine (132)	26	nigeria (282)
7	maroc (27)	17	senegal (152)	27	pakistan (311)
8	turquie (30)	18	russie (159)	28	hongrie (318)
9	algerie (32)	19	thailande (186)	29	portugal (319)
10	mali (34)	20	palestine (213)	30	lyon (348)

30 noms d'entreprises, partis politique, organisations et institutions les plus posés :

1	psg (8)	11	concordia (99)	21	dexia (215)
2	free mobile (14) / free (29)	12	france inter (101)	22	Al jazeera (219)
3	facebook (20)	13	seafeance (104)	23	Ump (221)
4	fn (38) / front national (62)	14	charlie hebdo (106)	24	Petroplus (235)
5	air france (39) / greve air france (112)	15	bfmtv (124)	25	Megaupload (259)
6	le canard enchaine (46)	16	areva (147)	26	Lejaby (282)
7	pacitel (51)	17	skyrock (198)	27	Coupe de France (317)
8	institut pour la justice (52)	18	renault (199)	28	Ifop (321)
9	sncf (53)	19	airbus (202)	29	Google (323)
10	youtube (92)	20	om (204)	30	Apple (326)

Les formulations des requêtes

Les formulations des requêtes sont généralement soit des uni-mots (par exemple, électricité), soit des uni-termes (par exemple, gaz de schiste), mais aussi des expressions, voire des phrases. Les formulations plus longues font en général partie des suggestions de Google, c'est-à-dire de l'autocomplétion de Google qui est elle-même liée aux questions des utilisateurs.

Exemples d'uni-mots, c'est-à-dire posés en un seul mot *(le chiffre entre parenthèses indique la position dans la liste extraite par XITI)* :

syrie (3), smic (60), retraites (231), viol (292), économie (516), électricité (517)…

Exemples d'uni-termes et de formulations plus longues *(le chiffre entre parenthèses indique la position dans la liste extraite par XITI)* :

institut pour la justice (52)
présidentielle 2012 (61)
thierry roland est mort (64)
greve 12 octobre (76)
elections cantonales 2011 resultats (93)
tuerie toulouse (94)
mariage gris (95)
trou du cul (98)
nouveau gouvernement (100)
tueur de toulouse (110)
grève air france (112)
tuerie de toulouse (119)
action discrete (120)
élection présidentielle (133)
prime à la casse (134)
concombre contaminé (138)
sondage présidentiel 2012 (162)
gaz de schiste (163)

Les formulations sont exprimées dans le langage naturel et directement liées au langage utilisé dans la presse et par le grand public.

1.3 Requêtes « liberation.fr »

30 noms de personnalités les plus posés :

1	guillon (1)	11	morano (86)	21	christophe ayad (160)
2	Dsk (6) / strauss kahn (163)	12	marine le pen (95) / le pen (165)	22	luc chatel (182) / chatel (183)
3	melanchon (12)/melenchon (128)	13	steve jobs (96)	23	villepin (187)
4	schneidermann (15) / schneiderman (178) /daniel schneidermann (197)	14	gueant (114)	24	jacky durand (188)
5	demorand (19) / nicolas demorand (166)	15	montebourg (115)	25	gbagbo (194)
6	berlusconi (22)	16	françois hollande (117)	26	michel serres (211)
7	hollande (25)	17	iacub (118)	27	willy le devin (212)
8	pierre-marcelle (28)	18	ben ali (131)	28	norman (213)
9	sarkozy (53)	19	segolene royal (136)	29	hessel (214)
10	bayrou (65)	20	carla bruni (159)	30	alain duhamel (217)

30 noms des pays et lieux les plus posés :

1	tunisie (3)	11	hongrie (37)	21	argentine (79)
2	maroc (6)	12	fukushima (38)	22	senegal (83)
3	grece (10)	13	quebec (42)	23	russie (87)
4	algerie (16)	14	iran (51)	24	rwanda (88)
5	japon (17)	15	italie (54)	25	cote d'ivoire (93)
6	syrie (18)	16	suisse (56)	26	mexique (94)
7	libye (20)	17	chine (64)	27	liban (107)
8	belgique (24)	18	haiti (69)	28	israel (113)
9	egypte (29)	19	lybie (70)	29	roumanie (116)
10	espagne (35)	20	mali (74)	30	mayotte (149)

30 noms d'entreprises, partis politiques, organisations et institutions les plus posés :

1	wikileaks (27)	11	apple (155)	21	acta (320)
2	dexia (32)	12	seafrance (164)	22	edf (325)
3	facebook (45)	13	merck (167)	23	societe generale (340)
4	free (50)	14	sncf (209)	24	msd (341)
5	servier (52)	15	fnac (226)	25	nkn (364)
6	google (84)	16	groupama (238)	26	npa (370)
7	air france (102)	17	carlton (241)	27	ratp (379)
8	renault (104)	18	ikea (254)	28	renault (390)
9	charlie hebdo (105)	19	areva (274)	29	pole-emploi (404)
10	megaupload (140)	20	skyrock (279)	30	lagardere (406)

Parmi les questions sur les personnalités, certaines sur le site portent naturellement aussi sur les noms des journalistes de *Libération*.

Rubriques

Les noms des rubriques ou des blogs du journal, dans l'ordre descendant, saisis dans l'onglet *Recherche* par les internautes sont les suivants *(le chiffre entre parenthèses indique la position dans la liste extraite par XITI)* :

1. Rebonds (3) : une des rubriques principales du journal consacrée surtout à l'actualité politique.

2. Portrait (5) : rubrique sur les portraits des personnalités connues ou moins connues.

3. 400 culs (37) : blog à succès sur le thème du sexe.

4. Météo (42)

5. Grand Angle (64) : rubrique consacrée aux divers phénomènes de société.

6. Désintox (131) : nom de la rubrique consacrée aux commentaires des paroles des politiques.

7. Editorial (135)

Remarque :

La rubrique Rebonds remonte dans la liste en premier, mais cette place est « faussée » par l'exemple de l'aide à la recherche, présent dans l'onglet Recherche : « ex. : presidentielles Rebonds ». L'atterrissage sur cette page ne prouve en aucun cas l'intérêt pour le contenu de la rubrique.

1.4 Observations générales et conclusion

Il n'y a pas de règles strictes et précises qui détermineraient la façon dont les internautes saisissent les noms propres des personnalités. En règle générale, il s'agit de la forme courante et très proche utilisée dans les médias, dans les débats informels et non officiels et à l'oral. La forme courte d'un nom propre est saisie plus souvent dans la mesure où elle est employée fréquemment dans les médias et la personnalité est très connue. Il s'agit la plupart du temps de personnalités politiques : Mélenchon, Hollande, DSK, Bayrou, Woerth, Morano.

On remarque également que les accents sont souvent ignorés par les internautes : Segolene Royal est largement plus utilisé que Ségolène Royal. De la même façon, les internautes recherchent avec « Melenchon » plutôt qu'avec « Mélenchon ». Les requêtes sont souvent posées en un seul mot quand il s'agit de noms propres.

Les moteurs de recherche externes proposent, en fonction de l'occurrence des requêtes posées par les internautes, des sujets/thèmes sous forme d'autocomplétion. Sur le site du journal, l'utilisateur, en posant sa requête, ne bénéficie pas de suggestions. Il n'est pas guidé dans sa recherche et ne peut pas l'affiner.

L'organisation du contenu de la base de données devra se concentrer tout d'abord autour de la politique actuelle (internationale et française). Par exemple, au moment où l'internaute pose comme requête « Grèce », il devra avoir le choix entre les articles sur « la crise grecque » ou « l'Union européenne », voire « l'Euro ». Il devrait pouvoir naviguer et chercher par personnalités politiques. De nombreuses requêtes liées aux entreprises mériteraient d'être soutenues et guidées par une information structurée dans ce domaine.

Les noms communs (ayant un sens très général et/ou du type des mots-outils du thésaurus) apparaissent donc peu dans les questions, mais au moment où l'on donne au visiteur la possibilité de naviguer par catégories du type « partis politiques », « politique intérieure », « politique étrangère », « économie politique », « littérature » ou « athlétisme », il y a de fortes chances qu'il se mette à explorer le site avec beaucoup plus d'intérêt.

2 Différentes voies

Il existe différentes options pour améliorer la recherche et la navigation sur le site de *Libération* mais qui ne sont pas, pour des raisons différentes, prioritaires, voire envisageables du point de vue de la faisabilité.

2.1 Recherche via formulaire

La recherche via un formulaire permettra d'affiner la recherche en s'appuyant sur la logique de la recherche booléenne (ET, OU, SAUF), et en sélectionnant les catégories sur lesquelles la recherche porte (titre, auteur, date). La recherche par formulaire est aujourd'hui possible sur le back-office du site pour les besoins internes du journal à travers les filtres suivants :

- papier (articles publiés dans le journal papier)
- web (articles publiés seulement sur le site web)
- tout

Seul, filtre par « papier » bénéficie de la recherche par mots-clés, car les articles du web ne sont pas indexés.

En revanche, la recherche par formulaires sur le site web du journal est, malgré sa pertinence et son efficacité, peu adaptée à un site d'actualité grand public, car elle peut paraître beaucoup trop complexe, trop longue et peu intuitive pour les utilisateurs du site. Enfin, pour une cohérence et un équilibre par rapport à la recherche par sujets, l'indexation devrait être effectuée pour tous les articles, y compris ceux du web. Mais ceci est valable pour tout autre type de recherches.

2.2 Topic Maps

Tout au début de mon stage, il a été question de développer un référentiel semblable à celui du site *The Guardian* dont l'architecte (professionnel de l'information) a rencontré les responsables informatique et documentaliste. En se renseignant de plus près sur le développement et la construction des cartes topiques en général [30] et en observant d'autres sites de la presse écrite (anglo-saxons et français), on constate qu'une telle démarche serait difficilement réalisable à *Libération* à partir du thésaurus existant.

La construction d'une carte topique, telle qu'elle a été développée par *The Guardian*, suppose une conception bien plus complexe que celle du thésaurus. Les cartes topiques sont en quelque sorte des index (ressemblants à ceux d'un livre) adaptés à l'environnement informatique, contrôlés et reliés entre eux. La structure est proche des ontologies dans la

mesure où une carte topique vise à englober les connaissances de tout un domaine. On s'aperçoit que beaucoup de bases de connaissances dans des sites anglais sont aujourd'hui construites autour des index, comme c'est le cas de *The Guardian*. En revanche, aucun site français – en tout cas, ceux de la presse française - ne fonctionne ainsi. Ces index sont développés autour de topics, sous-topics, etc. qui sont tous reliés entre eux par les associations, les sujets et les occurrences (Cf. Wikipédia). Le développement d'une telle base de connaissances nécessite plusieurs mois, voire plusieurs années, sans parler de la nécessité d'une collaboration étroite entre les professionnels de l'information, les informaticiens et les rédacteurs.

Le thésaurus de *Libération* n'a pas (ou n'a plus), comme on le verra plus loin (voir en page 44), d'associations qui relieraient les différents termes entre eux. En outre, pour optimiser la construction et la visualisation d'une carte topique, il est préconisé de se doter d'un logiciel approprié (par exemple *Ontopia* ou *ITM* de Mondeca). Un tel référentiel demande également une mise à jour quotidienne et un suivi permanent avec l'ensemble des rédacteurs, le tout administré par un professionnel de l'information.

3 Recherche et navigation à facettes

3.1 Principes

La recherche à facettes n'est pas une réelle méthode de recherche au sens documentaire du terme, mais plutôt une aide à la recherche. Elle se distingue de la recherche par formulaires qui est une recherche par critères. Les facettes sont des axes de recherche (par date, auteur, titre, sujet…). En principe, la recherche à facettes fait appel à un référentiel qui est élaboré et modélisé. Ce référentiel peut être plus large que le thésaurus dans la mesure où il inclut également les dates ou le support (web ou papier, par exemple).

Le principe de la recherche à facettes permet à l'utilisateur d'affiner sa requête grâce aux suggestions sous forme de mots-clés. Par exemple, avec la requête « François Hollande », il devra pouvoir associer et continuer sa recherche avec les thèmes :

- chef de l'état
- parti socialiste
- politique intérieure
- politique étrangère

Ces autres termes sont aussi des descripteurs faisant partie des résultats, au même titre que « François Hollande ».

En optant ensuite pour un terme, par exemple « Parti socialiste », le moteur de recherche ne fournira que les articles indexés avec ces deux descripteurs.

L'utilisateur pourra filtrer les résultats en fonction de ses besoins et affiner sa recherche. L'interface de recherche sera d'ailleurs développée et améliorée par rapport aux besoins spécifiques du site : l'internaute pourra ainsi associer « François Hollande » non seulement à un autre sujet comme ci-dessus, mais également à une personnalité (Nicolas Sarkozy, Ségolène Royale, Jacques Chirac…) ou à un événement (élections 2012, G20…).

C'est ce principe du filtrage par facettes relié à un développement IHM (interface homme-machine) ergonomique qui permet une navigation pertinente par mots-clés, directement à partir des articles eux-mêmes. Cette navigation peut considérablement optimiser la recherche et inciter la navigation sur le site. Les mots-clés peuvent être affichés dans la zone *ad hoc*. Ces mots-clés proposent de rebondir sur des sujets similaires. Ils permettent à l'utilisateur d'atterrir sur les pages traitant de sujets plus ou moins liés.

Par exemple, un article sur François Hollande dans lequel seront également cités des mots-clés comme « Nicolas Sarkozy » ou « G20 » (également termes du référentiel) offrira aux utilisateurs la possibilité de consulter les articles indexés avec ces descripteurs. Actuellement, il est possible, via un simple lien html (fait manuellement), de rebondir uniquement d'un article à un autre.

Le moteur de recherche pourra être associé à un référentiel, c'est-à-dire à un vocabulaire contrôlé. Dans ce cas, la fonction de ce référentiel sera d'abord liée à l'enrichissement du moteur de recherche pour mieux catégoriser des documents. La problématique du langage naturel (ambiguïté, par exemple) sera levée grâce aux documentalistes, et les articles seront

associés aux « bons » descripteurs et aux bonnes catégories. Il sera ensuite possible de proposer et/ou de publier ces descripteurs (mots-clés sur le site) représentant des contenus pertinents par rapport à la requête.

La recherche et la navigation à facettes offrent à l'internaute à la fois la possibilité de chercher sur un sujet indépendant de l'actualité du jour (et absent de la publication du jour) et la possibilité de naviguer et rebondir sur autre chose, à partir de l'article qu'il sera en train de lire. Le principe de recherche à facettes est fondé sur la même organisation du système d'information et sur la même gestion du contenu de la base de données, c'est-à-dire sur la classification à facettes. Le développement de la recherche/navigation à facettes n'est rien d'autre que la structuration de l'information et du contenu de la base de données par sujets. La recherche/ navigation par facettes peut s'appuyer sur plusieurs vocabulaires contrôlés (thésaurus, taxonomie, ontologie…). La question est de savoir quel vocabulaire contrôlé serait le plus adapté à la situation du quotidien, et plus particulièrement, au vocabulaire déjà en usage à *Libération*.

3.2 Vers la taxonomie de navigation

La recherche/navigation à facettes sur les sites web est aujourd'hui de plus en plus développée via la taxonomie de navigation. Quels sont les origines, les définitions et les principes fondamentaux d'une taxonomie de navigation ?

3.2.1 Taxonomie : définition

Le terme « taxonomie » vient de la botanique; la taxonomie permet de classer et de décrire des végétaux. [17]

Sur le plan documentaire et selon la (les) définition(s) du *Vocabulaire de la documentation*, publié par ADBS, « la portée du terme est finalement étendue à tout langage documentaire doté, exclusivement ou non, d'une organisation hiérarchique (…) D'un point de vue structurel, on parlera alors de taxonomies (de termes, de classes, de concepts) pour désigner 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. »[23]

La notion de taxonomie documentaire n'est donc pas simple car elle ne se limite pas à un seul type de langage documentaire, mais peut en représenter plusieurs. La taxonomie peut être basée et construite sur plusieurs formes de langages documentaires. La structure d'une taxonomie est celle d'un arbre, chaque terme dépend d'un autre et on peut créer autant de classes et sous-classes qu'on le souhaite.

Une taxonomie (en tant que vocabulaire contrôlé) relie les termes entre eux uniquement par les relations hiérarchiques, alors que dans un thésaurus, les termes sont reliés également par des relations d'équivalence et associatives.

[23] ADBS, consulté le 5/10/2012.

3.2.2 Un concept technologique lié au web

Le développement des taxonomies de navigation est étroitement lié à l'usage sur Internet et les sites web pour faciliter la recherche/navigation par sujets ou thèmes.

Moteur de recherche performant et CMS adapté

En principe, l'exploitation d'une taxonomie nécessite un moteur de recherche et un CMS (système de gestion de contenu) adéquats.

Le moteur de recherche actuel de *Libération* est un moteur de recherche « maison ». Il est développé à partir des technologies libres et ouvertes : un programme de base Python et un système pour stocker des données PostgreSQL permettant de gérer une base de données de plus de 700 000 articles (sur le front et en back-office). Le moteur de recherche possède les fonctionnalités les plus basiques de la recherche en texte intégral, mais *a priori* il ne permet pas de gérer une taxonomie de navigation de façon aussi avancée que les solutions adaptées (Mondeca, par exemple [19]).

Pour une gestion souple et une création de nouvelles facettes sur le site web, il est préconisé de se doter d'un CMS adapté, comme c'est par exemple le cas du CMS open source Drupal.

Il est à préciser que *Libération* s'est doté, en 2008, d'un moteur de recherche performant, Nstein, ainsi que du CMS Drupal[24]. Nstein n'a finalement jamais été installé et Drupal a été abandonné pour une solution maison « Djaz », basée sur un outil open source Django.

A côté du contexte purement documentaire, il reste à étudier dans quelle mesure l'environnement informatique actuel permettra de développer une taxonomie de navigation performante.

Remarque :

« Les taxonomies de navigation sont un outil dynamique au service de l'éditeur et doivent pouvoir évoluer rapidement en fonction de l'évolution des demandes et des attentes des utilisateurs. L'utilisation de logiciels spécialisés dans la gestion des taxonomies et référentiels, de moteurs de recherche spécialisés dans la recherche par taxonomie, et d'interfaces utilisateurs adaptées et ergonomiques est indispensable pour une gestion souple et évolutive des taxonomies de navigation dans les portails. » [19, Mondeca]

3.2.3 Fonctions et structure de la taxonomie sur le site

Les possibilités et les performances technologiques sont renforcées, perfectionnées et valorisées par un vocabulaire structuré (la taxonomie) et adapté aux besoins du site web.

L'une des fonctions de la taxonomie, et en même temps la fonction de base, restera la même que celle du thésaurus : celle d'un vocabulaire contrôlé. Elle permettra de contrôler les termes utilisés dans l'indexation des articles du journal. La présentation de ce nouveau vocabulaire contrôlé ne sera plus celle du thésaurus actuel et elle fera l'objet d'une modélisation et d'un développement documentaires, complexes et structurés.

---

[24] Source : « Quels systèmes de gestion de contenu (CMS) dans les rédactions web ? », publié le 11/04/2008, consulté le 5/10/2012.

Sa deuxième fonction (et la valeur ajoutée par rapport au thésaurus existant) sera l'exploitation en recherche par des utilisateurs du site (et non seulement par des documentalistes, comme c'est le cas aujourd'hui) et ceci à travers les facettes du front-office.

En indexant (de façon automatique et/ou humaine) les articles, à l'aide de la taxonomie, c'est-à-dire d'un vocabulaire contrôlé dont les termes seront répartis en catégories et sous-catégories, il sera ensuite possible d'offrir à l'internaute une recherche structurée et une navigation riche. La pertinence des articles « rangés » dans les bonnes classes dépendra de la pertinence de l'indexation.

Certaines facettes pourront néanmoins correspondre aux catégories de la taxonomie du back-office. Il sera par exemple possible de publier sur le front tous les articles indexés et classés sous la catégorie « Culture » de la taxonomie du back-office.

Mais la valeur ajoutée sera liée à la possibilité de filtrer la recherche et de trier les résultats. Par exemple, un utilisateur posant la question sur « gaz de schiste » devra avoir la possibilité de trier par date, pays et personnalités. Cela suppose la création de ces trois catégories qui elles-mêmes seront réparties en d'autres sous-catégories. Les articles traitant de la problématique du gaz de schiste en France seront indexés avec les termes « gaz de schiste » et « France » ; le terme « France » faisant partie de la catégorie « Pays ». L'utilisateur cherchant des articles sur Mélenchon devra pouvoir filtrer par exemple par date, personnalités, partis politiques ou événements politiques. Ces catégories feront partie de la taxonomie du back-office et seront publiées sur le front en fonction de la requête du lecteur.

3.2.4 Principe et création de sujets

La logique combinatoire permet donc de publier les documents indexés par un ensemble de termes représentant un sujet précis. Par exemple, les documents indexés à la fois par « crise » + « économie » + « Grèce » remonteront probablement tous les articles sur la « crise économique en Grèce », c'est-à-dire un sujet (ou un tag) pouvant être publié à proximité de l'article sur un sujet identique ou similaire. La présence d'un tel sujet incitera l'utilisateur à lire d'autres articles sur le sujet et augmentera son temps de navigation sur le site web.

Dans la même logique de création de sujets, certains termes de la taxonomie peuvent être corrigés, précisés ou affinés dans le back-office du site pour leur publication sur le front-office. Par exemple, comme dans un cas évoqué plus loin (p. 58), les articles indexés par « mutilation corporelle » pourront être consultables sous le nom de « mutilation rituelle ». Cette souplesse est nécessaire et confortable, mais il est tout de même préconisé de définir les termes de la taxonomie (vocabulaire contrôlé) de la façon la plus adaptée aux utilisateurs pour que, dans son ensemble, ce nouveau vocabulaire traduise au mieux les besoins du site web d'un quotidien d'actualité.

Le site Popline offre un bel exemple de taxonomie de sujets. Popline est la base de données bibliographique sur la population, la santé de la reproduction et la planification familiale. Elle contient plus de 300 000 références. La taxonomie de navigation s'appuie sur le thésaurus permettant d'indexer toutes les ressources. Le thésaurus Popline, d'ailleurs toujours consultable et autonome, permet de créer des « sujets » publiés sur le front-office. Comment ces sujets sont-ils créés ? La transformation en sujets s'opère grâce aux principes des profils de recherche. Les services de la DSI (diffusion sélective de l'information) utilisent, via des logiciels, des requêtes préprogrammées élaborées à partir des descripteurs du thésaurus. Le thésaurus Popline, avec 2 000 concepts regroupés par ordre alphabétique, ne

pouvait pas offrir un accès pratique aux ressources auprès des utilisateurs du site (et des utilisateurs en général). C'est seulement à travers la taxonomie, basée sur 12 catégories principales qui permettent ensuite d'aborder l'ensemble des ressources par sujets (à peu près 400), que les descripteurs du thésaurus, visibles d'ailleurs dans l'URL de la requête, ont pu trouver leur utilité dans la recherche sur le site. [25]

L'objet de recherche de notre étude est de trouver une structure de la taxonomie la plus adaptée au site web de *Libération*. Le premier pas vers la définition de ce nouveau vocabulaire contrôlé passe tout d'abord par l'évaluation de l'existant. C'est en grande partie l'analyse du thésaurus actuel qui permettra de mesurer sa reprise éventuelle dans la construction d'une taxonomie de navigation.

[25] Source : Sylvie Dalbin, *Dispositif d'accès à l'information et évolution des thésaurus – le cas de Popline*. Consulté le 9/10/2012.

Quatrième partie
Thésaurus de *Libération*

1 Présentation générale du thésaurus

1.1 Bref historique

Le thésaurus a été élaboré en 1995 par l'ensemble des documentalistes, et mis à jour au fur et à mesure en fonction des besoins de l'indexation et de la recherche. Les dernières modifications importantes, effectuées en particulier par l'ancienne responsable de la documentation, datent de 2006.

Le thésaurus a été intégré dans le logiciel de gestion et recherche documentaires Alexandrie de GB Concept, et ensuite dans le système NICA d'IBM. Les deux systèmes de gestion documentaire permettaient non seulement de gérer les candidats descripteurs (via le champ candidats descripteurs), mais ils géraient également les relations hiérarchiques et associatives du thésaurus. La navigation et la recherche ont été simplifiées grâce aux fonctions de l'autopostage (suggestions de termes à indexer en déployant l'arborescence autour du terme choisi).

1.2 Aujourd'hui

Depuis 2007, le thésaurus est ajouté au back-office du journal (actuellement appelé « Quai ») sous forme d'un simple fichier PDF. Seuls les documentalistes l'utilisent pour l'indexation. Il est donc indépendant de la base de données des articles archivés et en production, ce qui est aujourd'hui inutilisable pour le développement sur le site web du journal.

Le thésaurus est aujourd'hui un référentiel terminologique sans relations entre les termes, sans navigation possible entre les termes et avec une recherche limitée à la fonction « Rechercher » d'un fichier PDF de 284 pages.

Le fichier PDF est présenté comme une **liste alphabétique de termes regroupés par champs sémantiques (domaines).**

Les champs sémantiques sont regroupés par thèmes et non par facettes (nature des concepts). Les champs sémantiques contiennent uniquement les descripteurs **sans** :

- notes d'applications (définitions)
- non-descripteurs équivalents (EP = employé pour)
- termes associés (TA)
- indications des termes génériques (TG) et spécifiques (TS)

Il s'agit d'une simple liste de termes reliés entre eux par des relations hiérarchiques. Chaque niveau hiérarchique est marqué par un décalage vers la droite. Les termes à l'intérieur des domaines (ou des champs sémantiques ou des microthésaurus) sont classés par ordre alphabétique. Les champs sémantiques sont entre eux organisés également par ordre alphabétique.

A propos de la terminologie

La nouvelle norme ISO 25964 adapte les anciens termes aux besoins actuels. Même si la norme ISO 25964-1 n'existe aujourd'hui qu'en anglais, quelques présentations des professionnels d'information annoncent ces changements terminologiques (en français), confirmés dans les présentations[26] faites l'année de la publication de la norme, en 2011.

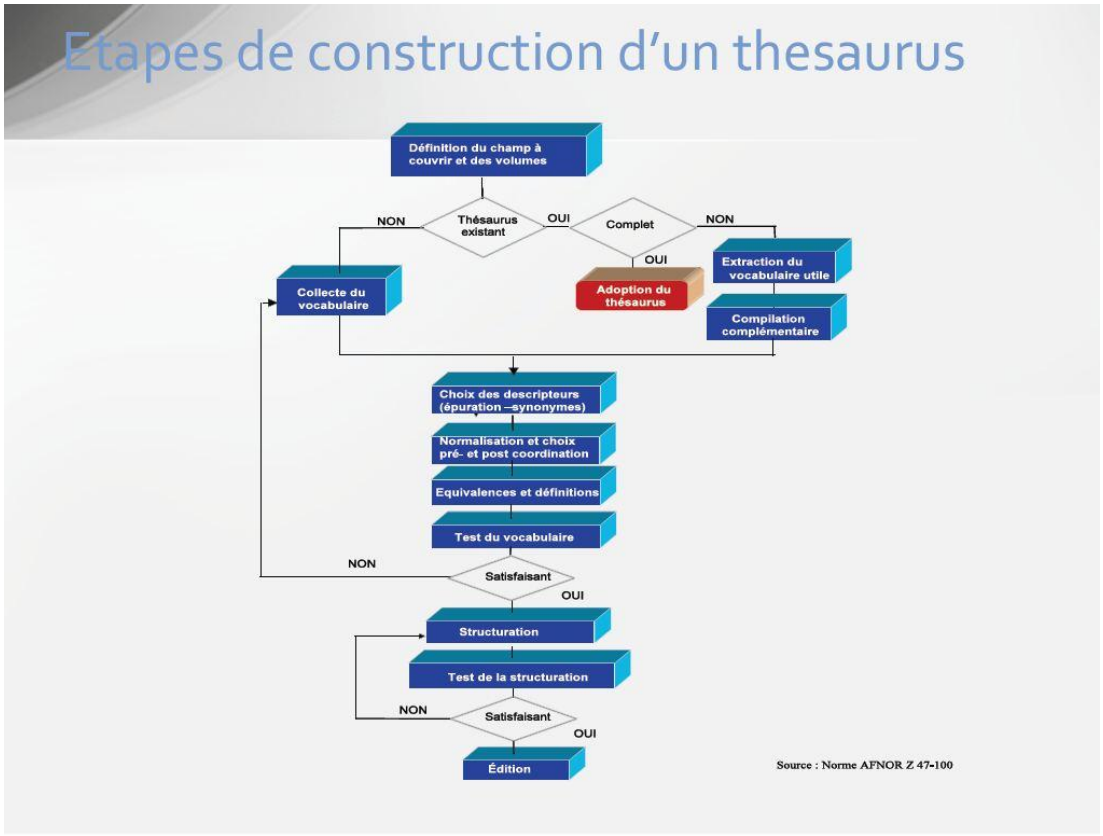
De ce fait, j'emploierai, dans l'analyse suivante, les termes nouveaux, et particulièrement : ***terme préférentiel*** pour descripteur, ***terme non préférentiel*** pour non-descripteurs et ***domaine*** pour champ sémantique.

[26] Par exemple, Norme ISO 25964, présentée le 15/3/2011. Consulté le 5/10/2012.

2 Analyse

2.1 Méthode à adopter

Le schéma[27] ci-dessous présente les étapes générales de la construction d'un thésaurus. Certes, dans le cas de *Libération*, il n'est pas question de reconstruire un nouveau thésaurus. Cette méthode ne sera donc pas adaptée à la construction elle-même d'une taxonomie de navigation. En revanche, le schéma donne une vision générale, même si très simplifiée, des aspects les plus importants d'un thésaurus. Pour l'analyse du thésaurus existant et pour pouvoir définir dans quelle mesure l'existant répond aux normes établies, certains points de ces étapes permettent d'orienter l'analyse.



Le schéma rappelle deux grands éléments qui sont ceux du vocabulaire et de la structuration. L'objectif de l'analyse est de repérer comment le thésaurus de *Libération* se situe par rapport à ces deux aspects.

[27] Schéma extrait du cours d'Hélène Rabault, [1].

Pour développer l'analyse ci-dessous plus en détail, je me suis essentiellement appuyée sur le cours de H. Rabault (le schéma en fait partie) qui lui-même s'inspire des normes destinées au thésaurus, en particulier ISO 25964[28].

C'est en fonction de ce cours et des premiers repérages de la présentation générale du thésaurus évoquée plus haut, que j'ai décidé de conduire l'analyse autour des points suivants :

- **Identification des domaines principaux**
- **Lexique**
- **Structure générale**
- **Absence des termes non préférentiels**
- **Mots outils**
- **Listes annexes**

2.2 Identification des domaines adoptés

En premier lieu, il s'est révélé nécessaire de parcourir l'ensemble du thésaurus pour faire ressortir les domaines principaux et repérer les anomalies. Cette étape de nettoyage a permis ensuite la réelle analyse de ce langage documentaire.

2.2.1 Repérage des doublons et des anomalies

Pour extraire la liste des domaines, j'ai été confrontée à une incohérence dans la présentation. Les non-descripteurs (ici les noms des domaines ou anciennement aussi les relais virtuels) sont, dans le thésaurus, précédés par « ZZZZ » (ils sont à peu près 20). En revanche, une partie des domaines n'est pas précédée de cette indication et est utilisée à l'indexation (ce qui est *a priori* contraire aux normes du thésaurus) : Agriculture, Aménagement du territoire, Faits divers, Politique... Le problème est lié à l'incohérence des concepts se trouvant en tête de hiérarchie. En principe, on peut, par exemple, utiliser à l'indexation « questions sociales », mais pas «questions financières ». Cette incohérence vient peut-être d'une démarche volontaire dans la construction du thésaurus, mais aussi d'une disparition et/ou d'un rajout des « ZZZZ » en tête de hiérarchie, dus à la mauvaise manipulation du fichier PDF. Dans la présente analyse, je considérerai tous les concepts en tête – précédés ou non par « ZZZZ » – comme étant au même niveau dans la hiérarchie.

Ci-dessous les domaines principaux, mots-outils et listes annexes confondus (avec numéro de la page entre parenthèses afin de faciliter l'orientation dans le fichier PDF) :

[28] Elle remplace la norme ISO 2788 de 1986. ISO 25964 est composée de deux parties : ISO 25964-1 (consacrée essentiellement au thésaurus et à sa construction ; publiée en août 2011) et ISO 25964-2 (consacrée à l'interopérabilité avec d'autres thésaurus et vocabulaires contrôlés ; aujourd'hui, elle est en cours de finalisation).

Important

[29] Le terme « classique » n'a aucune signification particulière, à part, probablement, la distinction du « Fonctionnement des organisations ».

Avant toute modification et modélisation, les doublons repérés (en rouge) devront être supprimés ! On constate également que la partie des mots-outils « Fonctionnement des organisations » n'est qu'un domaine à part.

2.3 Les domaines du thésaurus après « le nettoyage »

Ci-dessous la structure « nettoyée » (sans doublons) et répartie par domaines.

Domaines (31 actuellement) :

Agriculture (P 1)
Aménagement du territoire (P 2)
Civilisation (P 3)
Communication (P 5)
Culture (P 10)
Défense (P 15)
Economie (P 17)
Election (P 19)
Enseignement (P 20)
Entreprise (P 23)
Environnement (P 25)
Faits divers (P 25)
Fonctionnement des organisations (89)
Géographie (P 28)
Industrie (P 29)
Logement (P 31)
Loisir (P 32)
Matière première (P 33)
Organisme international (P 34)
Politique (P 34)
Questions sociales (P 37)
Questions financières (P 65)
Questions internationales (P 68)
Questions juridiques (P 70)
Science et technologie (P 73)
Religion (P 43)
Santé (P 45)
Secteur tertiaire (P 49)
Sport (P 50)
Transport (P 55)
Travail (P 56)

Listes

Partis et mouvements politiques (P 63)

- partis et mouvements politiques en France
- partis et mouvements politiques étrangers

Entreprises (P 76)

Personnalités (P 94)

Territoires étrangers (P 260)

Territoires français (P 273)

Mots-outils

- Mots outils classiques (P 90)

2.4 Lexique

Le lexique d'un thésaurus est évalué selon plusieurs points différents :

- dans son ensemble, en tant que vocabulaire représentant au mieux les contenus des documents ainsi que des questions posées par son public
- dans la façon dont ses termes sont présentés, c'est-à-dire selon la forme d'écriture
- dans un ensemble de concepts représentés par des termes

2.4.1 Mise à jour du vocabulaire

Avant de construire ou de mettre à jour le lexique du thésaurus, il est préconisé d'identifier des vocabulaires et des thésaurus existants et d'évaluer leur pertinence par rapport aux besoins du thésaurus à construire ou à mettre à jour [1]. Dans ce sens, H. Rabault rappelle les différents usages d'un thésaurus dont celui de « réservoir terminologique » ou de « liste de contrôle ». Tout vocabulaire structuré, et pas seulement les thésaurus, doit être mis à jour régulièrement et pour cela les référentiels existants sont une grande source d'inspiration.

2.4.2 Normalisation de la forme d'écriture

Les formes des termes recommandées sont :

- mot ou groupe de mots
- masculin (si le choix possible)
- singulier, sauf pour :
    - les mots au pluriel dans la langue française (archives)
    - les mots avec un sens différent au singulier ou au pluriel (fond/fonds; ciseau/ciseaux)
    - les termes en langue naturelle, avec les articles (droits de l'homme et non droit/homme)

Certaines règles d'écriture doivent également être adoptées et respectées au niveau de l'orthographe, du choix minuscule/majuscule, en fonction des besoins et du contexte.

Dans les choix des concepts et des termes, on évite les termes ambigus et on opte pour des termes en adéquation avec les questions des utilisateurs.

2.4.3 Termes : nombre et nature

Il est important d'évaluer le nombre total de termes d'un thésaurus et ensuite de distinguer les différentes natures de ces termes. Cette distinction concernera noms communs/noms propres [1].

Thésaurus de Libération

Du fait que le thésaurus n'a pas été mis à jour depuis 2006, il manque des termes, essentiellement des noms propres, surtout au niveau des organisations, des entreprises, des parties politiques ou des personnalités.

La forme des termes n'est pas, pour la plupart, contraire aux recommandations des normes. Les termes sont composés d'un seul mot ou d'un groupe de mots. La plupart sont au masculin et au singulier, et sont utilisés en langue naturelle.

On détecte néanmoins quelques incohérences, notamment au niveau de l'écriture des noms des organisations. Par exemple, l'écriture des partis politiques n'est pas unifiée ; elle est soit en développé, soit en sigle ; les deux (développé+sigle) étant rares.

D'après le responsable informatique, le thésaurus contient 24 000 termes dont 7 000 utilisés à l'indexation. Mais en étudiant le thésaurus de plus près, et en calculant tout simplement les termes du fichier PDF, nous arrivons approximativement à **18 300 termes** (y compris les listes annexes et les mots-outils), ce chiffre étant « décortiqué » plus loin.

Les domaines comptent approximativement **5650 termes**, principalement des noms communs malgré la présence de noms propres qui correspondent généralement aux noms des institutions et organismes.

Le nombre de noms propres est variable d'un domaine à l'autre. Par exemple, le domaine Politique ne contient que 8 noms propres, le domaine Economie 14, mais celui de Communication 160 (titres de presse, chaînes de télévision, radios…) et celui de Culture 100 (maisons d'édition, musées, institutions culturelles…). Le domaine Organisme international n'est finalement qu'une liste composée de 18 noms propres.

Les listes, constituées uniquement avec des noms propres, comptent 12 360 termes. La majeure partie des listes ne représente que des noms propres classés par ordre alphabétique.

En conclusion, on constate que le thésaurus compte environ :

- 910 noms communs
- 13 210 noms propres dont :
    - 850 dans les domaines
    - 12 360 dans les listes

Important

La distinction précise, à un terme près, entre les noms communs et les noms propres ne pourra être chiffrée qu'après avoir révisé l'ensemble du référentiel, terme par terme. C'est aussi valable pour le nombre exact de noms propres présents dans les domaines, qui est ici estimé à 850.

2.5 La structure générale

L'analyse de la structure générale a pour objectif de donner une idée de l'arborescence et d'identifier la présence ou l'absence de relations sémantiques.

2.5.1 Profondeur du thésaurus

La profondeur du thésaurus va principalement jusqu'aux 5 niveaux hiérarchiques avec 3 exceptions allant jusqu'aux 6 niveaux.

Exemple des 6 niveaux (P 50) :

- Secteur tertiaire (nom du domaine)
    - distribution (n. 1)
        - commercialisation (n. 2)
            - technique commerciale (n. 3)
                - vente (n. 4)
                    - vente à distance (n. 5)
                        - commerce électronique (n. 6)
                        - téléachat (n. 6)
                        - vente par correspondance (n. 6)

2 autres cas du 6ème niveau :

- descripteurs : Christie's / Drouot / Sotheby's (P 50)
- descripteurs : carburant de substitution / essence / essence sans plomb / gaz de pétrole liquéfié (P 33)

Les 31 domaines du thésaurus sont concernés par cette profondeur. Les termes sont très nombreux aux 2e et 3e niveau. Le 4e niveau est plus rare. Le 5e niveau contient 62 termes et le 6e uniquement 3.

Les niveaux de profondeurs sont très variables d'un domaine à l'autre. Par exemple, le domaine Agriculture contient 9 concepts dont 8 correspondent également aux catégories d'autres concepts.

Par domaine, le 1er niveau contient entre 4 à 16 concepts (voir tableau ci-dessous) qui sont pour la plupart également des catégories d'autres concepts. Les termes du 1er niveau sont majoritairement des noms communs, rarement des noms propres, excepté les noms des institutions et organismes.

Le nombre de termes du 1er niveau par rapport au nombre total de termes dans chaque domaine :

Nom du domaine	***Nombre de termes au niveau 1***	***Nombre total de termes dans le domaine***	***Taux (%)***
Agriculture	9	113	8
Aménagement du territoire	9	67	13
Civilisation	4	95	4
Communication	10	309	3
Culture	13	394	3
Défense	7	103	7
Economie	6	110	5
Election	3	59	5
Entreprise	7	78	9
Environnement	3	109	3
Faits divers	3	191	2
Géographie	16	50	32
Industrie	4	135	3
Logement	4	57	7
Loisir	5	64	8
Matière première	4	59	7
Organisme international	17	19	90
Politique	6	191	3
Questions sociales	10	369	3
Religion	7	104	7
Santé	11	270	4
Secteur tertiaire	6	127	5

Nom du domaine	***Nombre de termes au niveau 1***	***Nombre total de termes dans le domaine***	***Taux (%)***
Sport	5	243	2
Transport	13	94	14
Travail	8	172	5
Questions financières	9	192	5
Questions internationales	9	103	5
Questions juridiques	10	239	10
Science et technologie	9	152	6
Questions sociales	10	369	3

Dans 4 cas seulement, le nombre de termes du 1er niveau dépasse de 10 % celui de la totalité des termes dans le domaine : Organisme international, Géographie, Transport, Aménagement du territoire.

2.5.2 Relations entre les concepts

Normes et recommandations

Un thésaurus est caractérisé par les relations entre les concepts qui sont les suivantes :

- **Hiérarchique** : une « relation de subordination de type générique, partitif ou d'instance dans laquelle le terme supérieur représente une classe ou un tout et les termes subordonnés des éléments ou des parties. [30]»

- **Associative** : elle indique une proximité sémantique entre les termes ou les classes d'un langage documentaire. Elle peut être réciproque ou non. La plupart du temps, elle est réciproque.

- **D'équivalence** : une relation d'équivalence entre un terme préférentiel et un terme non préférentiel.

[1]

L'objectif de l'analyse des relations du thésaurus est d'étudier leur pertinence dans le cadre où le moteur de recherche exploitera ces relations. En effet, le moteur de recherche pourra exploiter ces relations de façon appropriée uniquement si elles sont justes et logiques. Les catégories de concepts pourront ainsi correspondre, dans la recherche ou la navigation sur le site web, aux thèmes censés intéresser les utilisateurs.

[30] Citation extraite du cours de H. Rabault [1].

Pour l'exploitation intelligente du thésaurus sur le plan informatique, il est important d'identifier/typer des relations hiérarchiques :

- typage de la relation tout/partie
- typage de la relation d'instance
- typage de la relation personnalisée (La nouvelle norme ISO 25964 autorise la personnalisation et la création de nouvelles relations, par exemple relation cause/effet.)

***Thésaurus de* Libération**

Les relations associatives et d'équivalence sont absentes. Les concepts sont dotés uniquement de relations hiérarchiques. De plus, le typage des relations génériques, partitives ou d'instances n'existe pas.

2.5.3 Relations d'instance

Le souhait des documentalistes est de créer une nouvelle liste « Institutions et organismes » dont les termes sont actuellement « dispersés » dans les différents champs sémantiques. En effet, les institutions et les organismes sont, dans le thésaurus actuel, souvent des termes spécifiques (noms propres) attachés à un concept général (nom commun) par le principe de la relation d'instance. Le problème est également celui de la présentation, ou plus précisément de la présentation unique. A noter que le logiciel de gestion documentaire *Alexandrie* permet les deux présentations [8]. Il est aujourd'hui impossible d'avoir la présentation de la liste des institutions et organismes en plus de la relation hiérarchisée.

Les noms des organismes sont reliés par une relation hiérarchique simple à un concept général. Les normes préconisent de classer les noms propres autour d'une catégorie et non autour d'un concept général, c'est-à-dire de donner un nom à l'ensemble des termes spécifiques. On constate ce problème à plusieurs reprises dans le thésaurus.

Exemple 1

- mutualisme TG
  - Fédération nationale de la mutualité française TSI
  - Maaf Assurance TSI
  - Macif TSI
  - Mnef TSI
  - Mutualité française TSI
  - Mutuelle des étudiants TSI
  - Mutuelles du Mans TSI

Ici, il serait possible de regrouper les TSI autour d'un TG appelé « institution mutualiste » au lieu de « mutualisme ».

Exemple 2

- droits de l'homme TG
  - Apartheid TSI
  - camp de travail TSI
  - Commission européenne des droits de l'homme TSI
  - Conseil des droits de l'homme TSI
  - Convention européenne sur les droits de l'homme TSI
  - Déclaration des droits de l'homme TSI
  - exil TSI
  - France Liberté TSI
  - torture TSI

Dans cet exemple, il sera préconisé d'extraire les noms des institutions (TSI) et de les regrouper autour d'une catégorie (TG), par exemple :

- Institution des droits de l'homme TG
  - Commission européenne des droits de l'homme TSI
  - Conseil des droits de l'homme TSI
  - Convention européenne sur les droits de l'homme TSI
  - France Liberté TSI
  - Ligue des droits de l'homme TSI

Le terme « Institution des droits de l'homme » serait associé à « droits de l'homme » pour ne pas perdre des relations et pour assurer la navigation à l'internaute.

Exemple 3

- géopolitique
  - ACP (états associés)
  - Alena
  - Asean
  - Benelux
  - CEDEAO
  - Commonwealth
  - Conseil de l'Europe
  - Forum de coopération économique Asie-Pacifique
  - Forum méditerranéen
  - Ligue arabe

Le concept général « géopolitique » serait remplacé par la catégorie « accord économique international ».

On pourrait ainsi envisager de créer un champ sémantique avec des catégories d'institutions et d'organismes, pour regrouper et mieux structurer tous les noms des institutions :

- Institutions mutualistes
- Accord économique international
- Instituts de sondage
- Institutions culturelles

2.5.4 Relations erronées

On constate des erreurs de relations dans le thésaurus. Ces erreurs ont pu être causées lors du report ou de l'intégration du thésaurus dans le fichier PDF. Dans les exemples suivants, non seulement aucune des relations hiérarchiques (partitive, générique, d'instance) ne pourra être appliquée, mais il est impossible d'en créer (ou d'en personnaliser) une autre pour que la hiérarchisation soit justifiée.

Exemple 1

- sang TG
    - appareil circulatoire TS

Ici, la relation hiérarchique n'a pas sa raison d'être.

Exemple 2

- métier TG
    - amateurisme TS

Amateurisme peut-il être considéré comme un type de métier ?

Exemple 3

- fonctionnement corporel TG
    - hygiène TS

L'hygiène ne constitue pas un type de fonctionnement corporel.

Exemple 4

- voile (nautisme) TG
    - Coupe de l'America TS (et TG de Coupe Louis Vuitton)
        - Coupe Louis Vuitton TS

Ici, les deux termes (*Coupe de l'America* et *Coupe Louis Vuitton*) seraient les TS du même TG (voile) : la relation hiérarchique entre *Coupe de l'America* et *Coupe Louis Vuitton* est à supprimer.

2.5.5 Prises de position

On constate également des relations qui résultent d'une interprétation trop subjective, ce qui s'écarte de la neutralité exigée dans la structuration de l'information, non seulement au sens documentaire du terme, mais également au sens de l'objectivité journalistique.

Exemple 1

- mutilation corporelle (TG)
  - amputation
  - circoncision

L'utilisateur ne verrait pas cette hiérarchisation directement car le moteur de recherche identifiera, par exemple, les articles sur la circoncision grâce à « circoncision ». En revanche, la catégorie « mutilation corporelle » (qui est aussi un concept général) suggèrera d'affiner la recherche par amputation ou circoncision, ce qui paraîtra illogique par rapport à la « circoncision ».

Une telle hiérarchisation s'éloigne de la neutralité et de l'objectivité des résultats à la recherche, ce qui peut être discutable pour un journal d'actualité. Il est fort probable que l'expression « mutilation corporelle » soit conçue dans le sens d'ablation ou d'amputation, c'est-à-dire dans le sens strictement médical et chirurgical, car il appartient au domaine Santé. Cette explication n'est pas - à la recherche - justifiée non plus, surtout dans un domaine journalistique grand public.

En s'inspirant de Motbis (thésaurus de l'Education nationale), on pourra remédier à cette problématique précise en créant un nouveau concept et en répartissant des concepts dans deux branches distinctes :

- mutilation rituelle (TG)
  - circoncision (TS)
  - excision (TS)
- mutilation corporelle (TG)
  - amputation (TS)

Il est également envisageable de procéder à un regroupement autour du concept « mutilation corporelle » :

- mutilation corporelle (TG)
  - amputation (TS)
  - mutilation rituelle (TS + TG)
    - circoncision (TS)
    - excision (TS)

Exemple 2

- travail précaire (TG)
  - travail intermittent (TS)
  - travail saisonnier (TS)
  - travail temporaire (TS)

Le TG « travail précaire » pourra être remplacé par « durée du travail » (cf. thésaurus Motbis) ou par « période de travail ». Le concept « travail précaire » serait à garder comme concept autonome d'une autre nature.

2.6 Absence des termes non préférentiels

Normes

Les termes non préférentiels sont des termes interdits ou rejetés à l'indexation ; c'est-à-dire qu'ils ne peuvent pas servir à l'indexation mais uniquement à la recherche. Ils sont traditionnellement indiqués par la mention EP (employer pour).

La relation entre le terme préférentiel et un terme non préférentiel est une relation d'équivalence, utilisée pour l'indexation et pour la recherche. Les termes non préférentiels sont des synonymes, quasi-synonymes, abréviations ou variantes orthographiques d'un concept. La relation d'équivalence peut, par exemple, associer à un terme préférentiel son sigle ou son développé, mais aussi des termes d'autres langues. [1]

***Thésaurus de* Libération**

L'absence des termes non préférentiels concerne une grande partie des termes des listes des entreprises et des partis politiques. Dans la liste des partis politiques, les noms des partis sont indiqués uniquement soit en développé, soit en sigle : « Front national », « UMP », « UDF », « Parti radical de gauche », « RPR »... Les termes de la liste Entreprises sont majoritairement en développé sauf quelques-uns : EDF, NPI, NTT, RATP, RDV... Certains d'entre eux sont notoirement connus (EDF, RATP), mais pour d'autres, il s'avère difficile de faire l'association avec une entreprise concrète (NPI, RDV, SMH, UAP, ZDF, RFF, AEG, ATT...).

2.7 Mots-outils

Normes

Les mots-outils ne sont pas obligatoires. Ils ne doivent jamais être utilisés seuls. [1]

***Thésaurus de* Libération**

Le thésaurus contient, suivant la structure visuelle, 261 mots-outils classés en deux parties appelées :

1) « *Mots outils classiques* » (199 descripteurs) : termes plus génériques, par exemple « victoire », « défaite », « rencontre », « température », « test », mais également « signature électronique » ou « université d'été »

2) « *Fonctionnement des organisations* » (62 descripteurs) : termes du type « assemblée générale », « débat », « démission », « mandat », « présidence » ...

   Ces termes, regroupés autour de « Fonctionnement des organisations », font partie d'un domaine à part entière qui correspond même à une discipline scientifique ; cette liste pourra également être développée et structurée. Elle ne correspond donc

pas à ce qu'on appelle liste de mots-outils ; on peut même considérer qu'il s'agit d'une erreur d'organisation dans le fichier PDF. En revanche, les documentalistes les emploient comme les mots-outils.

Les mots-outils aident les documentalistes à compléter l'indexation, à synthétiser le contenu des articles. Malgré leur utilisation fréquente à l'indexation, ils le sont en général beaucoup moins à la recherche. Les documentalistes, pour la plupart, affirment s'en servir beaucoup moins, voire dans certains pas du tout, à la recherche, mais ils en sont très demandeurs à l'indexation.

2.8 Listes annexes

Normes

Les listes annexes ne sont pas obligatoires. Elles sont constituées de noms propres. Il s'agit de termes utilisés à l'indexation et à la recherche. Les termes d'une liste annexe peuvent être reliés par les relations hiérarchiques (termes géographiques, par exemple) ou d'équivalence (sigles et développés des organismes). [1]

***Thésaurus de* Libération**

Le thésaurus contient 4 listes :

- **Partis et mouvements politiques** (120 termes = noms propres sauf 1)
- **Entreprises** (846 termes = noms propres)
- **Personnalités** (9900 termes = noms propres)
- **Territoires** (1500 termes = noms propres)

Les termes des listes *Entreprises* et *Personnalités* sont classés dans l'ordre alphabétique et ne représentent que des noms propres sans noms communs. Aucune hiérarchisation des termes n'est développée.

La liste des partis et mouvements politiques est en deux parties (français et étrangers sans distinctions entre les partis européens et étrangers), avec seulement quelques relations hiérarchiques pour certains partis précis, et aucune autre structuration n'est développée. Les noms des partis y sont classés par ordre alphabétique.

Les seules relations hiérarchiques sont appliquées pour *Partis et mouvements politiques* et pour *Territoires* :

- La liste *Partis et mouvements politiques* est répartie d'abord en deux listes :
  - *En France*
  - *A l'étranger*

A l'intérieur de ces deux listes, les termes sont également classés par ordre alphabétique et parfois hiérarchisés.

Exemple :

- UDF
  - Démocratie libérale (UDF)
  - Force démocrate
  - Parti populaire pour la démocratie française
  - Parti radical (UDF)

- La liste *Territoires* est également répartie en deux listes :
  - *En France*
  - *A l'étranger*

A l'intérieur des deux listes, les termes sont reliés par des relations hiérarchiques. Cette structure est celle d'une taxonomie : Continent / Région(Pays) / Régions nationales / Villes

Exemple :

- Europe occidentale
  - Allemagne
    - Bade-Wurtemberg
      - Baden-Baden
      - Heidelberg
      - Stuttgart
    - Basse-Saxe
      - Hanovre
    - Bavière
      - Dachau
      - Munich
      - Nuremberg

3 Forces et faiblesses du thésaurus

3.1 Forces

- l'identification des principaux domaines traités dans la presse quotidienne ainsi que dans le journal *Libération*
- un lexique couvrant une partie significative des sujets et des thèmes traités dans les publications du journal (papier et numérique)
- l'existence de relations hiérarchiques en tant que point de départ vers une structuration plus adaptée et plus précise
- la présence et le développement des listes pouvant donner lieu à une hiérarchisation et à une structuration plus aboutie en fonction des besoins du site

3.2 Faiblesses

- l'absence des termes non préférentiels (équivalences) ne permettant pas d'exploiter la diversité terminologique des internautes
- le lexique non révisé et non mis à jour
- la présence de concepts trop généraux qu'il faudra évaluer en révisant terme par terme
- les relations d'instance ne sont pas typées pour l'exploitation intelligente par un logiciel
- la présence de certaines relations hiérarchiques fausses
- la présence de mots-outils (200) difficilement exploitables pour la recherche sur le site web et/ou qui nécessiteront des modifications (transformation en locutions)

4 Préconisations principales

De l'analyse du thésaurus ressortent quelques points essentiels pouvant orienter la construction et la modélisation de la taxonomie. Tout d'abord, il est important de s'interroger sur la définition des catégories principales (base de ce nouveau vocabulaire contrôlé), à partir desquelles seront créées d'autres sous-catégories selon les niveaux de profondeur nécessaires. Le vocabulaire contrôlé (taxonomie) du back-office, utilisé en front-office (contrairement au thésaurus), sera structuré tout d'abord en grandes catégories principales dont la base pourra être celle des domaines du thésaurus. Ces domaines seront adaptés et ajustés en fonction des thématiques du journal et des requêtes des utilisateurs. Certains domaines pourront être regroupés. D'ailleurs, les catégories principales de la taxonomie pourraient éventuellement s'aligner par rapport aux domaines du thésaurus d'Eurovoc (voir Annexes 2, p. 93). Chaque catégorie principale sera ensuite divisée en d'autres catégories, classes et sous-classes, au fur et à mesure et en fonction des sujets traités par le journal.

4.1 Mise à jour du lexique : référentiels existants

Pour faciliter la structuration et la catégorisation, il est préconisé de s'appuyer sur des référentiels existants fiables et utiles à la mise à jour du lexique. Les requêtes des utilisateurs démontrent par exemple le besoin de renommer et/ou classer les mots-outils, ainsi que de structurer des noms propres pour une recherche et une navigation plus intéressantes et plus riches. Les questions des utilisateurs sur le site sont les premières à orienter la construction de ce lexique.

Les sujets traités par la presse d'actualité, et particulièrement dans les quotidiens, couvrent tous les domaines de la société. En tout cas, telles sont les attentes des lecteurs démontrées par la diversité des questions posées sur le site. L'ensemble des rédacteurs traitent *a priori* tous les événements jugés importants. *Libération* aborde effectivement tous les domaines de la société, même si le dosage d'intérêt peut varier : l'intérêt porté à la politique n'est pas le même que celui porté à l'économie et aux finances ; le sport, en dehors des événements comme les jeux olympiques, a moins de poids dans le quotidien d'aujourd'hui (et ce, contrairement aux années quatre-vingt) que la culture ; la culture peut être globalement plus orientée vers la littérature et le cinéma que vers la musique, le théâtre, etc. Puis, la ligne éditoriale de la presse est régie par des principes et une idéologie propres à chaque journal, mais également par des questions d'ordre purement médiatique liées à l'intérêt jugé « commercial » ou non d'un événement précis. L'ensemble de ces critères varient fréquemment, influencent le choix des sujets traités ainsi que la façon dont ils sont abordés.

Il est évident que la presse d'actualité ne traitera pas en profondeur des questions propres à la presse spécialisée. Il ne sera pas *a priori* utile d'avoir recours au vocabulaire très pointu dans le domaine de la physique, de la chimie, de la médecine, etc. Néanmoins, une base terminologique de chacun de ces domaines pourrait, ne serait-ce qu'occasionnellement, être utilisée par les rédacteurs. Certains événements liés à une découverte médicale importante justifieront même l'emploi de quelques termes très spécifiques.

Malgré les niveaux d'intérêt différents pour un domaine ou un autre, la granularité variable du développement de chacune des rubriques du journal et de la stratégie adoptée par la direction éditoriale, **l'actualité reste un espace très large** par son spectre de toutes sortes d'événements possibles et nouveaux, mais c'est également **un espace imprévisible** de par le caractère inattendu des événements.

Le vocabulaire de la presse d'actualité est en constante évolution et caractérisé par l'apparition incessante de termes et d'expressions nouveaux : nouveaux pays, nouveaux partis politiques, nouvelles entreprises, nouvelles personnalités, mais également les nouveaux termes liés aux changements naturels de la société ou aux débats sur des problématiques nouvelles (par exemple, nouvelles technologies, progrès médicaux, découvertes scientifiques, nouveaux comportements et mœurs de la société actuelle, etc.). Il est indispensable que le lexique d'un référentiel (thésaurus, taxonomie ou autre) représente toutes les notions principales des sujets traités par le journal. C'est pourquoi, la mise à jour régulière du vocabulaire contrôlé s'avère inévitable, voire indispensable.

Pour disposer de ressources complètes sur une question ou un domaine précis, les professionnels de l'information ont recours à des référentiels existants et des mises à jour régulières. L'immense variété de sujets traités par la presse d'actualité pourra demander plusieurs sortes de référentiels afin de lutter contre la rigidité du vocabulaire contrôlé.

A l'heure actuelle, le documentaliste et/ou le chargé de l'élaboration ou de la mise à jour d'un vocabulaire contrôlé, peut commencer par le très utile site Dmoz qui propose une liste de 52 thésaurus documentaires[31].

Termesciences, portail terminologique multidisciplinaire qui offre une sélection de ressources d'accès libre classées par domaine.

Quelques référentiels (thésaurus et vocabulaires structurés) sont également référencés (Annexes 2, p. 93) par rapport aux besoins du journal *Libération* et ils sont supposés pouvoir servir dans la construction et la modélisation de la taxonomie du back-office du journal.

4.2 Traitement et répartition des mots-outils

Les mots-outils (appelés « classiques » dans le thésaurus de *Libération*) représentent des concepts très généraux et larges. Même s'ils peuvent « compléter » un concept pour mieux résumer le contenu, il est très difficile de retrouver (et de se souvenir de) cette même association (concept d'un domaine + mot-outil) à la recherche. Pour l'utilisation sur le site, dans une taxonomie de navigation, la grande partie de ces mots-outils tels qu'ils existent créeraient avant tout une source de bruit/silence et n'amélioreraient aucunement la recherche ou la navigation. Les mots-outils en tant que tels n'apportent donc pas de valeur ajoutée et ils ne font pas l'objet de questions de la part des utilisateurs sur le site, tant qu'ils ne représentent pas un degré minimum de précision.

Certains de ces noms communs pourraient être transformés, à travers l'association de concepts et d'un mot-outil, en locutions : « débat politique » au lieu de « débat », « température record » au lieu de « température ». Ces locutions correspondraient ainsi aux concepts susceptibles d'intégrer les différents domaines du thésaurus. Cette démarche pourrait être complétée par l'analyse des fiches indexées : à étudier par exemple, quel type d'articles sont indexés par « débat » et quelle locution (pour préciser et affiner le concept) serait la plus appropriée.

Il serait également possible de voir, dans une analyse plus approfondie, si certains de ces termes pourraient faire partie des domaines existants du référentiel ou éventuellement s'il s'avérerait utile de créer des domaines appropriés.

[31] Liste de 52 thésaurus supervisée par Sylvie Dalbin.

Exemples :

Certains termes appartenant à la partie *Fonctionnement des organisations* sont classables dans le domaine « Travail » : « démission », « dirigeant », « effectif », « états-généraux »...

Les termes du type « négociation », « débat » ou « dérogation », pourraient donner lieu à la création d'un mini-champ sémantique *Processus économique et politique*.

Enfin, les mots-outils trop larges, à sens multiples, ambigus et impossibles à employer dans une locution appropriée, pourront être exclus du référentiel, car ils sont difficilement exploitables à la recherche. Dans une taxonomie de navigation, ces termes (par exemple, « abandon », « rapport », « entourage ») auraient un sens beaucoup trop large pour pouvoir structurer les documents de façon pertinente. Mais il est préconisé de s'interroger d'abord sur la possibilité d'accrocher ces mots-outils à un concept.

4.3 Créer des relations sémantiques

Pour mieux organiser les contenus (articles) et permettre aux utilisateurs du site une recherche et une navigation plus affinées, les listes mériteraient une structuration plus fine. Par exemple, les partis politiques pourraient être classés selon les grandes lignes idéologiques (gauche, droite, centre) et selon les partis actuels/anciens. Les entreprises pourraient être classées en publiques/privées, puis par français/étrangers (ou par pays) et par secteurs et/ou domaines d'activité (automobile, aviation, alimentaire, téléphonie, banque, édition...).

Les référentiels pouvant servir à la mise à jour de la liste des partis politiques sont, comme évoqué plus haut, par exemple, la liste Wikipédia (les plus grandes entreprises ou par pays), Fortune global 500 ou blog France-Politique (Voir Annexes 2, p. 95).

De la même manière, les noms propres de la liste des personnalités pourront être structurés et classés pour offrir aux utilisateurs une meilleure organisation des articles par types de personnalités. Par exemple, il serait intéressant d'avoir sur le site des facettes du type :

« personnalités politiques », « artistes », « écrivains » ou « sportifs »…

L'analyse d'un vocabulaire contrôlé ne devrait pas se limiter au lexique et à la terminologie mais elle doit porter sur les relations d'appartenance des concepts à un domaine (ou à plusieurs domaines, en cas de polyhiérarchie), car ces relations sont directement associées aux pratiques d'indexation ainsi qu'à la recherche. Puis, le thésaurus demande une attention particulière dans l'analyse minutieuse de ses relations sémantiques qui sont hiérarchiques et associatives. C'est effectivement l'existence et la qualité des relations sémantiques entre les concepts, qui constituent la particularité, la richesse et l'essence-même de la structure d'un thésaurus. La terminologie n'est donc, en aucun cas, dans un vocabulaire contrôlé, le seul point que les référentiels existants peuvent aider à améliorer et à enrichir.

4.4 Développer des équivalences

L'absence des termes non préférentiels pour déterminer les relations d'équivalence limite la pertinence de l'indexation et de la recherche.

La taxonomie de navigation permet, grâce aux technologies informatique et web (et liens hypertextes), de gérer des relations d'équivalence comme des synonymes ou sigles. A chaque sigle d'un parti politique ou d'une entreprise pourra être associé son développé, ce qui augmentera les chances à la recherche sur le site. L'identification des termes non préférentiels est, pour un site web, la garantie que le moteur de recherche repère également, en dehors du terme utilisé dans la presse le plus souvent, une autre forme du terme défini comme préférentiel.

Par exemple, si « UMP » avait pour EM le développé « Union pour un mouvement populaire », le moteur de recherche indexerait - et uniquement - par « UMP » également les contenus dans lesquels figure le développé « Union pour un mouvement populaire ». De la même manière, si « Le parti socialiste » avait pour EM le sigle « PS », le moteur de recherche remonterait également les documents dans lesquels ne figure que « PS ».

Cas de concepts proches

Pour certains cas de quasi-synonymie, il y a également la problématique des concepts proches qui sont dans la presse utilisés de façon très régulière. Par exemple, « insertion professionnelle » et « réinsertion professionnelle » pourraient être regroupés comme des équivalences documentaires d'un concept plus général, nommé par exemple « retour à l'emploi ».

5 Passer du thésaurus à la taxonomie

Le thésaurus de *Libération* ne suit donc pas les règles de la nouvelle norme. Il ne possède plus de relations sémantiques propres à ce langage documentaire. Aujourd'hui, avec les seules relations hiérarchiques (mais qui ne sont pas typées), il serait difficilement envisageable de reconstruire un thésaurus correspondant à la norme, c'est-à-dire un thésaurus fonctionnel approchant de la structure du thésaurus *Popline*, par exemple. Les moyens en termes financiers et humains seraient importants pour redresser le thésaurus actuel à un niveau exploitable. C'est pourquoi la solution la plus appropriée serait de faire évoluer le thésaurus par un vocabulaire contrôlé moins complexe… par une taxonomie (voir Taxonomie définition, p. 39).

5.1 Evaluer la reprise de l'existant

A *Libération*, le thésaurus évoluera vers la taxonomie alors que le site de la base de données *Popline* utilise son thésaurus. Cette reprise de l'existant est en quelque sorte entière à la différence de *Libération* où la reprise serait partielle. L'existant dans le cas *Popline* étant suffisamment performant et solide, il a pu être utilisé pratiquement tel quel. En revanche, l'existant de *Libération* a un très faible niveau pour pouvoir être repris en tant que langage documentaire. C'est pourquoi, il sera orienté et remplacé par la taxonomie.

Le passage du thésaurus actuel en taxonomie représente également un investissement non négligeable. Il reste néanmoins plus simple d'utiliser l'existant plutôt que de recommencer de zéro tout un développement d'une structure et d'un vocabulaire, particulièrement face à la reprise de l'indexation du fonds existant. La reprise de l'existant joue un rôle primordial dans la décision et l'orientation d'un projet à condition que l'évaluation de l'existant soit permise, prise au sérieux et effectuée correctement. Cette évaluation nécessite *a priori* l'appel des professionnels de l'information à travers un audit rigoureusement mené. Ce premier investissement humain et financier sera la base de la réussite du projet. Il permettra non seulement de confirmer et de préciser la faisabilité du projet fondé sur l'existant, de rappeler, voire redéfinir les besoins réels du site, mais également de définir des étapes essentielles du projet à mener.

Dans ce cas précis, après la première analyse effectuée plus haut, il apparaît que la reprise partielle de l'existant soit envisageable. Les points forts et faibles ont été dégagés et pourraient orienter la rédaction d'un cahier des charges technique, qui définira, de façon détaillée, le degré d'exploitabilité précis de l'existant ; en d'autres termes, il décrira dans quelle mesure le thésaurus actuel sera utilisable pour le développement de la taxonomie.

5.2 Définir les principales étapes du projet

Le cahier des charges technique pointera les principales étapes et proposera un planning détaillé.

Dans le cas du projet *Libération*, l'un des points essentiels de cette élaboration correspondra à la définition des catégories principales qui pourra se baser sur, ou du moins s'inspirer, des domaines du thésaurus.

Ensuite, l'une des étapes les plus laborieuses consistera dans le développement du vocabulaire qui reposera sur la transformation des concepts en sujets. Les chargés de cette mission devront créer des concepts suffisamment précis et en lien avec les questions des utilisateurs sur le site. Ce degré de précision sera préalablement discuté et défini avec l'ensemble des documentalistes, mais également avec les rédacteurs et les responsables éditoriaux du site. Par exemple, déjà évoqué, le terme « entourage » pourrait être « transformé » en « entourage d'une personnalité », mais aussi en « entourage d'une personnalité politique » ou encore en « entourage d'une personnalité politique en France »…

Enfin, l'une des difficultés de ce passage/transformation sera liée aux essences différentes de ces deux formes de vocabulaire : les domaines d'un thésaurus peuvent être coordonnés entre eux tout en gardant leur autonomie, alors que la hiérarchisation en catégories d'une taxonomie ne permet pas cette souplesse. Par conséquent, la hiérarchisation de concepts en catégories et sous-catégories d'une taxonomie ne remplacera pas la richesse des relations sémantiques. Un terme hiérarchisé dans une catégorie ne doit pas être utilisable pour une catégorie différente. Par exemple, le terme « effectif » placé dans la catégorie « Travail » empêchera l'utilisation de ce terme pour un article d'ordre purement économique et classable dans la catégorie « Economie ».

Cinquième partie
Perspectives pour l'activité documentaire dans les organisations

L'activité documentaire performante est aujourd'hui inséparable de l'aspect technologique. Les professionnels de l'information ne peuvent pas négliger ce point important qui contribue considérablement à la manière dont l'accès à l'information se présente aux utilisateurs. C'est effectivement à travers l'amélioration constante de l'accès à l'information – à l'aide des nouvelles technologies – que les enjeux des professionnels de l'information évoluent et ne se limitent pas au cadre strict d'un service de la documentation.

1 Importance de la technologie

Comme déjà évoqué plus haut, un CMS, adapté, performant, mais aussi facile à mettre à jour avec la possibilité d'ajouter les outils multimédias (insertion d'images, de son et de vidéo) et capable de gérer les vocabulaires contrôlés tels que la taxonomie ou le thésaurus, est indispensable dans la construction d'un site d'information. [29] Pour retrouver un contenu par un autre biais que les rubriques et les menus, le moteur de recherche s'impose. Certains CMS ont leur moteur de recherche intégré, mais les moteurs de recherche propriétaires et orientés vers les traitements de texte automatiques sont en général plus performants. Ils permettent, en les associant à un langage documentaire, d'améliorer considérablement la recherche sur le site. Enfin, l'une des réalités technologiques d'aujourd'hui, et celle qui constitue l'avenir dans l'accès à l'information, est le web sémantique. Cette technologie aura-t-elle une place dans l'activité documentaire et plus précisément dans la presse ?

1.1 Articulation de l'indexation automatique et de l'indexation humaine

1.1.1 Définitions et différences

Structurer une base de données d'un titre de presse uniquement par les champs titre, sous-titre, date et auteur ne suffit pas[32] ; l'évolution permanente de la langue, l'ambiguïté du langage naturel et plus particulièrement des discours journalistiques demandent une analyse du contenu qui, à *Libération*, est traitée par l'indexation effectuée uniquement manuellement par les documentalistes. Parallèlement à l'indexation humaine, l'activité documentaire peut bénéficier d'une indexation automatique grâce à un bon moteur de recherche. Certains titres de presse l'ont d'ailleurs entièrement adoptée : c'est le cas du *Monde*, du *Figaro* ou du *Parisien* qui se sont dotés du moteur de recherche Intuition de *Sinequa*. [33]

Les caractéristiques principales et les impacts de ces deux modes d'indexation sont les suivants :

L'indexation humaine et manuelle est profonde : il s'agit d'une opération intellectuelle que seul l'être humain est capable d'effectuer, grâce à sa compréhension de toutes les richesses et les finesses du langage naturel. L'indexation humaine peut être contrôlée ou libre (voir p. 16). Quand elle est contrôlée, elle repose sur un langage documentaire. La recherche (l'accès à l'information) passe par les métadonnées produites dans les champs créés à cet effet. En revanche, elle est très longue et coûteuse. Seuls les documentalistes peuvent l'effectuer.

---

[32] Cette structuration via l'indexation fonctionnelle dans les champs prédéterminés est, à *Libération,* gérée dans le back-office automatiquement (vérifié et éventuellement complété par les documentalistes), ainsi qu'au *Monde* d'ailleurs [26, Contat, p. 27-28].

[33] Wikipédia, consulté le 5/10/2012.

L'indexation automatique s'opère sur l'ensemble d'un texte. Elle est produite à partir du texte intégral et ne repose pas sur un langage documentaire mais sur des référentiels linguistiques adhoc. L'interrogation s'effectue en langage naturel. L'indexation automatique est sans doute bien plus rapide, mais elle ne peut pas résoudre tous les problèmes liés à l'ambiguïté du langage, quel que soit le niveau élevé de l'analyse automatique. La première génération des systèmes d'indexation automatique était fondée sur la création d'index (fichiers inversés et fichiers topologiques) permettant de localiser un terme ou un ensemble de termes au sein d'un corpus déterminé. Ces outils de TALN (traitement automatique du langage naturel) utilisent des méthodes statistiques (fréquence et occurrence des termes), linguistiques (analyse grammaticale, morpho-lexicale) ou informatiques (algorithmes de recherche). Chaque moteur de recherche exploite, développe et combine ces méthodes, à des niveaux très variés. [23].

Sinequa, mais aussi *Pertimm* ou *Lingua* disposent de moyens linguistiques, comme les dictionnaires spécifiques ou généraux, permettant l'analyse linguistique [22]. Cette dernière repose sur l'analyse morphologique, syntaxique, sémantique ou peut aller jusqu'à l'extraction d'entités nommées[34].

Il s'agit donc de deux fonctionnements fondamentalement différents. Cette différence correspond à l'opposition entre l'intellectuel (fait par l'homme) et l'automatique (fait par la machine). C'est sans doute précisément cette différence fondamentale qui crée la complémentarité nécessaire ; l'un et l'autre possèdent leurs avantages et leurs inconvénients qui sont complémentaires. D'ailleurs, certains professionnels de l'information affirment, depuis plus de 20 ans, que **les technologies de l'indexation automatique ne devraient pas s'opposer à l'indexation humaine :**

« Une autre façon de pallier les faiblesses de l'indexation Lexinet[35] serait de l'associer à une supervision d'un expert humain. Cela reviendrait à associer les deux modes d'indexation, cherchant ainsi à unir les caractéristiques positives de chacune des indexations: les techniques Lexinet apporteraient un "plus" pour la régularité, la spécificité et l'évolution des termes, l'indexation manuelle permettrait d'introduire des descriptions intermédiaires, apportant ainsi une meilleure accessibilité à l'information, et levant les ambiguïtés possibles de certains termes. »[36]

A la grande différence du moteur de recherche de *Libération*, le moteur de recherche *Intuition* du *Monde* permet une indexation automatique via un traitement linguistique qui s'appuie sur le dictionnaire des synonymes [26] ou sur le traitement du singulier/pluriel. Le moteur de recherche remonte par exemple les articles contenant à la fois « partis politiques » et « parti politique », tandis que sans cette gestion, l'utilisateur n'obtiendrait que les articles avec le terme correspondant uniquement à celui posé dans sa requête. En revanche, le vocabulaire contrôlé, ou un langage documentaire, n'y est plus utilisé.

Mais quelle est la place du langage documentaire, c'est-à-dire de l'indexation humaine, dans un système où l'indexation automatique existe ?

---

[34] Reconnaissance d'**entité nommée** est une sous-tâche de l'activité d'extraction d'information dans des corpus documentaires. Elle consiste à rechercher des objets textuels (c'est-à-dire un mot ou un groupe de mots catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc.). Wikipédia. Consulté le 5/10/2012.

[35] Lexinet était un système d'indexation automatique, fondé sur la méthode statistique, et un gestionnaire de lexique.

[36] [15, Chartron, Dalbin]

1.1.2 Bienfaits de la complémentarité

Ajouter l'indexation humaine à l'indexation automatique apporte de la valeur ajoutée uniquement là où les traitements automatiques ne remontent pas les termes liés à l'information recherchée.

Si, dans le contexte du traitement automatique, les termes de l'interrogation ne correspondent pas aux termes présents dans le contenu, le moteur de recherche ne les remontera pas. Par exemple, les questions sous forme de noms propres, comme Chirac ou Mitterrand, remonteront les articles pertinents à condition que ces deux noms fassent partie du contenu. Le problème sera, certes, celui de bruit et de silence. Pour atténuer le phénomène bruit/silence, certains logiciels bénéficient d'un traitement de la synonymie, mais également de traitements permettant l'analyse grammaticale et morpho-lexicale, la lemmatisation[37] [23].

O. Contat [26] rappelle les cas où le traitement linguistique et statistique des articles de presse peut difficilement remplacer entièrement l'analyse humaine et où la recherche de l'information pertinente pose des problèmes. Par exemple, l'orthographe des noms propres peut souvent varier d'un journal à l'autre, mais également d'un journaliste à l'autre. Ensuite, de nombreux cas d'homonymie (entreprise Orange, fruit orange) ne sont toujours pas détectés par les logiciels actuels : *Intuition* traite ce problème grâce au traitement majuscule/minuscule, mais si cette différence est ignorée par le journaliste ou si le terme « ORANGE » fait partie d'un titre en caractères majuscules, la distinction entre l'entreprise et le fruit ne sera pas effectuée. Enfin, les questions purement sémantiques ne sont pas prises en compte par un moteur de recherche comme *Intuition* : déclaration de Chirac à Jospin/Déclaration de Jospin à Chirac.

Une indexation humaine pourra compléter le traitement linguistique et statistique d'un logiciel. Et ceci à travers une liste des renvois ou d'un thésaurus, donc à l'aide d'un vocabulaire contrôlé.

« Que ce soit avec l'aide d'un thésaurus ou un index matière, l'indexation manuelle est la seule qui puisse transcrire des notions qui ne sont pas explicites dans un document. »[38]

L'utilisation d'un vocabulaire contrôlé permet également d'organiser les articles par sujets et par thématiques. Déjà évoqués, les topics maps (cartes topiques), comparés dans leur structure aux ontologies, font partie des systèmes de recherche les plus complexes et les plus aboutis. Le thésaurus est sans doute, grâce à sa capacité de gestion des relations, le langage documentaire le mieux adapté à cette organisation fine des connaissances. Enfin, les classifications telles que les taxonomies permettent aux documentalistes de « classer » les articles dans les bonnes catégories préalablement déterminées, et de contribuer à l'organisation d'un corpus important.

L'intégration d'un vocabulaire contrôlé n'empêche donc pas un recours à l'indexation automatique ; l'usage du langage documentaire est, en quelque sorte, son aboutissement pour un accès à l'information bien structuré et organisé. L'indexation automatique permet de gérer le bruit et le silence et d'économiser considérablement le temps ainsi que les moyens humains. En revanche, sans aucune indexation humaine, c'est-à-dire, sans aucune analyse intellectuelle, de nombreux textes, surtout ceux de la presse, ne seront pas, du fait de leur ambiguïté et de l'évolution du langage, associés à des concepts, des notions et des sujets qu'ils représentent réellement. Par conséquent, l'utilisateur de la base de données, et également celui du site web, passera à côté d'eux au moment de la recherche.

---

[37] Définition ADBS : identification de la racine d'un terme. Consulté le 6/10/2012.
[38] [26, Contat, p. 46]

En pratique, il est donc indispensable que les professionnels de l'information, accompagnés par les responsables informatiques, se tiennent informés des logiciels et solutions existants. En France, c'est par exemple Mondeca qui peut répondre aux besoins d'un logiciel de traitement automatique et à la fois complémentaire dans la gestion d'un vocabulaire contrôlé ou d'un langage documentaire. L'éditeur offre des solutions articulant l'indexation manuelle et l'indexation automatique. ITM de Mondeca permet d'adosser les thésaurus ainsi que les taxonomies et les vocabulaires contrôlés afin de gérer les corpus importants dans les entreprises ou la navigation et la structuration des portails web. [19]

Par ailleurs, ITM de Mondeca permet d'éditer les vocabulaires contrôlés sous des formats différents (Excel, HTML, PDF), mais il permet également d'importer ou d'exporter une taxonomie ou autre vocabulaire au format SKOS qui est le standard d'échange dans le cadre du web de données.

1.2 Ouverture au web de données

Le web de données, ou le web sémantique, fait aujourd'hui partie des technologies incontournables dans les réflexions sur les perspectives des applications et pratiques documentaires.

La première vision du web de données de Tim Berners-Lee[39] est celle « d'un immense espace d'échanges de ressources entre machines permettant à des utilisateurs d'accéder à des grands volumes d'information et à des services variés. » [25]

C'est le principe même du web et la puissance de l'hypertexte où « toute ressource peut être liée à toute ressource »[40] qui aide à comprendre l'objectif du web de données : lier les données entre elles (*linked data*) pour produire de l'information structurée. Le web sémantique, aussi appelé le web 3.0, est considéré comme le prolongement du web 2.0.

La nouveauté du web sémantique par rapport au web 2.0 se situe dans une évolution d'ordre technologique. On peut dire, de façon tout de même très raccourcie, que le web sémantique reprend les notions principales du web qui sont l'URI[41] (l'identifiant de la ressource) et HTTP[42] (le protocole permettant la présentation de la ressource), mais contrairement au web 2.0, le web sémantique n'utilise plus HTML[43] (le langage de balisage permettant la publication et la structuration du message). Le web 3.0 l'a remplacé par RDF (*Resource Description Framework*), le langage de description auquel s'ajoutent d'autres standards. Ce langage de description est souvent comparé à une grammaire permettant de décrire la relation qui lie deux objets entre eux. La différence avec le web 2.0 est essentiellement dans cette volonté de décrire la nature des relations, de prolonger en quelque sorte la logique de l'hypertexte, c'est-à-dire de traduire non seulement la structure du message, mais aussi de lui ajouter un sens. Cette formalisation logique des connaissances et la création des inférences entre les objets passent par la formalisation des balises (reconnaissables par la machine). [27]

---

[39] Tim Berners-Lee est le principal inventeur du World Wide Web et fondateur du W3C.
[40] [25, Charlet, p. 120]
[41] Universal Resource Identifier
[42] Hyper Text Transfer Protocol
[43] Hyper Text Markup Language

Le web de données tente de « créer le lien automatique pour relier les données qui sont stockées dans les différents fichiers et bases de données de nos ordinateurs. [44]»

« Chaque entreprise devra marquer toutes les données qu'elle veut publier sur le Web sémantique avec un descriptif.[45] » Ces descriptifs sont des standards ou des formats d'échange, ajoutés au descriptif RDF permettant de lier des concepts et des schémas entre eux. Le standard SKOS est celui qui peut représenter les concepts et les schémas des thésaurus et des autres vocabulaires contrôlés, ainsi que des topic maps ou des taxonomies :

« SKOS (*Simple Knowledge Organization System*) est une famille de langages formels permettant une représentation standard des thésaurus, classifications ou tout autre type de vocabulaire contrôlé et structuré. Construit sur la base du modèle de données standard RDF, son principal objectif est de permettre la publication facile de vocabulaires structurés pour leur utilisation dans le cadre du Web sémantique. »[46]

Les applications documentaires du web sémantique exploitent les vocabulaires contrôlés pour formaliser les balises, grâce aux relations d'un thésaurus par exemple, afin de créer les inférences voulues.

C'est donc la notion d'échange et le mécanisme d'interopérabilité qui sont au cœur du web de données, mais la première condition de l'interopérabilité est celle de l'ouverture de données (*linking Open Data*). Sans cette ouverture, le web de données ne peut pas se développer.

A l'heure actuelle, aucun titre de presse en France ne pratique ce qu'on appelle le web sémantique, alors que les sites d'information et les informations d'actualité sont dans leur essence même publics et s'inscrivent ainsi dans le principe fondamental du web sémantique. Que représenterait concrètement cette transformation des vocabulaires contrôlés des différents titres de presse en SKOS, pour s'introduire dans ce vaste espace qui est celui du web de données ? Un article de *Libération*, au moment où il a été publié sur le site web, devenu accessible par tous les utilisateurs de l'Internet, est immédiatement partageable tout simplement grâce à un lien hypertexte. Qu'apporteraient de plus la conversion au format SKOS et l'espace du web de données ?

L'immersion d'un article dans le web de données suppose qu'il soit attaché, via l'indexation, à la taxonomie, cette dernière convertie en SKOS. Les concepts de cette taxonomie permettront ainsi l'accès aux autres données structurées en RDF sur le web, par exemple aux données des catalogues de la BNF à travers le référentiel RAMEAU, lui-même en SKOS/RDF. Cet accès incitera le lecteur à enrichir ses connaissances sur une œuvre littéraire, sur un personnage ou autre, mentionnés dans le journal. L'apport du web sémantique dans un site d'information consisterait dans l'accès à l'information structurée et dans cet enrichissement perpétuel de la base de connaissances qu'un site pourrait offrir à ses utilisateurs, à travers des vocabulaires contrôlés comme taxonomie ou thésaurus.

La technologie et les progrès du web sémantique concernent également l'innovation des moteurs de recherche qui ne se limitent plus uniquement à la recherche en texte intégral. Le principe du fonctionnement de ces nouveaux moteurs consiste dans l'exploitation de référentiels métiers (en SKOS/RDF) afin de les relier aux données en RDF, comme c'est le cas de la plate-forme scientifique Isidore[47], développée par Adonis.

---

[44] [25, Charlet, p. 125]
[45] [25, Charlet, p. 126]
[46] Wikipédia, consulté le 6/10/2012.
[47] Isidore, consulté le 6/10/2012.

Si le web sémantique représente un avenir prometteur pour l'activité et les techniques documentaires, il contribue nécessairement aux changements fonctionnels et organisationnels dans les organisations où les vocabulaires contrôlés n'ont pas dit leur dernier mot.

2 Changements fonctionnels et organisationnels

Les vocabulaires contrôlés participent effectivement, et de façon non négligeable, à la construction des systèmes d'organisation des connaissances (SOC). [32] La conceptualisation, l'évaluation et les mises à jour des vocabulaires contrôlés ont toujours été effectuées par les documentalistes et les professionnels de l'information en général, qui y trouvent aujourd'hui des enjeux nouveaux. L'étroite collaboration et le rapprochement avec d'autres professions présentes dans les entreprises semblent inévitables, voire s'imposent.

2.1 Enjeux pour les documentalistes

Aujourd'hui, dans le secteur de l'info-doc qui est de plus en plus numérique, ce sont encore les documentalistes qui alimentent, organisent et structurent au mieux les métadonnées. Cette organisation et cette structuration des métadonnées sont en lien étroit avec le suivi des vocabulaires contrôlés ainsi qu'avec leur adaptation aux besoins nouveaux. C'est majoritairement le savoir-faire des documentalistes qui peut créer un vocabulaire contrôlé adapté, spécialement conçu et sur mesure pour représenter tel ou tel corpus de documents et de questions des publics. De la même manière, le développement et la modélisation d'un nouveau vocabulaire permettant une navigation à la fois riche et structurée sont aujourd'hui plus que jamais essentiels dans la bonne gestion d'un site d'information : les sites d'information doivent effectivement offrir aux utilisateurs une recherche et une navigation efficaces avec en priorité les résultats pertinents.

Peu importe de savoir si la presse en ligne remplacera un jour la presse écrite ou si cette dernière finira par être son complément ou si encore elle acquerra une position tout à fait différente de celle d'aujourd'hui. Sa place dans l'avenir reste difficile à prévoir. En revanche, il est certain que la presse en ligne fait partie de l'immense processus de numérisation de nos sociétés et elle ne pourra plus disparaître ou même faire marche arrière. Par ailleurs, la gestion et la structuration des bases de données en ligne doivent, à mon sens, faire inévitablement appel aux professionnels de l'information. Les professionnelles de l'info-doc continuent, par ce biais, à avoir leur rôle à jouer dans la création et la mise à jour des vocabulaires contrôlés, que ce soit à l'aide des référentiels existants (Annexes 2, p. 93) ou à l'aide des différents schémas proposés par l'IPTC (voir p. 1411). La bonne structuration et l'organisation des données numériques font partie des compétences et du savoir-faire des documentalistes, car ils ont une vision claire et précise de l'ensemble du corpus qu'une organisation ou une entreprise possède. Par ailleurs, l'ouverture des fonds documentaires dans le cadre du web sémantique demande une implication importante de la part des documentalistes, dans la mesure où ce fonds doit être structuré et organisé et doit pouvoir répondre aux recommandations du W3C.

Pour répondre au développement d'un environnement totalement numérique, une nouvelle formation vient d'être créée à l'ENS de Lyon, Master Architecture de l'information. Ce master « procure une formation innovante pour concevoir, organiser et présenter l'information aux utilisateurs dans un environnement numérique, interactif et mobile[48] ». [40]

[48] Citation extraite du site de l'ENS de Lyon, [consulté le 5 octobre 2012].

Dans le contexte des *nouveaux* vocabulaires contrôlés, il faudrait mettre l'accent sur l'adjectif « *nouveau* », car la transformation du thésaurus en taxonomie pointe justement des enjeux pour les documentalistes et leur rôle à jouer. En effet, avec l'évolution des langages documentaires dans les systèmes informationnels et la nécessité d'élaborer des vocabulaires contrôlés, le rôle des documentalistes dans les organisations d'aujourd'hui ne se limite plus à l'utilisation et au maintien d'un vocabulaire contrôlé mis en place, mais ils doivent prendre l'initiative dans le choix et l'adaptation de ces vocabulaires aux besoins nouveaux. La construction de la taxonomie de navigation est l'occasion unique de montrer leurs compétences. Leur participation dans la construction d'un nouveau vocabulaire contrôlé devrait être active et soutenue, car ce sont eux qui maîtrisent et connaissent le mieux l'ensemble des documents de leur base de données et les questions des utilisateurs, que ce nouveau vocabulaire contrôlé est censé représenter.

Une fois la taxonomie structurée et mise en place, les documentalistes contrôleront, orienteront et complèteront l'indexation automatique grâce à l'analyse intellectuelle via l'indexation manuelle. Ils pourront ainsi participer activement à la création de sujets publiés sur le front-office du site, en créant des profils de recherche, comme c'est déjà le cas d'autres professionnels de l'information dans les entreprises et précisément des services de la DSI (*Diffusion sélective de l'information*). En fonction des équations de recherche (fondées sur les termes de la taxonomie et/ou les requêtes des utilisateurs), ils créeront des sujets susceptibles d'intéresser les lecteurs. Le principe consiste dans l'association d'un sujet précis à un dossier, à « la bannette », contenant les résultats pertinents par rapport au sujet défini.

Une telle activité qualifiée de veille et de création de sujets valorisera considérablement leur activité par rapport à celle effectuée actuellement, par exemple à *Libération,* et contribuera à les placer au centre de l'activité de l'entreprise et à les rapprocher davantage des autres acteurs du journal, notamment les journalistes et les informaticiens.

Le cas concret de *Libération* soulève et confirme l'importance de la gestion de projet par les professionnels de l'info-doc eux-mêmes pour pouvoir faire face à l'ampleur qu'un tel projet représente. En effet, malgré l'anticipation et l'implication du service informatique et web qui était l'initiateur du projet et avait, en tant que premier, l'idée de se servir des termes du thésaurus ainsi que des descripteurs afin d'améliorer la recherche et la navigation sur le site, c'est naturellement le service de la documentation qui doit solliciter aide, appui et chercher la solution auprès des professionnels de l'info-doc. Au cœur de la problématique du projet est le thésaurus, un vocabulaire contrôlé, c'est-à-dire un élément d'aspect purement documentaire et que seuls les documentalistes utilisent et maîtrisent. Le responsable du service informatique, en se positionnant en chef de projet, ne peut effectivement proposer, seul, des solutions suffisamment satisfaisantes et appropriées dans la structuration et la modélisation d'un nouveau vocabulaire contrôlé. Il paraît donc primordial que le service de la documentation, accompagné du service informatique certes, mais également des journalistes, au titre des producteurs de l'information, puisse confier la gestion de projet à un professionnel de l'info-doc et que les documentalistes restent entièrement impliqués dans le projet. Par ailleurs, le chef de projet, dans un premier temps, favorisera la communication et s'appliquera à installer une collaboration étroite entre les trois univers qui doivent aujourd'hui être perçus comme complémentaires.

Comment renforcer, voire créer cette communication entre les documentalistes, les informaticiens et les journalistes ? L'expérience à *Libération* a démontré que les questions des utilisateurs aient une importance non négligeable, voire qu'elles soient au cœur des préoccupations des documentalistes afin de créer, structurer et mettre à jour régulièrement un vocabulaire contrôlé (un thésaurus, une taxonomie). Les questions des utilisateurs servent ainsi de lien indispensable entre les deux services. C'est en travaillant étroitement et régulièrement avec les informaticiens que les documentalistes ont les moyens de compléter, enrichir et mettre à jour un vocabulaire contrôlé par rapport aux besoins des lecteurs : les outils d'analyse informatiques permettent d'identifier quelles questions sont posées par les

utilisateurs et les documentalistes les intègrent ou non dans le vocabulaire contrôlé. Cette utilisation et transformation des questions en mots-clés pointe le problème lié à la question de l'évolution des termes dans le temps. Il s'agit d'une question très complexe à laquelle les documentalistes réfléchissent et répondent tout en tenant compte de leurs pratiques d'indexation en même temps que du fonds déjà indexé. Toute cette activité documentaire donc impose que les documentalistes se rapprochent également des autres activités de l'entreprise.

2.2 Rapprochement indispensable des différents savoir-faire

Le cas de *Libération* illustre effectivement le besoin de complémentarité des différents métiers et montre en quelque sorte la porosité qui existe entre les objectifs des documentalistes, des informaticiens et des journalistes. Les trois métiers doivent obligatoirement trouver un terrain d'entente afin de mieux structurer le site d'information et leur base de données qui représente pour eux, mais surtout pour les lecteurs, une base de connaissances. Un site d'information ne se limite plus à l'accès immédiat à l'actualité, publiée et lue au même moment, et ensuite « enfouie » pour toujours dans les archives du journal. C'est la structuration qui participe à la réutilisation des documents en cas de besoin, et grâce au web sémantique, à l'enrichissement des informations publiées sur le site par d'autres informations complémentaires, nouvelles ou inédites. Il est effectivement dans l'intérêt de tous qu'une structuration efficace et maintenue quotidiennement par les professionnels de l'info-doc soit mise en place.

Les compétences des informaticiens sont primordiales dans le fonctionnement d'un site, car indispensables dans le développement de services et le paramétrage technique des différents logiciels et solutions. Les informaticiens participent au choix d'un CMS ou d'un moteur de recherche, et surtout, ils les installent, les paramètrent et les mettent à jour en fonction des besoins du site et de la granularité de sa structuration. Et ce sont justement ces besoins qui sont tout d'abord repérés par les documentalistes. L'informaticien gère les aspects fonctionnels et techniques et le documentaliste le guide. Encore une fois, c'est surtout le documentaliste qui maîtrise le vocabulaire contrôlé, avec ou sans relations sémantiques d'un thésaurus, et c'est également lui qui a la meilleure vision de l'ensemble du contenu de la base de données que ce vocabulaire contrôlé représente, ainsi que des relations plus ou moins fines entre le thésaurus et le corpus qu'il indexe.

A travers la création de sujets, le documentaliste suit également les questions des utilisateurs et sait lui proposer des documents en conséquence ; il est ainsi proche des goûts du lecteur et par ce biais, il peut trouver sa place à côté des rédacteurs. Non pas pour les remplacer dans leur métier de journaliste, mais pour les orienter vers les nouveaux sujets à traiter, tout en constituant des agendas prévisionnels riches et de qualité. La rédaction ne lui est pas interdite, bien sûr, et il peut coiffer la casquette du rédacteur-documentaliste telle qu'elle existe dans des agences de publicité et de presse depuis de nombreuses années. Le rédacteur-documentaliste peut ainsi rédiger ou du moins participer à la rédaction d'un certain type de dossiers pour lesquels il collecte et sélectionne les informations nécessaires.

C'est, encore une fois, en s'approchant davantage des besoins des utilisateurs que les documentalistes peuvent occuper, eux-aussi, des postes à responsabilités importantes et décisionnelles dans la gestion de l'entreprise, tout en restant documentalistes. Dans ce contexte, le métier de documentaliste n'est plus considéré comme un service qui doit trouver

l'information pertinente, mais comme celui qui la gère et l'organise, afin de créer des services et des produits nouveaux, attractifs et attendus de la part des utilisateurs.

Conclusion

La fonction des documentalistes change et avec elle leur place dans les entreprises. Le documentaliste inséré dans le service de la documentation d'un titre de presse n'est plus obligatoirement et uniquement au service des journalistes, mais il peut aujourd'hui participer pleinement à la création de la valeur ajoutée à travers le site d'information, en offrant des services et en proposant des produits destinés directement aux lecteurs du journal. Ce nouveau rôle est indéniablement un apport pour n'importe quelle organisation, y compris pour un titre de presse.

La question qui se pose est celle du statut et de la place qu'une entreprise donne à ces professionnels de l'info-doc. L'entreprise et particulièrement ses principaux décideurs sont-ils toujours conscients de ce que les documentalistes peuvent apporter à leur structure? Il est certain que le cas de Libération n'est pas un cas isolé. Il fait partie de ces nombreuses entreprises où, à travers des plans sociaux, inévitables ou non, préserver les postes des documentalistes ne constituait pas une priorité. Les documentalistes, très appréciés certes par certains journalistes et informaticiens, ne peuvent pas aujourd'hui bénéficier d'une reconnaissance plus large, car ils ne sont pas impliqués dans le processus de production comme ils le souhaiteraient, ou en tout cas, comme ils le pourraient. Mais pour qu'ils soient motivés dans l'appropriation d'un nouveau rôle, une approche nouvelle semble indispensable.

Mais est-ce aux documentalistes eux-mêmes de convaincre de leur utilité, de leur rôle stratégique et de leurs compétences indispensables dans la structuration et l'organisation du système informationnel ? Ne serait-ce pas davantage l'approche générale des autres services de l'entreprise qui devrait changer radicalement ? A mon sens, les deux seront nécessaires pour qu'un regard nouveau soit posé sur le métier de documentaliste. Un regard à la fois plus profond dans le sens de la reconnaissance du savoir-faire que les documentalistes maîtrisent, et plus moderne du fait de ce que les nouvelles technologies représentent pour l'activité documentaire.

Le projet web est un travail d'équipe : ce n'est pas nouveau et en théorie tout le monde le sait. En revanche, il n'est peut-être pas notoirement connu que les documentalistes peuvent et doivent y participer. Et puis, c'est à travers la recherche des produits sur mesure pour l'utilisateur, les compétences liées à la structuration du contenu et de la base de donnés, et la gestion du flux informationnel - si longtemps considérée comme la préoccupation des informaticiens avant tout - que consiste la valeur ajoutée des documentalistes qui doit être défendue par eux-mêmes et soutenue par les autres métiers, ainsi que par les directions en place.

En termes de compétences, il est certain qu'un documentaliste qui n'a jamais participé à la réalisation d'un vocabulaire contrôlé ou à sa mise à jour, aura besoin de quelques heures de formation et d'accompagnement par une personne spécialisée en la matière. C'est donc également et nécessairement à travers des moyens déployés pour la formation et la conduite d'accompagnement au changement, et à travers un investissement rationnel et efficace, que la valeur ajoutée des professionnels de l'info-doc pour l'entreprise sera acquise.

Certains secteurs ont sans doute mieux compris que d'autres l'utilité des professionnels de l'info-doc ; le secteur scientifique par exemple. Etant de par sa nature même très structuré et rationnel, il est par conséquent plus enclin à tirer profit des compétences des documentalistes. Des secteurs comme celui de la presse, en France, ne considèrent pas toujours que l'activité des documentalistes pourrait, elle-aussi, être dans de nombreux cas stratégique et profitable, et participer activement à la gestion de l'entreprise.

Bibliographie

La présente bibliographie analytique est constituée autour de 6 thématiques :

- Thésaurus et langages documentaires
- Indexation(s)
- Taxonomie/taxonomie de navigation
- Technologies et web sémantique
- Presse et édition
- Activités et domaines des professionnels de l'info-doc.

Certains documents traitent naturellement plus qu'une de ces thématiques ; dans ce cas-là, ils sont répartis en fonction de la thématique la plus utilisée pour la rédaction du mémoire.

La bibliographie a été arrêtée le 15 septembre 2012.

Thésaurus et langages documentaires

[1] RABAULT Hélène. Langages documentaires et langage naturel. Support de cours INTD, 2011-2012, Document PowerPoint, 76 p.

Document définissant le langage documentaire par rapport au langage naturel. La deuxième partie du document concerne le thésaurus, sa définition, les étapes de la construction et ses relations. L'auteur donne quelques exemples des thésaurus en ligne.

[2] DALBIN Sylvie. Norme ISO 25964 : Présentation à l'attention d'éditeurs de logiciel de la future norme. Thésaurus pour la recherche documentaire. Elaboré par le groupe de travail sur la norme. 15 mars 2011. Document Powerpoint in Slideshare. [consulté le 2 octobre 2012]
http://fr.slideshare.net/Dalb/iso-25964-thsaurus-pour-la-recherche-documentaire

Le document explique les objectifs de l'élaboration de la nouvelle norme et l'importance du thésaurus pour la recherche documentaire.

[3] DALBIN Sylvie. ISO TC46-Information et documentation, ISO 25964 – Thésaurus et interopérabilité avec d'autres vocabulaires, Thésaurus pour la recherche documentaire, Présentation du projet de norme. Version du 21 novembre 2009, document Powerpoint in Slideshare. [consulté le 2 octobre 2012].
http://fr.slideshare.net/Dalb/presentation-du-projet-de-norme-iso-dis-259641-sur-les-thsaurus

Le document rend compte des différences entre les anciens documents normatifs et la nouvelle norme de la conception et la maintenance d'un thésaurus. Il met l'accent sur la notion de l'interopérabilité.

[4] DALBIN Sylvie. Thésaurus et informatique documentaires. Partenaires de toujours ? Documentaliste. 2007, n°1 – vol. 44. Pages 42-55.

L'article riche sur le sujet du thésaurus utilisé en informatique documentaire. Il est intéressant pour la compréhension des usages et fonctionnalités logicielles. L'auteur pointe les nouveaux enjeux des vocabulaires contrôlés à travers les technologies actuelles.

|5] DALBIN Sylvie. Thésaurus et informatique documentaires. Des Noces d'Or. Documentaliste. 2007, n°1-vol. 44. P. 76-80.

L'article est construit autour de l'usage du thésaurus et des vocabulaires contrôlés face à la technologie et sous l'angle historique.

[6] MANIEZ Jacques. Les langages documentaires et classificatoires : conception, construction et utilisation dans les systèmes documentaires. Les éditions d'organisation. Paris, 1987. 291 p. ISBN 2-7081-0833-6.

L'ouvrage décrit les grandes classifications et leur apport dans l'activité documentaire. Il retrace la naissance et l'évolution des langages documentaires, et particulièrement des langages d'indexation, dont le thésaurus, et les replace dans le contexte des systèmes documentaires.

[7] WILL Leonard. The ISO 25964 Data Model for the Structure of an Information Retrieval Thesaurus. Bulletin ASIS&T. [en ligne]. Avril-mai 2012 [consulté le 3 octobre 2012]. http://www.asis.org/Bulletin/Apr-12/AprMay12_Will.html

Avec la publication de la nouvelle norme du thésaurus, l'article présente, dans le contexte international et anglo-saxon, les principales différences par rapport aux documents normatifs précédents ainsi que les raisons conduisant à son élaboration et liées aux enjeux des nouvelles technologies dont celle du web sémantique.

[8] LACHANA Evanghélia. La mise à jour d'un thésaurus : éléments et propositions de méthode à partir de la mise à jour du thésaurus du Centre de documentation sur la formation et le travail du CNAM. 2001. Mémoire INTD. 2001.

Le mémoire développe le projet de la mise à jour du thésaurus. Il ne s'agit pas du même contexte que celui de la presse, mais l'auteur y donne des informations intéressantes sur les étapes liées au projet de la mise à jour d'un langage documentaire. Le mémoire aborde également les aspects technologiques liés aux fonctionnalités de la gestion informatique et particulièrement du logiciel Alexandrie.

[9] CHAUMIER Jacques. Les ontologies, antécédents, aspects techniques et limites. Documentaliste. 2007, vol. 44, n° 1. P. 81-83.

L'article définit les ontologies, fait le rapprochement avec le thésaurus et conçoit la construction des ontologies comme la modélisation des langages formels.

[10] RABAULT Hélène, ZYSMANN Hélène. Une nouvelle norme pour le thésaurus : modèle et perspectives à l'ère du web sémantique. ADBS. Février 2011, [en ligne], présentation PowerPoint in Slideshare [consulté le 2 octobre 2012]. http://fr.slideshare.net/ADBS/adbs-57-thesaurus-1

La présentation insiste sur la nécessité d'une nouvelle norme dans le contexte du web sémantique.

[11] DEFRICHES DORIA Orélie. La recherche d'information du grand public : l'évolution des langages documentaires ou l'avenir des ontologies. Création d'un thésaurus dynamique pour le site d'assistance en de France Télécom Orange. 2008. 115 p. Mémoire INTD. 2008.

Le mémoire utile pour l'historique et les fonctions des langages documentaires et riche dans l'approche du projet de créer un thésaurus pour la recherche d'information sur le site d'assistance de France Télécom Orange.

Indexation(s)

[12] VAN SLYPE Georges. Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires. Les éditions d'organisation. Paris, 1987. 277 p. ISBN 2-7081-0760-7.

Un ouvrage toujours utile pour la définition des langages d'indexation : leur typologie, les caractéristiques et leurs conceptions.

[13] RAïS Nadia. Identifier et décrire une ressource : de l'ISBD aux métadonnées. Documents imprimés et électroniques. Support de cours INTD-CNAM. Novembre 2011.

Cette note technique donne des notions essentielles de la description d'un document et permet de comprendre l'impact des métadonnées dans la description des ressources électroniques.

[14] QUESNEL Odile, FRANCIS Elie. Indexation collaborative et folksonomies, Documentaliste. 2007, vol. 44, n° 1, p. 58-63.

Les auteurs décrivent quatre modes d'indexation collaborative, très utilisée sur le web mais ayant ses limites dans la structuration des données dans les organisations.

[15] CHARTRON Ghislaine, DALBIN Sylvie, MONTEIL Marie-Gaëlle, VERILLON Monique. Indexation manuelle et indexation automatique. Dépasser les oppositions. Documentaliste. 1989, vol. 26, n° 4-5. P. 181-187.

Cette étude de comparaison des deux indexations, à travers le thésaurus EDF et un système d'indexation automatique (LEXINET), expose des avantages et des inconvénients de chacune. Ce document publié en 1989 est la preuve que certains professionnels de l'information pointent depuis bien longtemps la nécessité d'articuler les indexations manuelle et automatique alors qu'en 2012 cette complémentarité n'est pas toujours vue comme étant utile et indispensable.

[16] DADOU Nathalie. Indexation pour le web : usages et applications au fonds documentaire des éditions Techniques de l'Ingénieur. 2011. 105 p. Mémoire INTD. 2011.

Le mémoire rappelle l'importance de l'indexation et son utilité croissante dans le contexte du web. L'auteur dans ce sens-là développe la réflexion autour de la structuration du fonds documentaire de l'éditeur, naturellement différent de la presse, mais ayant le même objectif qui est celui de faciliter l'accès à l'information.

Taxonomie/taxonomie de navigation

[17] 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 histoiredesarts.culture.fr. 2011. 118 p. Mémoire INTD. 2011.

Avant de développer la notion de l'indexation sémantique, l'auteur rappelle la fonction des langages documentaire, décrit les origines de la « taxonomie » et regroupe plusieurs usages de ce terme.

[18] LE TARGAT Gaëlle. Langages classificatoires et recherche d'information sur les portails d'entreprises : quels apports pour les salariés ? Les taxinomies du portail Intralignes d'Air France. 2005. 168 p. Mémoire INTD. 2005.

Le mémoire s'interroge sur les besoins en accès à l'information sur le web et en entreprises en s'appuyant sur le projet de la mise en place de la taxonomie pour le portail d'Air France.

[19] MONDECA site. Les Taxonomies de navigation – La recherche à facettes : Définition, utilisation, objectifs, mise en œuvre. [en ligne] In site Mondeca. [consulté le 4 octobre 2012].
http://mondeca.wordpress.com/2007/10/07/les-taxonomies-de-navigation-la-recherche-a-facettes-definition-utilisation-objectifs-mise-en-oeuvre/

Les taxonomies de navigation - en tant que solution pour l'organisation du contenu des sites web - font l'objet du développement des outils logiciels. En France, la société Mondeca est bien placée pour son appui en matière de modélisation des taxonomies.

[20] RICCI Christian. Developing and Creatively Leveraging Hierarchical Metadata and Taxonomy. Boxesandarrows. [en ligne]. 22 mai 2004. [consulté le 5 octobre 2012].
http://www.boxesandarrows.com/view/developing_and_creatively_leveraging_hierarchical_metadata_and_taxonomy

Organisation des métadonnées autour des taxonomies, tel est le sujet de cet article traité du point de vue ergonomique, l'angle indispensable dans la création des sites web et qu'un architecte de l'information (*information architect*) doit prendre en compte. La bonne navigation sur le site demande une organisation du contenu qui est gérée sous deux aspects essentiels :

1) construire le site sur la base d'une structure universellement reconnue (*Universal Hierarchy*)
2) cartographier le contenu (*Content Mapping*)

[21] MINKOVSKY Natalya. Taxonomy : Content Strategy's New Best Friend. [en ligne] Johnny Holland, 18 avril 2012 [consulté le 5 octobre 2012].
http://johnnyholland.org/2012/04/taxonomy-content-strategys-new-best-friend/

Un autre article avec l'approche ergonomique mais faisant référence à l'indispensable recours aux architectes de l'information.

Technologies et web sémantique

[22] DEBONNE Eric, Formation Moteurs de recherche CNAM-INTD. 2012. Support de cours, 2 documents en fichier PowerPoint.

Documents permettant de comprendre les bases des principes et des technologies utilisés dans le développement des moteurs de recherche. L'auteur initie aux compétences nécessaires pour le choix de solution devant les besoins et objectifs précis en exposant les points essentiels pour l'élaboration d'un cahier des charges.

[23] CHAUMIER Jacques, DEJEAN Martine. Recherche et analyse de l'information textuelle. Tendances des outils linguistiques. Documentaliste. 2003, n° 1-vol. 40, p. 14-24.

Le document date en termes des logiciels répertoriés sur le marché français, mais il reste intéressant pour comprendre le principe des outils d'analyse textuelle à travers l'approche du traitement linguistique de l'information utilisé dans les logiciels d'indexation.

[24] PIERRE Béatrice. L'avenir des langages documentaires dans le cadre du web sémantique : conception du thésaurus iconographique pour le Petit Palais. 2010. 115 p. Mémoire INTD. 2010.

Le mémoire étudie la transposition de l'usage des langages documentaires dans le contexte des nouvelles technologies liées au web et particulièrement au web sémantique. L'auteur donne les notions essentielles propres aux topic maps ainsi qu'aux folksonomies et s'interroge sur les bénéfices de l'articulation des indexations libre et contrôlée dans le contexte culturel.

[25] CHARLET Jean, AUSSENAC-GILLES Nathalie, LAUBLET Philippe, BACHIMONT Bruno, VANDENBUSSCHE Pierre-Yves. Ontologie, terminologie, web sémantique. 8 février 2012. Support de cours INTD, fichier PowerPoint. 140 p.

Le cours donne une vision historique des ontologies et présente la méthodologie pour leurs constructions. Le document aborde, à travers le chapitre « Thésaurus *versus* ontologie formelle », l'articulation de la structure d'un thésaurus avec la description formelle d'une ontologie et rappelle les perspectives des ontologies pour le web sémantique.

[26] CONTAT Odile. Langages documentaires et nouvelles technologies : l'avenir des langages et leur positionnement au cœur des systèmes d'informations dans le contexte de la presse. 2003. 89 p. Mémoire INTD. 2003.

L'auteur définit les fonctions des langages documentaires et expose le fonctionnement de l'indexation automatique. Elle décrit l'accès à l'information dans le contexte de la presse et donne des exemples des différents systèmes d'informations (*Bayard*, *Le Monde*, *Les Echos* et *Le Nouvel Observateur*). Le mémoire insiste sur la nécessité de l'indexation humaine non seulement en tant qu'un besoin indispensable pour structurer le contenu, mais également en tant que la nécessité liée à la préservation de la connaissance.

[27] Web sémantique, web de données. Dossier in Documentaliste. Bruno Bachimont, Fabien Gandon, Gautier Poupeau, Bernard Vatan, Raphaël Troncy, Stépane Pouyllau, Ruth Martinez, Michèle Battisti, Manuel Zacklad, Sylvie Dalbin, Emmanuelle Bernès, Antoine Isaac, Romain Wenz, Yann Nicolas, Tayeb Merabti, Anila Angeli, Dominique Cotte, collab. 2011, vol. 48, n°4, p. 24-61.

Le dossier complet consacré aux notions principales du web de données et à ses enjeux dans l'activité documentaire. Les différents articles du dossier sont constitués autour de deux grandes parties :

1) *Enjeux et technologies : des données au sens*
2) *Approches documentaires : priorité aux contenus*

[28] AMAR Muriel, MENON Bruno. Web sémantique, web de données : quelle nouvelle donne ? Dossier in Documentaliste. 2012, vol 48, n° 4. P. 20-76.

Ce deuxième dossier sur le web sémantique rappelle ses principes et insiste sur les enjeux dans les organisations et entreprises ainsi que dans le secteur culturel. Le dossier pointe le principe fondamental du web sémantique qui est celui de l'ouverture des données.

[29] Bien mener son projet Web, guide pratique. Patrice Bourlon, Guillaume Nuttin, Xavier Delengaigne, Benoît Neuts, Dominique Guenuer, Christophe Dutheil, Michel Remize, Bruno Texier, collab. Archimag. 2012. N° 44. 48 p.

Ce guide pratique édité par Archimag est entièrement consacré aux aspects autour d'un projet web : compétences et connaissances nécessaires pour la gestion de projet, notions technologiques indispensables à connaître pour un professionnel de l'information, retours d'expérience (Bayard, Rue89, Vigipalia Soin palliatif).

[30] CHAUSSANEL Jean, CAHIER Jean-Pierre, ZACKLAD Manuel, CHARLET Jean. Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web Sémantique ? Conférence Ingénierie des Connaissances, IC 2002, [en ligne], mai 2002, Rouen, [consulté le 3 octobre 2012]. http://perso.limsi.fr/Individu/turner/Master2005_UMLV/Zacklad_WebSemantic.pdf

Les auteurs traitent les aspects propres au développement et à la modélisation des topic maps (cartes topiques) en tant que le formalisme de la représentation des connaissances, comme étant l'un des objectifs du web sémantique.

[31] ZACKLAD Manuel. Classification, thésaurus, ontologies, folksonomies : comparaison du point de vue de la recherche ouverte d'information (ROI). Congrès annuel de l'Association Canadienne de l'Information. Montréal. 2007.

L'auteur compare différents systèmes d'organisation des connaissances qui s'appuient sur le thésaurus, les ontologies mais aussi sur les folksonomie, dans la recherche documentaire et la recherche d'information en général.

[32] ZACKLAD Manuel. Evaluation des systèmes d'organisation des connaissances. Lavoisier. Cahiers du numérique. 2010/3 – vol. 6. P. 133-166.

Après avoir rappelé les principaux systèmes d'organisation des connaissances (SOC), l'auteur voit la nécessité d'y ajouter les « SOC automatiques » (fondés sur les index des moteurs de recherche). Il fait un comparatif entre les SOC fondés sur le thésaurus, les ontologies et les techniques collaboratives.

[33] CHATEAURAYNAUD Francis. Moteurs de (la) recherche et pragmatique de l'enquête, Les sciences sociales face au web connexioniste. Documentaliste. 2006/2 – n° 82. P 109-118.

L'auteur s'interroge sur les différentes façons dont on envisage et mène la recherche et comment l'Internet et le web influencent nos comportements. L'article est intéressant pour sa réflexion autour de la toute puissante technologie, en particulier informatique, et qui ne doit pas s'attaquer à l'organisation rationnelle du savoir et des connaissances.

Presse et édition

[34] MARINO Cristina. De la presse écrite à la presse électronique. Vers un nouveau média ? ADBS Editions. 1996. 143 p. ISBN 2-901046-94-0.

Cet ouvrage date, mais rend compte de la vitesse et des changements avec lesquels les sites d'information de la presse se sont en quinze ans développés et ont trouvé leur place dans la

vie quotidienne des lecteurs. L'ouvrage rappelle les premiers sites (dont celui de *Libération*) ainsi que leurs contenus et leurs stratégies éditoriales. Malgré le fait que la question de l'accès à l'information grâce aux vocabulaires contrôlés n'y est pas posée, on constate que le débat autour des versions papier et/ou électronique n'est pas aujourd'hui terminée.

[35] TRONCY Raphaël. Explorer les actualités multimédia dans le web de données. 20èmes journées Francophones d'Ingénierie des Connaissances (IC 2009), [en ligne]. Hammamet, Tunisie, 2009. [consulté le 3 octobre 2012].
http://ic2009.inria.fr/docs/papers/Troncy_IC2009_58.pdf

L'auteur retrace l'historique des efforts liés à la structuration des métadonnées de la presse, notamment à travers l'IPTC, et expose les problèmes d'interopérabilité dont la solution sera dans la formalisation des connaissances et le web de données.

[36] CONSTANT Jérôme. Les services d'accès à l'information de presse en ligne : de l'innovation à l'usage. Le cas du site web Lesechos.fr. 2008. 157 p. Mémoire INTD. 2008.

Le mémoire traite de la problématique des sites d'information de la presse d'actualité et se concentre prioritairement sur « la stratégie des services » proposés aux lecteurs. L'auteur met l'accent sur l'importance d'accompagner le lecteur dans la recherche d'information.

[37] ALBERT Pierre, La presse française. La Documentation française. Paris, 2008. 215 p. ISSN 1763-6191.

Un ouvrage qui dresse un tableau sur la complexité de la presse en France, s'interroge sur ses fonctions, ses nouveaux concurrents, ses lecteurs ainsi que sur son rôle et ses modalités après la révolution de l'Internet.

Activités et domaines des professionnels de l'info-doc

[38] RANJARD Sophie. Vers une politique de produits et services documentaire. Support de cours INTD. Novembre 2011, document PowerPoint.

Ce support de cours permet de se faire la première idée de la palette des différents services offerts par les professionnels de l'information.

[39] CHAUMIER Jacques, SUTTER Eric. Documentalistes, ajoutez de la valeur à vos services ! ADBS éditions. 2007. 63 p. ISBN 978-2-84365-095-6.

Un ouvrage autour de l'évolution du métier de documentaliste à l'ère numérique et des opportunités à saisir dans la proposition de nouveaux services.

[40] Architecture de l'information. Colloque Jacques Quartier. [en ligne] in site ENS de Lyon. [consulté le 5 octobre 2012].
http://archinfo.ens-lyon.fr/colloque-international-148317.kjsp?RH=ARCHINFO_PRES&RF=1333776895746

Le mouvement des architectes de l'information, lié à la prolifération des sites web, existe et se développe depuis 15 ans (!) aux Etats-Unis et dans les pays anglo-saxons. Il commence à se déployer en France et a contribué à la création d'une nouvelle formation « Master Architecture de l'information » à l'ENS de Lyon.

Annexes

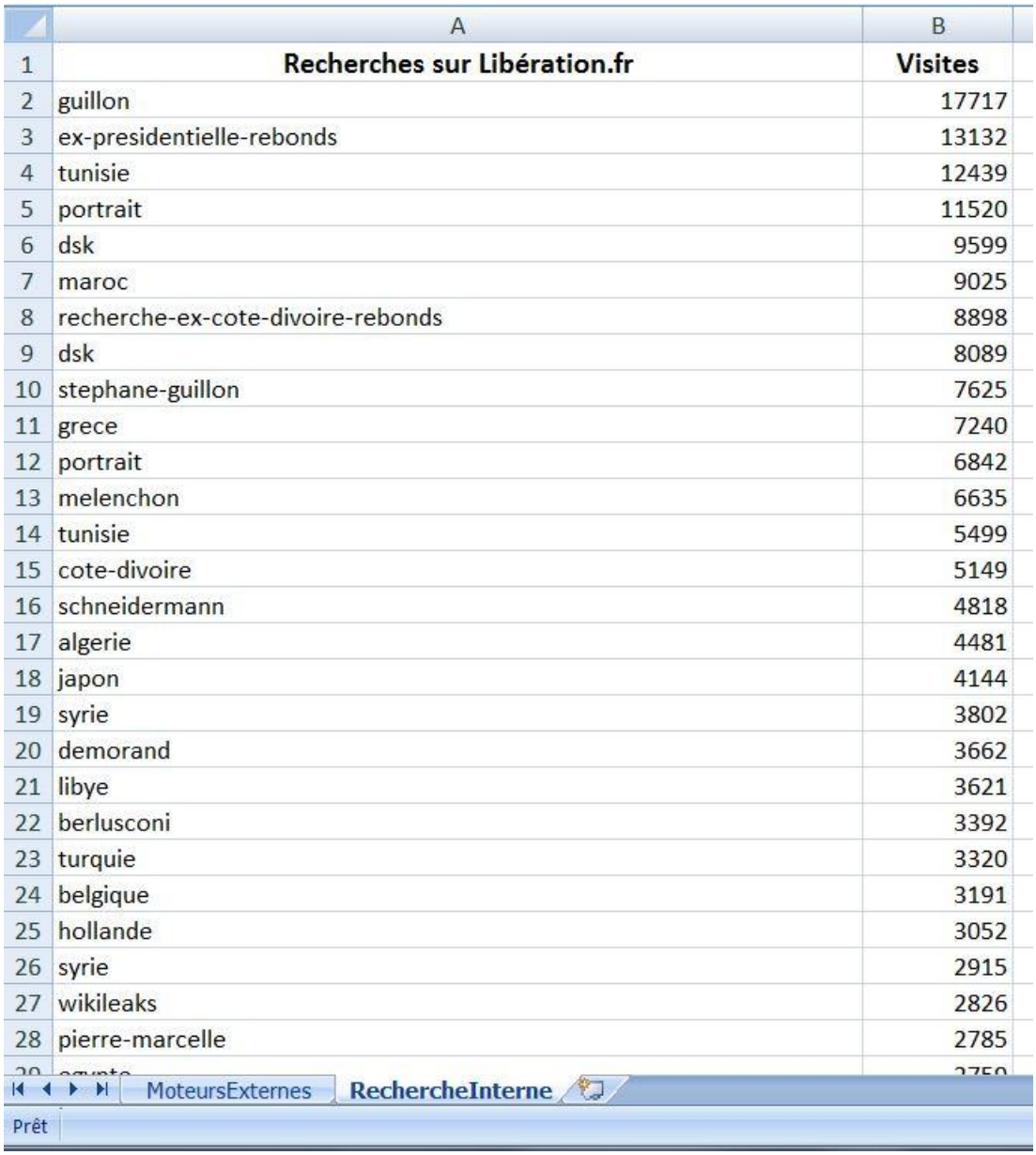
Annexe 1 Requêtes des utilisateurs

Extrait de la liste des requêtes dans les moteurs de recherche externes pendant la période du 01/07/2010 et 30/03/2012.

	A	B
1	**Expressions clés Moteurs de recherche**	**Visites**
2	dsk	223174
3	actualité	147563
4	marine le pen	105801
5	mélenchon	94187
6	actualités	86870
7	tunisie	83987
8	sarkozy	81167
9	psg	80684
10	syrie	70257
11	iran	65208
12	hollande	57399
13	kadhafi	56688
14	actualites	55715
15	free mobile	55442
16	melenchon	55258
17	sondages	54892
18	sondage	49856
19	japon	47373
20	actu	45241
21	facebook	44278
22	costa concordia	42772
23	fukushima	42543
24	toulouse	39993
25	bayrou	33986
26	meteo france	33499
27	silence on joue	29618
28	maroc	29496

MoteursExternes | RechercheInterne

Extrait de la liste des requêtes sur le site liberation.fr pendant la période du 01/07/2010 et 30/03/2012.



	A	B
1	**Recherches sur Libération.fr**	**Visites**
2	guillon	17717
3	ex-presidentielle-rebonds	13132
4	tunisie	12439
5	portrait	11520
6	dsk	9599
7	maroc	9025
8	recherche-ex-cote-divoire-rebonds	8898
9	dsk	8089
10	stephane-guillon	7625
11	grece	7240
12	portrait	6842
13	melenchon	6635
14	tunisie	5499
15	cote-divoire	5149
16	schneidermann	4818
17	algerie	4481
18	japon	4144
19	syrie	3802
20	demorand	3662
21	libye	3621
22	berlusconi	3392
23	turquie	3320
24	belgique	3191
25	hollande	3052
26	syrie	2915
27	wikileaks	2826
28	pierre-marcelle	2785

Annexe 2 Référentiels existants

Ci-dessous quelques référentiels (thésaurus et vocabulaires structurés) évalués par rapport aux besoins du journal Libération et censés pouvoir servir dans la construction et la modélisation de la taxonomie du back-office du journal.

EUROVOC

Eurovoc est le thésaurus multilingue couvrant toutes les activités de l'Union européenne (disponible en 22 langues) avec un accent sur l'activité parlementaire.

Les domaines principaux d'Eurovoc :

- Vie politique
- Relations internationales
- Communautés européennes
- Droit
- Vie économique
- Echanges économiques et commerciaux
- Finances
- Questions sociales
- Education et communication
- Sciences
- Entreprise et concurrence
- Emploi et travail
- Transports
- Environnement
- Agriculture, sylviculture et pêche
- Agro-alimentaire
- Production, technologie et recherche
- Energie
- Industrie
- Géographie
- Organisations internationales

Tous les domaines d'Eurovoc peuvent être utiles pour affiner le référentiel de Libération car les concepts représentant les discours de la presse y sont présents.

Le domaine Organisations internationales permettrait de structurer et de développer le domaine des Institutions et des organismes qui constitue l'une des faiblesses du thésaurus Libération. Ce domaine est dans Eurovoc construit autour de 5 catégories principales :

- Nations Unis
- Organisations européennes
- Organisations extra-européennes
- Organisations mondiales
- Organisations non gouvernementales

MOTBIS

Motbis est le thésaurus du ministère de l'Education nationale. Motbis étant représentatif des sciences de l'éducation, offre un vocabulaire couvrant des principales activités de la société dont les domaines susceptibles de pouvoir enrichir ceux du vocabulaire contrôlé d'un quotidien de presse, par exemple :

- politique
- vie politique
- population et société
- économie
- économie internationale
- entreprise
- nature et environnement
- démographie et population

Listes annexes :

- lieux
- France
- sites géographiques

UNBIS

Le Thésaurus multilingue UNBIS est un outil multidisciplinaire, abordant l'ensemble des domaines d'action des Nations Unis.

Domaines principaux pouvant également servir dans la construction du futur vocabulaire :

- Questions politiques et juridiques
- Développement économique et financement du développement
- Ressources naturelles et environnement
- Agriculture, foresterie et pêche
- Industrie
- Transports et communications
- Commerce international
- Population
- Etablissements humains
- Santé
- Education
- Emploi
- Assistance humanitaire
- Questions sociales
- Culture
- Sciences et techniques
- Descripteurs géographiques
- Questions organisationnelles

L'UNESCO

Le thésaurus de l'UNESCO contient des domaines suivants :

- Education
- Science
- Culture
- Sciences sociales et humaines
- Information et communication
- Politique, droit et économie

BDSP

BDSP est le thésaurus de la Santé Publique. Domaines pouvant enrichir ou réactualiser le référentiel Libération :

- Politique santé
- Politique sociale
- Protection sociale
- Sciences éducation

Blog France politique

Blog France politique est élaboré par Laurent de Boissieu (journaliste politique) : site de référence avec non seulement les partis politiques actuels mais également l'historique des partis et de la vie politique française.

Partis politiques françaises : liste Wikipédia

Il s'agit de la liste des partis politiques actuels (y compris avec des sigles). Cette partie est scindée en trois groupes : les partis parlementaires, les partis non parlementaires et les partis locaux (régionaux).

Entreprises : liste Wikipédia

- classement par les plus grandes entreprises
- par pays (par exemple, la liste des 35 plus grandes entreprises françaises)

Entreprises : liste Fortune global 500

Fortune global 500 est une liste de 500 entreprises mondiales classées selon leur chiffre d'affaires. Elle est publiée chaque année par le magazine Fortune qui publie également la liste Fortune 500 (uniquement les entreprises américaines).