Documentation et bibliothèques

Exploitation d'une cartographie sémantique à des fins de validation : application à l'indexation experte de corpus documentaires

The Use of Semantic Cartography for Validation: The Application of Indexation Expertise to Documents

Utilización de una cartografía semántica con fines de validación: aplicación de la indexación experta en corpus de documentos

Eric Kergosien, Marie-Noelle Bessagnet and Mauro Gaio

Volume 57, Number 1, January–March 2011

URI: https://id.erudit.org/iderudit/1028961ar
DOI: https://doi.org/10.7202/1028961ar

See table of contents

Publisher(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (print)
2291-8949 (digital)

Explore this journal

Cite this article

Kergosien, E., Bessagnet, M.-N. & Gaio, M. (2011). Exploitation d'une cartographie sémantique à des fins de validation : application à l'indexation experte de corpus documentaires. *Documentation et bibliothèques*, *57*(1), 19–28. https://doi.org/10.7202/1028961ar

Article abstract

The authors' intention is to support the work of indexation by librarians using concepts from a controlled vocabulary with the relationship of meanings in the descriptive cataloguing. During the course of our research, we defined the TERRIDOC thesaurus using its specifics of terms that "make sense" to the librarian when drafting the descriptive cataloguing. During the conceptualisation, the different modules were defined in order to automatically create the semantic structure representing the work of indexation experts. Then, interfaces were developed that visually represent this work in order to validate the indexation done by librarians and to navigate in the library holdings. Following the analysis and creation of this first instrument, we suggest means to control and validate the indexation. Our work is part of the field of document classification and "networked" users.

DOCUMENTATION BIBLIOTHÈQUES

Exploitation d'une cartographie sémantique à des fins de validation : application à l'indexation experte de corpus documentaires

ERIC KERGOSIEN
UPPA, Domaine Universitaire Laboratoire LIUPPA
Faculté des Sciences, PAU
eric.kergosien@univ-pau.fr

MAURO GAIO
UPPA, Domaine Universitaire Laboratoire LIUPPA
Faculté des Sciences, PAU
mauro.Gaio@univ-pau.fr

MARIE-NOELLE BESSAGNET
UPPA, Domaine Universitaire Laboratoire LIUPPA
IAE, PAU
marie-noelle.bessagnet@univ-pau.fr

RÉSUMÉ | ABSTRACTS | RESUMEN

Notre approche vise à aider le travail d'indexation des bibliothécaires via les concepts provenant d'un vocabulaire contrôlé par des relations de sens contenues dans les notices descriptives. Dans notre travail de recherche, nous définissons automatiquement le thésaurus TERRIDOC en exploitant les spécificités du corpus liées à des termes qui « ont fait sens » au bibliothécaire lors de la constitution de la notice descriptive. Une phase de conceptualisation a permis dans un premier temps de définir en détails les différents modules pour créer automatiquement la structure sémantique représentant le travail d'indexation des experts, puis dans un second temps de proposer des interfaces représentant visuellement ce travail à des fins de validation du travail d'indexation des bibliothécaires et de navigation dans le fonds documentaire. À la suite de l'analyse et de la création de ce premier outil, nous proposons des moyens de contrôler et de valider le travail d'indexation. Notre travail s'inscrit dans le champ de la classification documentaire et de ses usages en environnement « réseauté ».

The Use of Semantic Cartography for Validation : The Application of Indexation Expertise to Documents

The authors' intention is to support the work of indexation by librarians using concepts from a controlled vocabulary with the relationship of meanings in the descriptive cataloguing. During the course of our research, we defined the TERRIDOC thesaurus using its specifics of terms that "make sense" to the librarian when drafting the descriptive cataloguing. During the conceptualisation, the different modules were defined in order to automatically create the semantic structure representing the work of indexation experts. Then, interfaces were developed that visually represent this work in order to validate the indexation done by librarians and to navigate in the library holdings. Following the analysis and creation of this first instrument, we suggest means to control and validate the indexation. Our work is part of the field of document classification and "networked" users.

Utilización de una cartografía semántica con fines de validación : aplicación de la indexación experta en corpus de documentos

Nuestro enfoque tiene como objetivo colaborar con la tarea de indexación de los bibliotecarios mediante conceptos de un vocabulario controlado por relaciones de sentido incluidas en instrucciones descriptivas. En nuestro trabajo de investigación, definimos automáticamente el tesauro TERRIDOC, investigando las especificidades del corpus relacionadas con términos que « han adquirido sentido » para el bibliotecario durante la elaboración de las instrucciones descriptivas. Una fase de conceptualización permitió, en un primer momento, definir en detalle los diferentes módulos para crear automáticamente la estructura semántica que representa el trabajo de indexación de los expertos. En segundo lugar, se propusieron interfaces que representan visualmente este trabajo para validar el trabajo de indexación de los bibliotecarios y la navegación en los documentos. Luego del análisis y de la creación de esta primera herramienta, proponemos métodos para controlar y validar el trabajo de indexación. Nuestro trabajo se inscribe en el campo de la clasificación de documentos y de sus usos en un entorno de « red ».

Introduction

DANS LES BIBLIOTHÈQUES ET LES MÉDIATHÈQUES, le souci majeur a toujours été de structurer au mieux la connaissance afin d'en faciliter l'accès et le partage par une communauté d'utilisateurs. Ces institutions renferment des corpus documentaires conséquents qui deviennent de plus en plus facilement disponibles au format électronique, mais leur accessibilité, afin de comprendre ces données, reste encore problématique.

Nous proposons dans nos travaux de mettre en place un Système à Bases de Connaissances (SBC) que nous nommons TERRIDOC[1], offrant tout d'abord aux experts du domaine (les bibliothécaires) une représentation sémantique graphique synthétisant leur travail d'indexation. Nous proposons d'extraire automatiquement les éléments caractérisant un territoire à partir des

1. TERRIDOC fait partie d'un projet plus général (projet PIV) dont l'objectif général est de développer des techniques et des outils informatiques de marquage sémantique à des fins d'indexation par les contenus territorialisés des documents et de leurs notices associées. Ces outils permettront à un lecteur-utilisateur de découvrir et de s'approprier un document ou un ensemble documentaire dans une perspective « territoriale ». Cette finalité nécessite un marquage des contenus documentaires selon les deux facettes : spatiale et/ou temporelle.

documents et des notices descriptives qui leur sont attachées et qui contiennent des informations riches. Les notices descriptives sont réalisées à l'aide d'un thésaurus faisant autorité dans le milieu, le thésaurus RAMEAU[2] de la Bibliothèque nationale de France (BNF). Nous cherchons ensuite à proposer à tous types d'utilisateurs de naviguer dans un territoire à travers les documents. Le fonds documentaire sur lequel nous évaluons notre approche est constitué de différents types de documents (image, son, vidéo et texte) décrivant les Pyrénées entre le XVIII^{e} et le XX^{e} siècle. Ce choix n'est pas innocent car ces documents renferment de nombreuses références géographiques.

Pour répondre au premier usage consistant à proposer aux experts du domaine une représentation sémantique synthétique de leur travail d'indexation, nous proposons d'extraire des notices descriptives les résultats du travail d'analyse et de représentation de façon automatique plutôt que de traiter directement les documents associés.

Le deuxième usage assigné au SBC est de faire découvrir aux utilisateurs d'un centre documentaire un espace et une période historique implicitement décrits par les documents constituant le fonds. Nous proposons pour cela d'enrichir la structure sémantique obtenue par des éléments décrivant un territoire.

Nous pensons que la représentation cartographique des connaissances expertes extraites automatiquement des résultats du travail d'indexation apporte un premier élément de réponse aux attentes, en offrant une synthèse exhaustive d'un état de l'indexation de la base documentaire. Nous incorporons dans cette synthèse un ensemble de cas d'erreurs identifiés dans les notices descriptives. Nous nous appuyons ensuite dans notre démarche sur cette représentation cartographique pour proposer aux experts des outils permettant de valider de façon semi-automatique les termes employés dans les notices descriptives afin de décrire le contenu des documents qui sont identifiés comme potentiellement erronés. Nous nous situons dans une problématique de validation semi-automatique du travail manuel d'indexation de fonds documentaires en exploitant les connaissances expertes extraites automatiquement des notices, problématique basée sur deux étapes principales :

- Structuration du domaine du fonds documentaire (Tricot *et al*, 2006) : définition d'un thésaurus exploitant les spécificités du travail d'indexation liées à des termes qui ont du « sens » pour le lecteur-bibliothécaire ;
- Navigation et interrogation du fonds documentaire : proposition de techniques de visualisation adaptées aux besoins des experts. Dans ces objectifs, la visualisation de l'information apparaît comme l'une des voies les plus intéressantes pour « produire du sens » dans l'observation des masses de données. Norman (1993) souligne que l'utilisation d'aides externes et notamment visuelles augmente considérablement la puissance cognitive de l'esprit humain.

Dans cet article, nous énoncerons d'abord les problématiques et les objectifs de notre recherche. Nous expliciterons ensuite nos expérimentations liées à un outillage pour la navigation et l'exploration. Nous présenterons enfin nos propositions à des fins d'aide à la validation de l'indexation dans le corpus basées sur le thésaurus enrichi.

Problématique et objectifs

L'intérêt de disposer d'un fonds documentaire et de pouvoir ensuite proposer à des utilisateurs d'accéder aux informations nécessaires pour leur activité est primordial. Cela implique, d'une part, la nécessité d'identifier les informations pertinentes et, d'autre part, la possibilité de fournir des moyens pour y accéder. Il faut donc exploiter la structuration du fonds documentaire. Les possibilités pour organiser, classer et structurer un ensemble de documents sont nombreuses. Nous positionnons nos travaux dans la structuration sémantique de domaine, car nous faisons appel à des ressources externes sur lesquelles nous appuyer.

Dans le domaine de la gestion documentaire, l'usage des thésaurus à des fins d'indexation et de recherche d'information s'intensifie et il est courant, voire obligatoire, dans les systèmes de centres documentaires de type bibliothèques, médiathèques, etc.

Le travail réalisé par les bibliothécaires se décompose généralement en deux phases distinctes :

- une première phase définie par les bibliothécaires comme la phase de « description » consistant à décrire le document analysé en renseignant notamment le ou les auteur(s), le titre, la légende, le type de document, etc. ;
- la deuxième phase définie par les experts comme la phase d'« indexation » aboutissant à la sélection d'un ensemble de termes décrivant le contenu du document.

La première phase nécessite obligatoirement une saisie manuelle. La deuxième phase, plus contraignante, nécessite des connaissances du langage d'indexation RAMEAU et l'utilisation de la ressource thésaurus associée.

Le thésaurus est un outil qui, comme les listes d'autorités ou encore les lexiques d'indexation, fait partie de la famille des vocabulaires contrôlés permettant l'accès par sujet aux catalogues et aux bases de données bibliographiques (Nieszkowska, 2003). Un vocabulaire est dit contrôlé s'il a été soumis à trois niveaux de contrôle :

2. Répertoire d'autorité-matière encyclopédique et alphabétique unifié, accessible à l'URL <http://rameau.bnf.fr/>.

- le niveau sémantique : vérifie qu'un élément du vocabulaire contrôlé n'aura qu'un seul sens précis dans le vocabulaire d'indexation et que cet élément sera le seul à avoir ce sens. Cela implique le contrôle de la synonymie, de l'homonymie et de la polysémie ;
- le niveau terminologique : détermine le choix de la forme du terme (choix de la langue, du pluriel, de la spécificité plus ou moins technique du terme) ;
- le niveau syntaxique : donne la possibilité, lors de l'indexation, de définir des sujets nouveaux ou complexes.

Nous nous intéressons au thésaurus RAMEAU, utilisé notamment par la Bibliothèque nationale de France (BnF), par le Sudoc (catalogue collectif des bibliothèques universitaires), par un nombre important de bibliothèques spécialisées ainsi que dans les bibliothèques municipales et départementales françaises. RAMEAU est également utilisé à l'étranger, notamment au sein des bibliothèques publiques de la Communauté française de Belgique et à la Bibliothèque Nationale de Tunisie. Nous parlons ici de connaissances expertes plutôt que de connaissances métiers, car RAMEAU n'est pas défini pour un domaine précis : il couvre l'ensemble des disciplines scientifiques et contient aussi les termes traitant des loisirs, des arts, etc.

Les bibliothécaires doivent respecter les termes du thésaurus pour établir les notices descriptives. Cela peut s'avérer rapidement très contraignant car, à l'heure actuelle, un usager désirant consulter le thésaurus doit utiliser le moteur de recherche RAMEAU disponible sur le site de la BnF[3]. Lorsque, par exemple, un bibliothécaire indexe un document image présentant la façade de la station thermale de Barèges, il ira alors examiner via la recherche par mots-clés dans le catalogue en ligne les termes à utiliser. Dans ce cas précis, parmi les propositions présentes dans RAMEAU, les termes adéquats seront « Stations climatiques, thermales », etc., et Barèges (Hautes-Pyrénées). Cette sélection nécessite alors deux phases de recherche au catalogue distant (une par terme), ce qui est relativement fastidieux pour les bibliothécaires.

Il n'est pas toujours simple pour le bibliothécaire de trouver le terme adéquat pour décrire le contenu du document qu'il a en mains. Ainsi, on obtient au final des distorsions entre les divers documents, des termes erronés, etc. et donc des problèmes lors de la recherche d'information. La connaissance qu'a le bibliothécaire du vocabulaire utilisé pour indexer le fonds documentaire est un gage de bon usage et donc d'une indexation recevable et efficace.

À la différence des ressources disponibles sur le Web, une bibliothèque a pour rôle de contrôler, de valider et de préserver ses ressources selon des critères spécifiques (notamment ceux préconisés par RAMEAU), l'objectif étant d'assurer aux utilisateurs une qualité et une fiabilité d'usage tant au niveau des ressources que de l'accès à celles-ci.

Objectifs

Dans le but d'offrir aux experts bibliothécaires (notamment les bibliothécaires de la Médiathèque Intercommunale à Dimension Régionale (MIDR) de Pau et son agglomération) un outil de validation de l'utilisation du langage contrôlé mis en œuvre pour harmoniser la formulation des termes d'indexation, nous avons élaboré, dans notre démarche, deux phases préalables :

- extraction et structuration des connaissances du domaine du fonds documentaire ;
- navigation et interrogation du fonds documentaire avec représentation cartographique de ce dernier.

La première phase permet d'extraire et de structurer les connaissances de notre corpus composé de notices descriptives et du thésaurus RAMEAU à des fins de création d'un thésaurus TERRIDOC pour la recherche d'information.

L'une des difficultés majeures est d'appréhender de façon automatique à la fois la structure du thésaurus RAMEAU, défini pour traiter une multitude de domaines d'activités, ainsi que le langage spécialisé, associé au thésaurus, qui est utilisé par les bibliothécaires pour indexer les documents via les notices descriptives. Le vocabulaire constituant le thésaurus est contrôlé, ce qui nous permet de l'exploiter en nous basant sur les connaissances d'experts. Nous considérons ici que le thésaurus RAMEAU est constitué d'un vocabulaire contrôlé fermé car, même s'il est enrichi de façon périodique, nous l'utilisons, dans notre démarche, dans un état figé.

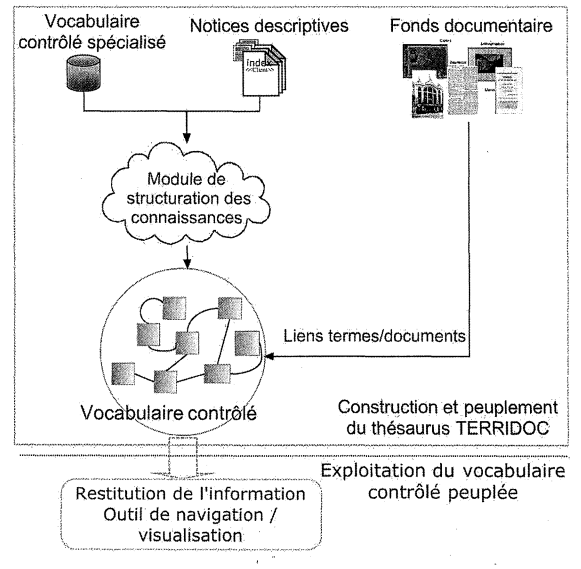
Dans ce travail d'extraction et de structuration, nous nous rapprochons des travaux de Schatz *et al.* (1996) qui se sont attachés à mixer les termes d'un thesaurus et des termes d'autres sources afin de mieux cibler la recherche d'information dans les systèmes documentaires. L'un de nos objectifs est la mise en place d'un processus pour passer d'un thésaurus classique à une base de connaissances (Elghoul, 1990). Ainsi, nous proposons de créer automatiquement une structure représentant, sous forme de thésaurus, le travail d'indexation des bibliothécaires en nous appuyant sur les notices descriptives pour identifier les termes, et sur RAMEAU pour extraire les relations entre ces termes. Notre corpus de référence est un ensemble fini constitué de près de 5 000 documents, ce qui permet de restreindre l'étendue des termes utilisés pour l'indexation. Nos premières expérimentations ont été menées

3. <http://catalogue.bnf.fr/jsp/recherche_autorites_rameau.jsp;jsessionid=oooonsXnlbFvITFOQM2IqwLD-5U:-1?nouvelleRecherche=O&host=catalogue>.

sur un corpus de 750 notices descriptives et leurs documents associés.

La deuxième phase consiste à proposer une représentation cartographique de ce thésaurus sous forme de carte de concepts. Barsalou (1999) montre que l'œil permet de percevoir un ensemble de signaux simultanément et d'effectuer un grand nombre de traitements instantanément avant même de mettre en place des mécanismes cognitifs comme le raisonnement ou la mémorisation. La carte représentée joue alors le rôle de support de la pensée (Card, Mackinlay et Shneiderman, 1999). De nombreuses solutions, dont le but est de faciliter la navigation visuelle au sein d'un fonds documentaire, existent. Card, Mackinlay et Schneiderman (1999) et Jacquemin *et al.* (2005) les divisent en deux groupes : les techniques de visualisation intra-documents et les techniques de visualisation d'un ensemble de documents. Nous souhaitons dans nos travaux représenter les connaissances extraites d'un ensemble de documents constituant le fonds documentaire. Nous mettons de côté les travaux cherchant à représenter la structuration de façon non traduite, tels ceux de Cubaud et Topol (2001) avec Bib3D, car nous cherchons à proposer aux experts une représentation visuelle permettant de comprendre et de valider la structuration qu'ils ont mise en place lors de la phase d'indexation. En cherchant à représenter dans notre structure les termes utilisés pour décrire le contenu des documents et les relations entre ces termes, nous nous rapprochons plutôt de la représentation de type nœud-lien telle que décrite par Tricot *et al.* (2006). Nous souhaitons ajouter à notre solution un module de représentation hiérarchique sous forme d'arborescence simple, afin de faciliter la navigation, en offrant à l'utilisateur en parallèle une vision globale de la structure. Zizi et Beaudoin-Lafon (1995) proposent une visualisation sous forme de cartes conceptuelles interactives dynamiques, où des informations (documents, mots-clés, thèmes, etc.) sont positionnées sur des cartes de telle sorte que les informations les plus similaires sont voisines. Parmi les divers types de cartes développés pour des usages divers, les cartes de thésaurus sont des cartes obtenues par l'analyse d'une collection de documents permettant l'extraction d'une structure de termes. Il en résulte un thésaurus visualisable sous forme de carte sémantique. La masse d'information contenue étant, pour des centres bibliographiques, souvent trop importante, nous nous orientons ensuite vers les techniques d'interactions (Zizi et Beaudoin-Lafon, 1995) pouvant être couplées avec la technique de visualisation pour faciliter l'appréhension des connaissances afin de combiner la vue globale (contexte) et la vue locale (focus). Au final, nous utilisons les techniques de cartographie sémantique (Kohonen, 2006) pour répondre, entre autres, au besoin de visualisation des informations pertinentes à plusieurs échelles : d'une part, sur la structure globale de notre carte de concepts et, d'autre part, sur une structure locale.

Figure 1
Méthodologie de construction de vocabulaire contrôlé.



Ainsi, l'interface graphique proposée permet aux bibliothécaires d'appréhender plus aisément les résultats du travail d'indexation ; elle permet également de mettre en avant des erreurs possibles de saisie et d'utilisation de RAMEAU. À terme, nous souhaitons offrir aux experts du domaine un outil de validation de l'utilisation du langage contrôlé (sous-ensemble RAMEAU enrichi de termes spécifiques au fonds documentaire) qu'ils ont exploité pour indexer les documents. On peut aider les experts du domaine pour l'attribution de thèmes décrivant les documents en vérifiant qu'ils proviennent de la ressource (le thésaurus RAMEAU ici). En effet, lors de nos discussions avec les professionnels de la Médiathèque Intercommunale à Dimension Régionale (MIDR) de Pau, nous avons noté plusieurs erreurs de traitement. Le bibliothécaire ne dispose pas de moyens efficaces pour visualiser le travail d'indexation de façon globale afin de contrôler et de faciliter l'action de validation. Il lui est difficile, voire impossible, d'identifier les éventuelles erreurs de saisie. Enfin, certains termes nécessaires à l'indexation du document peuvent être spécifiques au fonds documentaire et donc ne pas faire partie de RAMEAU ; on parle alors de termes candidats à l'intégration dans RAMEAU. L'outil de validation doit faire apparaître ces derniers.

TERRIDOC

Présentons à travers un exemple (Figure 1) la méthodologie adoptée pour enrichir un premier vocabulaire de termes provenant du thésaurus RAMEAU afin de créer un thésaurus adapté, l'objectif étant de

Tableau 1
Relations entre termes dans un thésaurus (d'après Hernandez 2005).

t_1 **Terme préféré dans** $t_2, t_3,\ldots,t_N$	t_1 est le terme préféré pour désigner l'ensemble des synonymes $t_2, t_3,\ldots,t_N$.	
t_1 **Note** texte	Remarque sur le terme t_1(usage exceptionnel, contexte d'utilisation)	NA
t_1 **Utiliser plutôt** t_2	t_2 est utilisé pour désigner t_1	EM
t_1 **Utilisé pour** t_2	t_1 est utilisé pour désigner t_2	EP
t_1 **Plus spécifique que** t_2	Le terme désigné par t_1 est plus spécifique que le terme désigné par t_2	TS
t_1 **plus générique que** t_2	Le terme désigné par t_1 est plus générique que le terme désigné par t_2	TG
t_1 **est lié à** t_2	t_1 est un terme lié ou associé à t_2	TA

représenter l'information à l'aide des connaissances expertes extraites automatiquement des notices.

Ressources

Dans notre démarche, nous nous appuyons sur la base de notices descriptives correspondantes aux documents ainsi que sur le thésaurus RAMEAU.

Selon l'AFNOR (L'Association française de normalisation) (1981), un thésaurus est un langage documentaire fondé sur une structuration hiérarchique d'un ou plusieurs domaines de connaissances et dans lequel les notions sont représentées par des termes d'une ou plusieurs langues naturelles et, les relations entre les notions, par des signes conventionnels. Le Tableau 1 présente les relations les plus usuelles.

Le thésaurus RAMEAU regroupe un ensemble de sujets et de relations entre ces sujets et, dans des notes d'application, les indications qui permettent de construire des « vedettes-matière » dans un fichier bibliographique. Il propose 256 000 termes, dont 88 000 noms communs et 46 000 noms géographiques. Les « vedettes-matière » sont les points d'accès au sujet dans les notices descriptives réalisées par les experts bibliothécaires, construites selon les règles et à l'aide des termes proposés dans RAMEAU.

RAMEAU est un langage d'indexation précoordonné qui ne doit pas être employé comme un ensemble de mots-clés dans un ordre aléatoire. Sa syntaxe est précise et doit être respectée pour assurer la cohérence et l'homogénéité de l'indexation. La tête de vedette, premier élément de la vedette-matière, exprime l'essence du sujet. Elle peut se suffire à elle-même ou être suivie de subdivisions de sujet, de lieu, de temps ou de forme. Chacun des éléments constituant une vedette-matière (tête de vedette et subdivisions éventuelles) est qualifiée d'« autorité-matière » dans RAMEAU.

Le thésaurus RAMEAU traite les homonymes et les synonymes. Les formes d'autorité constituant le thésaurus proviennent de plusieurs domaines de la connaissance reliés par des relations connues et arrêtées. Ainsi, à partir d'une autorité-matière, nous pouvons exploiter des relations de type :

- Employé pour ;
- Terme(s) générique(s) ;
- Terme(s) associé(s) ;
- Terme(s) spécifique(s).

Dans notre phase d'extraction et de structuration des connaissances, l'exploitation de ces relations nous permet de construire le thésaurus TERRIDOC.

Conception d'un thésaurus TERRIDOC

La première étape du traitement consiste à identifier et extraire automatiquement tous les termes (autorités-matières RAMEAU) utilisés pour décrire le

Figure 2
Extrait de notice descriptive 1.

<DOC_DEE>Stations climatiques, thermales, etc. -- Barèges (Hautes-Pyrénées) -- 18e siècle**</DOC_DEE>**

<DOC_DEE>Eaux minérales -- Pyrénées (France) -- 18e siècle**</DOC_DEE>**

<TITRE>Précis d'observation sur les eaux de Barèges et les autres eaux minérales de Bigorre et du Béarn **</TITRE>**

<LEGENDE>Médecin du XVIIIème siècle, Théophile de Bourdeu est à l'origine de la mode du thermalisme pyrénéen**</LEGENDE>**

<date>2007-04-16**</DATE>**

Figure 3
Extrait de notice descriptive 2.

<DOC_DEE>Œuvres scientifiques -- Stations climatiques, thermales, etc. -- Bagnères-de-Bigorre (Hautes-Pyrénées) -- 19e siècle**</DOC_DEE>**

<TITRE>Recherches sur les propriétés physiques, chimiques et médicinales des Eaux minérales de Bagnères de Bigorre**</TITRE>**

<GEOREF>Hautes-Pyrénées, Bagnères-de-Bigorre**</GEOREF>**

<date>2007-04-19**</DATE>**

contenu d'un document dans sa notice descriptive au format XML. Lors de la phase d'indexation, ces autorités sont sélectionnées par les bibliothécaires et utilisées dans les notices via la ou les balise(s) DOC_DEE (Figures 2 et 3) (Kergosien, Bessagnet et Gaio, 2008 ; Bessagnet, Kergosien et Gaio, 2009).

Chaque balise DOC_DEE correspond à une vedette-matière composée d'une ou plusieurs autorités séparées par l'élément « -- ». Chaque vedette-matière correspond à un thème estimé par l'expert comme important (l'autorité décrivant le thème est utilisée en tête de vedette) dans le document (Stations climatiques, thermales, etc. et Eaux minérales dans la figure 2). Nous obtenons en résultat de ce premier traitement un ensemble de termes. L'extraction automatique de cet ensemble de termes et leur mise en correspondance grâce au thésaurus RAMEAU dans un graphe conceptuel nous permet d'obtenir une première représentation sémantique du fonds documentaire. Si nous reprenons les extraits de notices descriptives présentées dans les Figures 2 et 3, nous obtenons entre autres les termes : Eaux minérales, Stations climatiques, thermales, etc., Barèges (Hautes-Pyrénées), Bagnères-de-Bigorre (Hautes-Pyrénées).

Pour chaque terme extrait d'une notice, nous attachons, dans la structure définie, un lien vers le document auquel il a été assigné. Les autorités représentent alors le niveau conceptuel et, les documents, le niveau physique. Le document décrit par la notice descriptive présentée en Figure 3 sera relié aux autorités Œuvres scientifiques, Stations climatiques, thermales, etc., Bagnères-de-Bigorre (Hautes-Pyrénées) et 19e siècle.

Cette liste de termes est une première étape vers la définition d'un thésaurus représentant les résultats du travail d'analyse des bibliothécaires. Il reste maintenant à identifier l'ensemble des relations entre ces termes pour transformer la liste en thésaurus TERRIDOC.

En exploitant le thésaurus RAMEAU, nous enrichissons automatiquement le vocabulaire obtenu ci-dessus avec :

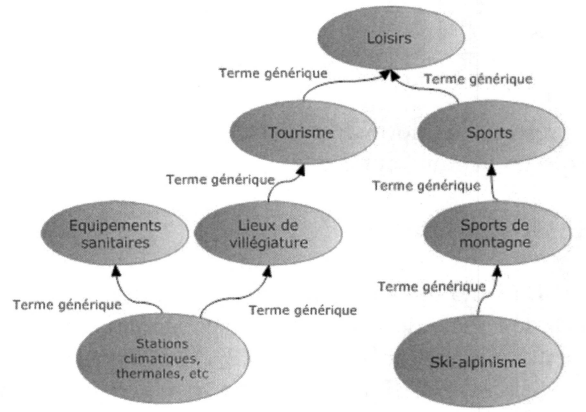
- les termes « génériques » et « employés pour » ;
- les relations qui leur sont associées ;
- les relations entre termes associés s'il en existe.

La règle est la suivante : nous recueillons dans un premier temps pour chaque terme du vocabulaire les termes « employés pour », « associés » et « génériques » et nous les relions entre eux respectivement par les relations « employé pour », « terme associé » et « terme générique ». Il faut noter que les relations hiérarchiques incluent la relation générique (genre-espèce), la relation partitive (tout-partie), la relation d'instance et les relations poly-hiérarchiques. Fischer (1998) souligne cette ambiguïté par le fait que la définition de ces relations « terme plus spécifique », « terme plus générique » est orientée par l'utilisation faite des thésaurus, c'est-à-dire l'aide au travail du documentaliste (indexation, recherche), et non par la formalisation de la connaissance du domaine.

Figure 4
Extrait de notice descriptive 3.

<DOC_DEE>Ski-alpinisme -- Barèges (Hautes-Pyrénées) -- 18e siècle**</DOC_DEE>**

Figure 5
Enrichissement du vocabulaire par les termes « génériques »



Le but visé par l'enrichissement du thésaurus, par ces termes génériques, est de permettre le regroupement en une seule structure des termes extraits. Prenons l'exemple d'un document image dont le sujet principal représenté par le bibliothécaire est indiqué à la Figure 4. Le traitement proposé ci-dessus permet d'obtenir une structure sémantique enrichie (Figure 5).

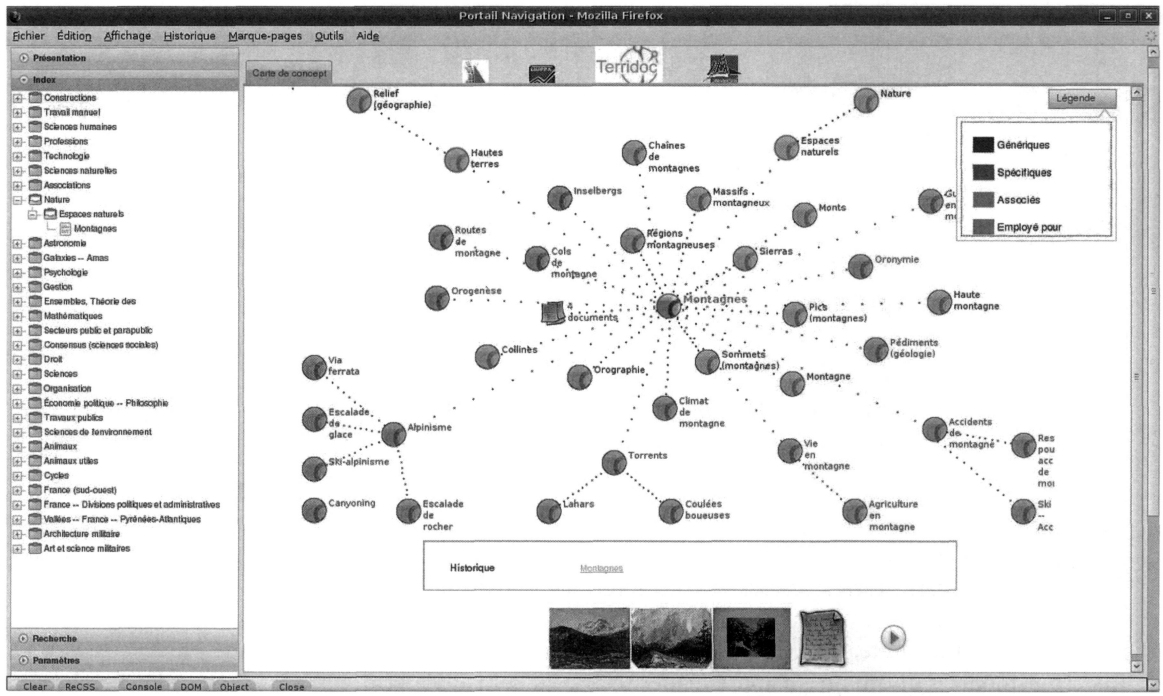
Le lien défini explicitement entre le terme « Ski-alpinisme « et les termes qui lui sont associés facilite la navigation en offrant la possibilité de lier les documents traitant de ce sujet aux documents traitant de « Stations climatiques ».

Le thésaurus propose une première représentation sémantique de structure triviale de notre fonds documentaire, définie automatiquement, représentant de façon globale la connaissance extraite des notices indexées par les bibliothécaires.

Nos premières expérimentations ont été menées sur un corpus de 750 notices descriptives et leurs documents associés. Nous obtenons un ensemble de 1 449 termes que nous enrichissons ensuite par les termes « employés pour », « associés » et « génériques » suggérés par RAMEAU. Le thésaurus que nous développons ainsi offre une première structure synthétique représentant le travail des bibliothécaires.

Sous sa forme la plus simple, cette structure ne peut être exploitée pour observer et analyser l'ensemble des saisies. Nous en proposons donc une représentation cartographique pour permettre aux experts d'appréhender de façon synthétique l'ensemble du travail d'in-

Figure 6
Interface de navigation et de visualisation du thésaurus TERRIDOC : recherche par Index



dexation d'un fonds documentaire donné, tel que réalisé par les différents bibliothécaires y ayant contribué.

Représentation cartographique du thésaurus TERRIDOC

Le premier usage que nous assignons à la représentation cartographique du fonds documentaire est de proposer aux experts une navigation dans le fonds, guidée par la structuration, définie automatiquement, de leurs propres connaissances expertes. Nous sommes alors dans une problématique de représentation au sens large de l'information synthétisée dans la structure sémantique et nous cherchons à faciliter les tâches suivantes (Hascoët et Beaudoin-Lafon, 2003) : exploration rapide de l'ensemble de la structure sémantique, mise en évidence de relations et de structures dans les informations et enfin mise en place des chemins d'accès aux informations pertinentes.

Nous proposons une interface composée d'une vue globale et d'une vue locale. La vue globale permet aux bibliothécaires d'appréhender les connaissances expertes structurées dans le thésaurus de façon intégrale, alors que la vue locale permet d'acquérir des informations sur un sous-ensemble du thésaurus. La combinaison de ces deux vues facilite la navigation et la relecture du travail d'indexation opéré sur un fonds documentaire. Nous proposons en parallèle, en nous basant sur les travaux de Zizi et Beaudoin-Lafon (1995), une vue locale sous forme de carte de concepts qui met en évidence un sous-ensemble de la structure sémantique et qui nous permet de représenter explicitement les liens existants entre les termes du thésaurus (niveau conceptuel) et les documents du fonds attachés à ces termes (niveau physique).

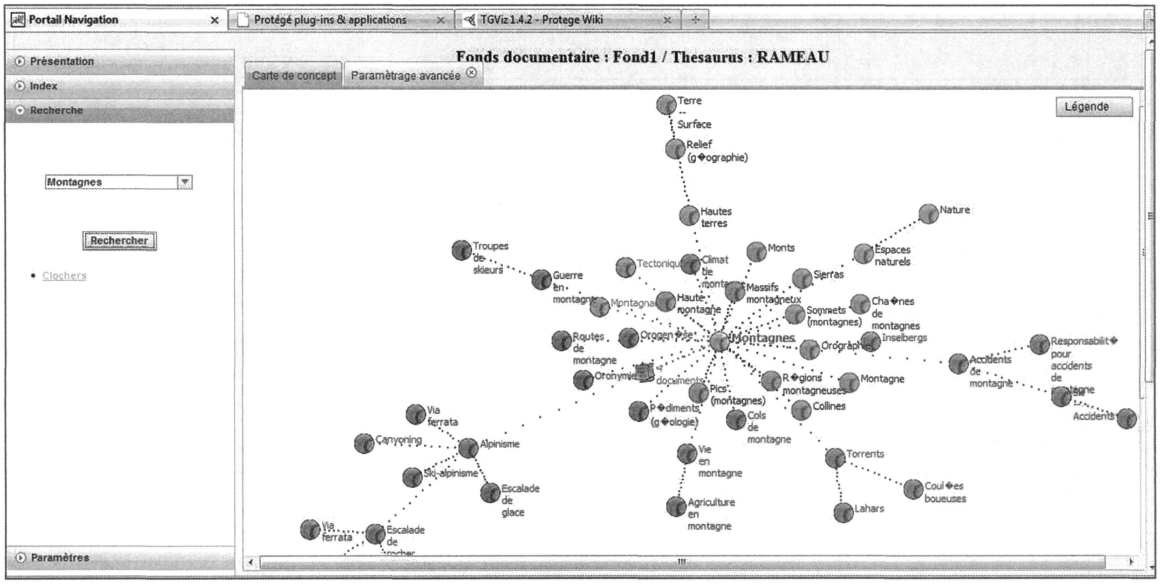
Nous avons mis en place une application Web nommée TERRIDOC permettant à l'utilisateur d'explorer la collection pour trouver les documents pertinents sans avoir à exprimer son besoin sous forme de requête conventionnelle. L'usager navigue dans la structuration via les concepts de haut niveau déterminés par notre traitement préalable en double cliquant sur le concept désiré (Figure 6).

Un premier accès peut être obtenu à partir de la liste de termes présentée sous forme hiérarchique à la gauche de l'écran. Lorsque l'utilisateur sélectionne un terme dans la liste, la carte de concepts est mise à jour, avec en son centre le terme sélectionné. L'utilisateur peut choisir le nombre de niveaux à afficher ainsi que les relations classiques d'équivalence, de hiérarchie et d'association que l'on trouve dans un thésaurus.

Un second accès est proposé via une saisie du terme recherché (Figure 7).

Une icône nommée « *Documents* » représente le nombre de documents physiques dans le fonds documentaire auxquels a été assigné un terme d'indexation particulier. Nous affichons également dans un onglet supplémentaire les informations de la notice d'autorité RAMEAU par un simple clic sur le terme.

Figure 7
Interface de navigation et de visualisation du thésaurus TERRIDOC : recherche par Mot-clé



À l'aide de cette maquette, les bibliothécaires de la MIDR se sont montrés en mesure d'appréhender de façon globale, puis détaillée, l'ensemble du travail d'indexation réalisé et d'identifier ainsi plus aisément les erreurs d'indexation.

Nous nous devons de noter quelques limites dues à la saisie opérée par les experts de la médiathèque. Certaines associations ne pourront hélas être explicitées automatiquement, certains documents ayant subi une phase de description, mais non la phase d'indexation. Ainsi, ces documents ne sont pas décrits par des vedettes provenant de RAMEAU et ne pourront être reliés à notre thésaurus. De plus, nous avons identifié des erreurs de saisies (orthographe, pluriels, termes exclus) ainsi que des termes ne semblant pas provenir du thésaurus. Un travail plus approfondi doit nous permettre d'exploiter au mieux ces connaissances cachées.

Structure et contraintes

Dans ces différentes étapes, nous nous sommes appuyés sur le thésaurus RAMEAU initialement stocké au format UNIMARC. Le format UNIMARC a l'avantage d'être universel en répondant à la norme MARC 21 définie principalement pour le monde des bibliothèques sur le plan international. Cependant, l'utilisation de RAMEAU nécessite de parcourir constamment des fichiers importants, posant à la fois des problèmes de temps, de traitement et d'accès aux ressources.

Aussi, nous avons défini une base de données, respectant la norme UNIMARC et permettant de gérer un ou plusieurs vocabulaires contrôlés qui puissent être utilisés dans des applications d'indexation de fonds documentaire importants, mais aussi et surtout au sein d'applications enrichies de recherche d'informations. Nous sommes en relation actuellement avec des chercheurs québécois afin d'obtenir une version du Thésaurus RASUQAM de l'Université de Québec à Montréal, dont l'exploitation nous permettrait de prouver la généricité de notre approche.

Plusieurs objectifs ont guidé ce traitement :

- respect du format UNIMARC : le thésaurus RAMEAU étant pour nous un prérequis important dans nos traitements, nous avons utilisé le document « format des notices d'autorité UNIMARC Deuxième édition revue et augmentée » définissant le format UNIMARC pour définir la structure de notre base de données. Les thésaurus qui seront stockés dans la base de données THESAURUS devront respecter cette structure, ce qui pourrait exiger des traitements de conversion ;
- intégration dans MYSQL et PostGreSQL : la structure de la base de données présentée en annexe a été créée avec MYSQL Server 5.1 et ultérieur ainsi qu'avec PostgreSQL 8.2.6 et ultérieur.

Propositions pour une assistance à la validation de l'indexation

Les premiers retours d'expérience des experts de la MIDR sont encourageants. Les bibliothécaires apprécient la possibilité de visualiser de façon plus globale les résultats de leur travail d'indexation. La structure définie sous forme de thésaurus représente en effet une base de connaissances résultant du travail d'analyse

Tableau 2
Résultat du traitement de validation de l'indexation de 750 documents.

Autorités RAMEAU	1 108
Reprises à faire (majuscule)	85
Reprises à faire (pluriel)	116
Termes non RAMEAU	341

d'un fonds documentaire donné, impossible à réaliser jusqu'à maintenant étant donné le nombre trop important de documents et de notices descriptives associées. De plus, les bibliothécaires étant régulièrement amenés à travailler sur des sites distants les uns des autres (sept sites dépendent de Pau), cette base de connaissances facilite les échanges lors des phases de description et d'indexation du fonds documentaire.

Un premier contrôle

Lors de la création automatique du thésaurus, une phase de vérification est lancée de façon automatique afin de s'assurer que les termes utilisés dans les notices descriptives pour indexer le document sont bien présents en tant qu'autorités-matière dans RAMEAU. Nous avons ainsi pu identifier quatre cas d'erreurs liées à l'utilisation de la ressource RAMEAU :

1. notices ne contenant pas de termes d'indexation : certaines notices ne contiennent pas de descriptifs du contenu utilisant la ressource RAMEAU. Il est donc impossible de lier les documents décrits par ces notices dans notre cartographie ;
2. dénomination des autorités : des erreurs de nommage des autorités sont apparues dû à une mauvaise utilisation de la casse et du singulier/pluriel : le langage d'indexation impose des règles de désignation de concepts en utilisant le pluriel pour désigner ce qui est dénombrable (par exemple Montagnes, Châteaux) et le singulier pour ce qui ne l'est pas (Paysage ou Mer). Un nombre important d'erreurs de ce type est identifié dans l'utilisation de la casse et du pluriel (Tableau 2) ;
3. gestion des termes exclus : en tant que langage contrôlé, RAMEAU limite l'usage du vocabulaire en proscrivant l'emploi des termes considérés comme équivalents (synonymie et quasi-synonymie). Parmi ces termes équivalents, une seule forme est retenue pour représenter un concept. Elle est le seul terme autorisé pour l'indexation. Les termes exclus peuvent être des synonymes ou des quasi-synonymes (mots de sens voisin, ou ayant entre eux des rapports d'inclusion) et sont reliés à l'autorité par une relation d'équivalence exprimée sous la forme « employé pour ». Nous avons identifié des termes exclus utilisés pour décrire les documents (par exemple Maisons qui devrait être remplacé par l'autorité Habitations) ;
4. gestion des homonymes. Une autorité doit représenter un seul concept. Cependant un terme peut avoir différents sens et la distinction des homonymes se fait en utilisant des qualificatifs (par exemple Grues (appareils) et Grues (animaux)). Actuellement, nous savons distinguer l'ensemble des propositions possibles provenant de RAMEAU lorsque le terme est utilisé dans la notice sans ces qualificatifs alors qu'ils sont nécessaires (le terme Grues ne peut être utilisé seul dans RAMEAU par exemple).

Vers une assistance à la correction

Une fois identifiés, les cas d'erreurs sont stockés pour une phase de correction ultérieure via l'interface Web.

À partir du prototype permettant de naviguer dans le fonds documentaire présenté aux Figures 6 et 7, il est possible d'accéder via un onglet à une liste de termes identifiés (d'après les différents cas d'erreurs énoncés précédemment) comme ne faisant pas partie de RAMEAU. Pour chaque terme, le nombre d'occurrence de l'erreur est affiché. En ce qui concerne les cas liés à des erreurs de nommage et à l'utilisation de termes exclus, des solutions sont proposées. Lorsque l'usager, par exemple, souhaite corriger le terme Grues qui figure sur la liste des termes, il clique sur le terme défini comme un lien afin d'accéder à une nouvelle page présentant la ou les solution(s) proposée(s). Dans le cas cité en exemple, trois solutions de remplacement sont proposées : Grues (appareils), Grues (animaux), et Ajout à la liste de termes candidats.

La correction d'un terme peut être faite de façon locale (pour une notice en particulier) dans le cas où l'utilisation du terme est spécifique au document traité, ou de façon globale lorsque l'erreur est le pluriel (par exemple à l'autorité Montagnes) et que l'on souhaite reporter cette correction sur l'ensemble des notices descriptives.

Il est possible d'indexer des documents en utilisant des termes ne faisant pas partie de RAMEAU. Prenons le cas d'une photographie de la Place Royale à Pau ; l'expert qui fait l'indexation pourra choisir d'utiliser le terme Place Royale (Pau), terme spécifique au fonds documentaire que le comité chargé de maintenir RAMEAU n'a pas incorporé au thésaurus. Ce terme apparaîtra dans la liste des erreurs puisqu'il ne se trouve pas dans RAMEAU. Lors de sa correction, l'usager peut choisir de l'ajouter à la liste de termes candidats. Cette liste apparaît comme une nouvelle source de connaissances sur laquelle pourront s'appuyer à l'avenir les bibliothécaires pour de nouvelles tâches d'indexation. La liste des termes candidats peut aussi être formatée et proposée au comité de décision de la BnF, en vue de l'enrichissement du thésaurus.

Pour les cas d'erreurs liés à l'homonymie, nous limitons l'assistance à la correction en proposant, comme pour d'autres cas, une liste de termes contenant le mot qui signale un cas d'erreur (par exemple Grues sans qualifiant supplémentaire). Il est envisageable de prévoir des solutions tentant d'identifier, en calculant la fréquence des qualifiants possibles (appareils et animaux pour l'exemple de Grues) la solution qui semble la plus probable dans la notice descriptive voire même, pour les documents texte, dans le contenu du document.

Tous les cas recensés fournissent une base d'expérience que nous pouvons exploiter pour, d'une part, aider à l'indexation en analysant à posteriori ce travail et, d'autre part, pour mettre en place des scénarios pédagogiques afin de faciliter le travail des bibliothécaires néophytes dans l'apprentissage de l'utilisation de RAMEAU pour indexer des documents.

Conclusion

Nous avons présenté nos travaux sur la structuration et l'adaptation d'un thésaurus comme outil d'aide à l'indexation. À l'aide d'un fonds documentaire représenté par 750 notices descriptives, nous avons pu valider nos premières expérimentations auprès des experts du domaine, que sont les bibliothécaires, par le biais de l'interface décrite ici. Ces retours d'expériences nous encouragent à prévoir des enrichissements de l'interface de validation du contenu des notices descriptives et, plus précisément, les éléments liés à l'indexation basée sur le thésaurus RAMEAU.

La structure définie sous forme de thésaurus représentant de façon globale ce travail d'indexation offre, dans un premier temps, la possibilité aux usagers experts de naviguer dans le fonds documentaire via leur propre représentation de ce fonds. La structuration, incluant une phase de vérification des termes utilisés, permet, dans un second temps, de proposer des outils d'aide à la correction des cas d'erreurs identifiés, facilitant ainsi le travail de validation des notices descriptives.

Nos travaux actuels portent sur une démarche qui vise à définir une ontologie à partir d'un thésaurus. En effet, nous souhaitons intégrer de nouvelles connaissances caractérisant le territoire dans notre structure (ajout de toponymes et de relations entre concepts). À terme, l'objectif est de proposer une interface permettant à toutes les catégories d'utilisateurs, désireux de découvrir un territoire décrit par des documents, de naviguer dans le fonds documentaire via une ontologie du domaine. ⊙

Sources consultées

AFNOR. 1981. *Documentation : Règles d'établissement des thésaurus monolingues* (NF Z47-100). Paris : AFNOR.

Barsalou, L.W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences* 22, (4) : 577–660.

Bessagnet, M.-N., E. Kergosien et M. Gaio. 2009. Extraction de termes, reconnaissance et labellisation de relations dans un thésaurus, *in CIDE 2009, 12e Colloque International sur le Document Electronique*, octobre 2009, 275-286.

Card, S.K., J.D. Mackinlay et B. Shneiderman. 1999. Information visualization. In *Readings in information visualization : using vision to think*. Burlington, Mass. : Morgan Kaufmann, 1-34.

Cubaud, P. et A. Topol. 2001. A VRML-based user interface for an online digitalized antiquarian collection. In *Proceedings of the sixth international conference on 3D Web technology*. Paderbon, Germany : ACM Press.

Elghoul, M. 1990. *Méthodologie de conception d'un SIAD pour la gestion documentaire, aide à l'indexation, aide à la construction du thésaurus, aide à la recherché et aide à l'apprentissage*. Thèse Université Paris X Nanterre.

Fischer, D. H. 1998. From Thesauri towards Ontologies. In *Structures and Relations in Knowledge Organization : Proceedings of the 5th International ISKO Conference*, W.M. Hadi, J. Maniez, S. Pollitt (Eds.),Würzburg : Ergon, 18-30.

Hascoet, M. et M. Beaudouin-Lafon. 2003. Visualisation interactive d'information. *Revue I3* (1) : 650–659.

Hernandez, N. 2005. *Ontologies de domaine pour la modélisation du contexte en recherche d'information*. Thèse de doctorat, Université Paul Sabatier de Toulouse.

Jacquemin, C., H. Folch, K. Garcia et S. Nugier. 2005. Visualisation interactive d'espaces documentaires. *Revue I3* 5,(1) : 59-84.

Kergosien E, M.-N., Bessagnet et M. Gaio. 2008. Semantic cartography : towards helping experts in their indexation task. In *Sixteenth International Conference on Knowledge Engineering : practice and patterns, Acitrezza, Catania, Italy, 29th September-3rd October, 2008*.

Kohonen, T. 2006. *Self-Organizing Maps*. Berlin : Springer.

Nieszkowska, E. 2003. Quelle indexation pour une bibliothèque spécialisée. Mémoire d'étude, sous la direction de Max Naudi, BNF.

Norman, D. A. 1993. *Things That Make Us Smart : Defending Human Attributes in the Age of the Machine*. Reading, Mass. : Addison Wesley.

Pepper, S., et G. Moore. 2001. *XML Topic Maps (XTM) 1.0 Specification, TopicMaps*. En ligne à l'URL <http://www.topicmaps.org/XTM/1.0.>.

Schatz, B. R., E.H. Johnson, P.A. Cochrane et H. Chen. 1996. Interactive term suggestion for users of digital libraries : Using subject thesauri and co-occurrence lists for information retrieval. In *Proceedings of the 1st ACM Digital Library Conference*, Bethesda, MD, 126–133.

Tricot, C., C. Roche, C.E. Foveau, et S. Reguigui. 2006. Cartographie sémantique de fonds numériques scientifiques et techniques. *Revue Document Numérique*, Numéro thématique Visualisation pour les bibliothèques numériques, 13-35.

Zizi, M., et M. Baudoin-Lafon. 1995. Hypermedia Exploration with Interactive Dynamic Maps, *International Journal of Human Computers Studies* (43) : 441-464.