Des documents vers la connaissance d'un territoire

Eric Kergosien

Des documents vers la connaissance d'un territoire

Eric Kergosien

UPPA
Laboratoire LIUPPA
64000 PAU
http://liuppa.univ-pau.fr
eric.kergosien@univ-pau.fr

RÉSUMÉ. *Nos travaux s'inscrivent dans la description d'un territoire par le biais de documents patrimoniaux numérisés. Dans notre travail de recherche, nous exploitons les spécificités du corpus liées à des termes qui « ont fait sens » aux bibliothécaires lors de la constitution de la notice descriptive. Le réseau de termes qui constitue une version minimaliste d'une ontologie a été construit à partir des notices descriptives faites par les bibliothécaires. Notre approche comprend trois étapes : la première consiste à définir automatiquement un thésaurus en utilisant deux types de ressources ; la deuxième étape consiste à enrichir ce thésaurus par des concepts par la mise en relation d'entités toponymiques via des ressources externes et enfin la dernière phase de l'approche consiste à associer ces entités aux concepts provenant du thésaurus généré dans la première étape par des relations de sens contenues dans les notices descriptives afin d'obtenir une structure couvrant de manière intéressante un domaine cible.*

ABSTRACT. *Our work concerns the description of a territory through* digitized *heritage documents. In our research work, we exploit specificities of the corpus linked to terms which "made sense" to the librarians when defining the descriptive notice. The network of terms which constitutes a minimalist version of an ontology was built starting from the librarians' descriptive notices. Our approach involves three steps: the first one defines automatically a thesaurus using two types of resources; the second one consists in enriching this thesaurus by concepts about toponymic entities and finally the last one associates these entities with the concepts issued from the thesaurus by semantic relationships contained in the descriptive notices in order to obtain a global structure of our target area.*

MOTS-CLÉS : *Ingénierie des connaissances, Ontologie, Territoire, Indexation, Navigation.*

KEYWORDS: *Knowledge Engineering, Ontology, Territory, Indexing, Navigation.*

1 Introduction

Un grand nombre de centres documentaires tels que les médiathèques et les bibliothèques universitaires possèdent un fonds documentaire à la fois riche et hétérogène, relatant d'un territoire portant sur une période plus ou moins importante. Leurs besoins sont importants tant dans la gestion du fonds documentaire (organisation du fonds documentaire) que dans l'extraction de l'information,

l'indexation automatisée (ou semi-automatisée) ou encore dans la recherche et la navigation qui doit pour être attractive intégrer les technologies proposées par le Web. Afin de revitaliser la valeur des documents constituant ce type de fonds, il nous a paru important d'identifier et de formaliser leur potentiel informatif. L'objectif est ici d'en extraire une représentation sémantique mettant en avant un territoire en nous appuyant sur le travail d'indexation des bibliothécaires. Ce travail regroupe les connaissances expertes provenant d'un vocabulaire contrôlé externe – RAMEAU[1] en l'occurrence. Parmi les différentes structures existantes, permettant de modéliser la connaissance, que sont les vocabulaires contrôlés, la taxonomie, le thésaurus ou encore l'ontologie (du plus simple au plus spécifique), notre choix s'est porté sur l'ontologie. En effet, comme le vocabulaire contrôlé, l'ontologie permet de définir un ensemble de termes caractérisant le contenu du fonds documentaire ; comme la taxonomie, elle permet de représenter les relations de type spécifique et générique existantes entre ces termes ; comme le thésaurus, l'ontologie permet de représenter les relations entre termes connexes. L'ontologie donne la possibilité supplémentaire de décrire explicitement ces relations entre concepts, ce qui permet d'enrichir de façon importante la structure sémantique caractérisant le fonds documentaire. Nous faisons l'hypothèse de pouvoir déterminer grâce à une ontologie du domaine un ensemble de ressources terminologiques propre au territoire implicitement mis en avant par un fonds documentaire. Une étude bibliographique réalisée dans les domaines de la géographie et de la sociologie nous permet de proposer §3.1 une première définition de ce que nous entendons par « territoire ».

Dans une première partie (§2), nous présenterons les travaux relatifs à la création d'ontologie relatifs à nos recherches. Nous développons ensuite les problématiques liées à notre approche et la démarche proposée (§3) permettant de définir une ontologie du domaine offrant une représentation du territoire décrit dans le fonds documentaire. Nous verrons (§4) les perspectives résultantes de nos travaux actuels.

2 Travaux liés

Dans un but d'offrir aux experts bibliothécaires un outil de recherche et de navigation permettant d'appréhender un territoire décrit par le fonds documentaire mis à notre disposition, nous proposons de créer quasi-automatiquement une ontologie que nous nommons ontologie PIV[2]. Depuis plusieurs années, les ontologies sont créées et utilisées dans le domaine de l'ingénierie des Connaissances. Le champ d'application est très large (Gruber, 1993) : d'une manière générale dans l'indexation et la recherche d'information (Genest, 2000) ou plus particulièrement

---

[1] Répertoire d'autorité-matière encyclopédique et alphabétique unifié; http://rameau.bnf.fr/. Thésaurus défini au sein de la Bibliothèque Nationale de France (BNF). Il est notamment utilisé à la BnF, par le Sudoc (catalogue collectif des bibliothèques universitaires) et par un nombre important de bibliothèques (municipales, départementales ou encore spécialisées).

[2] Ces travaux entre dans le cadre du projet Pyrénées Itinéraires Virtuels ; http://liuppa.univ-pau.fr/spip/article.php3?id_article=114

dans le domaine touristique (Velardi et al, 2001) par exemple. La conception automatique d'ontologies commence à émerger comme un sous-domaine de l'ingénierie des connaissances. Afin de créer ces ontologies, il existe diverses approches et méthodes. Certains travaux reposent sur l'analyse de textes (Bourigault et al, 2004) afin d'aider à la construction semi automatique des ontologies. (Bourigault et al, 2004) décrit les quatre étapes de la méthodologie de construction d'une ontologie à partir de textes (constitution du corpus à partir d'une analyse des besoins, étude linguistique afin d'identifier les termes et relations constituant la structure sémantique, normalisation sémantique définissant dans un langage formel les concepts et relations identifiées, validation par des spécialistes du domaine étudié). On peut remarquer que pour bâtir une ontologie à partir de textes, on utilise soit des ressources linguistiques externes, soit le corpus constitué des documents. Les outils supportant ces méthodes utilisent des techniques linguistiques pour retrouver les formes terminologiques dans l'analyse des textes. (Maedche & Staab, 2001) décrivent différents types d'approches distinguées en fonction du support sur lequel elles se basent : les plus courantes sont comme ci-dessus à partir de textes, de dictionnaires, d'autres à partir de bases de connaissance, ou encore de schémas semi-structurés et de schémas relationnels.

Pour notre part, nous nous intéressons plus particulièrement aux méthodes permettant de créer une ontologie du domaine à partir d'un thésaurus (Hernandez, 2005 ; Chrisment et al, 2006). Ces travaux simplifient l'opération de création d'ontologie à travers une approche permettant d'enrichir un thésaurus pour créer une ontologie à partir de sources de connaissances du domaine (vocabulaires, thésaurus, etc). Ces sources formalisées, contenant des termes représentant le domaine et (pour les thésaurus) des relations entre ces termes, apportent alors un plus sémantique indéniable à la représentation du domaine étudié. Le thésaurus sur lequel nous nous appuyons pour créer une première ontologie minimaliste est défini implicitement dans le travail d'indexation réalisé par les bibliothécaires et nous verrons que la première difficulté consiste à extraire et modéliser cette connaissance.

3 Notre approche pour construire une ontologie du territoire

Nous allons aborder dans cette partie les problèmes de recherche soulevés pour la mise en œuvre d'une ontologie du territoire et la démarche suivie pour y répondre.

3.1 Problèmes liés à la construction automatique d'ontologie

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. Les possibilités pour classer et structurer un ensemble de documents sont nombreuses. Le travail d'indexation réalisé par les bibliothécaires nécessite des connaissances du langage d'indexation RAMEAU et l'utilisation de la ressource thésaurus associée. Le thésaurus est un outil qui fait partie de la famille des vocabulaires contrôlés permettant l'accès par sujet aux catalogues et aux bases de données bibliographiques (Nieszkowska, 2003). Les vocabulaires contrôlés sont

notamment caractérisés par le fait qu'un terme le composant n'a qu'un seul sens précis et cet élément sera le seul à avoir ce sens, impliquant le contrôle de la synonymie, de l'homonymie et de la polysémie. Nous faisons références ici aux connaissances *expertes* plutôt qu'aux 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. Ainsi, un premier problème est identifié : quelle méthodologie et quels outils doivent être mis en œuvre pour définir automatiquement une structure sémantique représentant de façon globale le travail de l'ensemble des bibliothécaires ?

D'autre part ces connaissances expertes seules ne permettent pas de donner une représentation du territoire. En effet, les bibliothécaires n'ont pas pour objectif d'indexer un fonds documentaire en mettant directement en avant un territoire. Les fonds documentaires mis à disposition par un grand nombre de centres culturels sont constitués d'une quantité importante de documents territorialisés. Nous entendons ici par document territorialisé, tout document qui décrit/raconte un territoire : récits de voyage, contes, cartes postales. En nous appuyant notamment sur les travaux (Bonnemaison et al, 1996 ; Soergel et al, 2004 ; Elissalde, 2002), nous définissons le territoire comme : (i) *un ensemble de lieux formant un réseau délimitant l'espace* : nous cherchons ici à identifier un ensemble de toponymes présents dans le fonds documentaire afin de définir l'espace décrit par l'ensemble du fonds; (ii) *un ensemble d'éléments temporels* nous permettant d'identifier les dates et périodes clés pour ensuite délimiter la période globale mise en avant par le fonds documentaire ; (iii) *un ensemble d'éléments thématiques.* Bien que la composante thématique regroupe différents aspects tels que les coutumes, les activités pratiquées, etc., nous nous restreignons aux informations permettant d'identifier des personnages qui ont un rôle plus ou moins important dans le temps et dans l'espace identifiés dans les deux premières composantes. Un deuxième problème se pose alors à nous : comment identifier les éléments relatifs à la composante Lieux qui nous permettront ainsi d'enrichir notre structure sémantique en utilisant comme point d'entrée le territoire?

Afin de tenter de répondre à ces deux problèmes, nous avons dégagé trois phases importantes : (i) l'extraction de connaissances, (ii) la structuration du domaine et l'enrichissement de cette structure par des informations sur le territoire implicitement décrit par le fonds documentaire, (iii) la représentation visuelle de ce fonds à des fins d'acquisition de connaissances. Dans cet article, nous nous intéressons aux aspects extraction et gestion des connaissances.

3.2 Principe général de la démarche

Afin de répondre aux deux problèmes décrits (§3.1) liés à l'extraction et la gestion de connaissances, nous avons dégagé une démarche reposant sur deux étapes principales : (i) **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. Les notices descriptives sont les annotations sémantiques des documents ; (ii) **Transformation**

du thésaurus en ontologie : identification des éléments relatifs à la composante Lieux. Nous emploierons le terme lieux qui correspond dans les SIG[3] aux toponymes. Les documents territorialisés se caractérisent notamment par une omniprésence des noms de lieux relatifs à un territoire particulier. L'originalité de notre travail repose sur le fait que l'ontologie sera bâtie automatiquement à partir de trois sources de connaissances formalisées: celle issue du vocabulaire contrôlé, celle issue des notices descriptives rédigées manuellement et celles issues de gazetteers[4]. En effet, l'une des difficultés pour bâtir une ontologie est la collecte des documents sur lesquels reposera cette construction afin de capturer la connaissance. Présentons notre méthodologie.

3.3 Démarche

Nous proposons d'extraire un premier vocabulaire de termes provenant des notices descriptives, l'enrichir ensuite sous forme de thésaurus par les relations pouvant exister entre les termes extraits, et créer enfin une ontologie en enrichissant la structure par des concepts par la mise en relation d'entités toponymiques via des ressources externes offrant une première représentation du territoire implicitement décrit par le fonds documentaire.

3.3.1 Conception d'un thésaurus PIV

La première étape consiste à identifier et extraire automatiquement tous les termes utilisés dans les notices descriptives XML pour décrire le contenu du document. Lors de l'indexation, ces termes sont sélectionnés par les bibliothécaires dans RAMEAU et utilisées dans les notices via les balises DEE (figure 1).

< **DEE**>Stations climatiques, thermales, etc. -- Barèges (Hautes-Pyrénées) -- 18e **</DEE>**
< **DEE**>Eaux minérales -- Pyrénées (France) -- 18e siècle**</ DEE**>
<TITRE>Observation sur les ***eaux minérales de Bigorre et du Béarn*****</TITRE>**

<DEE> Stations climatiques, thermales, etc. -- Bagnères-de-Bigorre (Hautes-Pyrénées) -- 19e </**DEE**>
<TITRE>Recherches scientifiques sur les ***Eaux minérales de Bagnères de Bigorre*****</TITRE>**

Figure 1. *Extrait de notices descriptives 1 et 2*

Chaque balise DEE correspond à une liste de termes séparés par l'élément « -- ». Nous obtenons un ensemble de termes offrant une représentation sémantique du fonds documentaire : dans notre exemple, Stations climatiques, thermales, etc, Bagnères-de-Bigorre (Hautes-Pyrénées), Barèges (Hautes-Pyrénées), Eaux minérales. Pour chaque terme extrait d'une notice, nous attachons dans la structure un lien vers le document correspondant. Les termes représentent alors le niveau conceptuel et les documents le niveau physique.

---

[3] Système d'Information Géographique
[4] Répertoire toponymique fournissant les coordonnées géographiques correspondant au nom d'un lieu.

En exploitant le thésaurus RAMEAU, nous enrichissons automatiquement l'ensemble de termes en définissant pour chacun d'eux les termes « employés pour », « associés » et « génériques » et nous les relions entre eux respectivement par les relations « employé pour », « terme associé » et « terme générique » (figure 2).

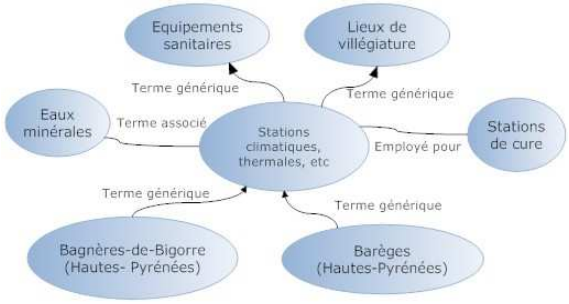

Figure 2. *Extrait du thésaurus PIV généré*

Nous considérons chaque terme du vocabulaire comme concept de bas niveau car rattaché directement à des documents et nous enrichissons le thésaurus avec les concepts plus génériques du thésaurus RAMEAU en ajoutant les relations de type « générique » liées. L'ambiguïté concernant les relations hiérarchiques est levée ici car les relations sont définies préalablement comme des relations génériques dans le thésaurus RAMEAU. Le but visé par l'enrichissement du thésaurus par ces termes génériques est de permettre le regroupement en une seule structure des termes extraits. Le thésaurus PIV propose une représentation sémantique de structure triviale de notre fonds documentaire, définie automatiquement, représentant de façon globale la connaissance extraite des notices indexées par les bibliothécaires. Nous cherchons maintenant à enrichir cette première structure sémantique par des informations renseignant sur le territoire implicitement décrit par les notices descriptives attachées aux documents.

3.3.2 Vers la connaissance d'un territoire

En exploitant les notices descriptives, notre méthode permet ensuite de capturer les entités spatiales (Lesbegueries, 2006) ainsi que tous les termes précédant ces entités spatiales à partir d'un traitement linguistique. Une entité spatiale est ici définie comme un terme ou un groupe de termes ancrés sur un toponyme définissant un lieu auquel nous associons une représentation géographique. En prenant comme exemple l'expression *eaux minérales de Bagnères de Bigorre*, *eaux minérales* est un terme provenant du thésaurus PIV et le terme *Bagnères de Bigorre* sera identifié dans notre chaîne de traitement puis validé automatiquement à travers des ressources toponymiques (gazetteers, SIG, etc.) comme une entité spatiale, devenant alors un concept dans notre ontologie enrichie. La dernière phase de l'approche consiste à associer les concepts identifiés au thésaurus PIV par des relations de sens contenues dans les notices descriptives. Ainsi en reprenant les extraits de notices présentées figures 1, sont aussi retenus comme entités spatiales candidates les concepts *Bigorre* et *Béarn* que nous validons ensuite en tant qu'entités spatiales via l'appel au

SIG. Un lien sémantique est alors créé entre les concepts *Eaux minérales*, *Bigorre* et *Béarn* » que l'on nomme « localisation » (figure 4).

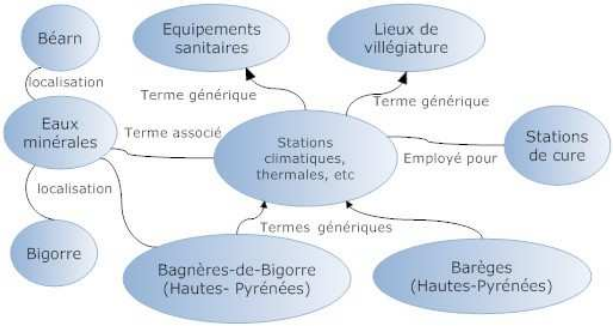

Figure 4. *Extrait de l'ontologie générée*

Nos travaux actuels cherchent à typer explicitement ce type de relation. Dans notre cas, nous faisons appel à un SIG qui va nous permettre d'identifier, par calculs géométriques, les relations spatiales entre concepts.

4 Conclusion

Nous avons présenté dans cet article nos travaux concernant la création d'un thésaurus en nous appuyant sur les connaissances expertes des bibliothécaires, et la transformation d'un thésaurus en une ontologie caractérisant un territoire. Parmi les différents formalismes existants et notamment OWL[5] permettant des calculs d'inférence, nous utilisons le formalisme XML Topics Map (Pepper & Moore, 2001) pour représenter cette ontologie par sa capacité à différencier aisément les niveaux conceptuels et physiques (les documents eux-mêmes). Nos premières expérimentations ont été menées sur un corpus de 750 notices descriptives et leurs documents associés. Nous obtenons un ensemble de 1449 termes constituant notre ontologie et 344 de ces termes sont des entités spatiales formant ainsi un premier réseau de lieux délimitant l'espace implicitement décrit par le fonds documentaire. Nous avons effectué une présentation d'un premier prototype (auprès de personnels de la MIDR[6]) permettant d'explorer la collection de documents via la carte de concepts pour trouver les documents pertinents.

Un prolongement à ces travaux est de proposer un outil d'aide à l'indexation en nous appuyant sur (van der Sluijs et al, 2008) qui propose notamment de réutiliser les annotations (tags) réalisés par les usagers afin de les soumettre, aux experts documentalistes pour validation Nous souhaitons à plus long terme intégrer de nouvelles connaissances caractérisant le territoire en exploitant les composantes Temps et Thématiques. L'objectif est à terme de proposer une interface permettant

[5] Web Ontology Language
[6] Médiathèque Intercommunale à Dimension Régionale de Pau

aux bibliothécaires de naviguer dans le fonds documentaire via une ontologie enrichie du domaine. L'ontologie ainsi créée offre notamment aux bibliothécaires la possibilité d'organiser la base documentaire selon un deuxième point de vue qu'est le territoire, le premier étant les notices descriptives résultantes de leur travail d'indexation.

5 Bibliographie

(Bonnemaison et al, 1996) J. Bonnemaison et L. Cambresy *G&C*, 20 : 8, 1996.

(Bourigault et al, 2004) D. Bourigualt, N. Aussenac-Gilles, J. Charlet. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. Revue d'Intelligence Artificielle (RIA). Numéro spécial sur les Techniques Informatiques et Structuration de Terminologies. PIERREL J.M. et SLODZIAN M. (Ed.). Paris : Hermès. 18 (1), pp 87–110. 2004.

(Chrisment et Al, 2006) C. Chrisment, F. Genova, N. Hernandez et J. Mothe. « D'un thesaurus vers une ontologie de domaine pour l'exploration d'un corpus ». *ametist*, Numéro 0AMETIST. http://ametist.inist.fr/document.php?id=152.

(Elissalde, 2002) B. Elissalde. « Une géographie des territoires », *L'information Géographique*, 65, 3, p. 193-205), 2002.

(Genest, 2000) D. Genest : Extension du formalisme des graphes conceptuels pour la recherche d'information. Thèse de doctorat, Université Montpellier II, 2000.

(Gruber, 1993) Thomas R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199-220, 1993.

(Hascoët et al, 2003) M. Hascoet et M. Beaudouin-Lafon. « Visualisation interactive d'information ». Revue I3, 1(1):77-108, 2003.

(Hernandez, 2005) N. Hernandez, Ontologies de domaine pour la modélisation du contexte en Recherche d'information, Thèse de doctorat UPS, décembre 2005

(Lesbegueries et al, 2006), J. Lesbegueries, C. Sallaberry, and M. Gaio, « Associating spatial patterns to text-units for summarizing geographic information ». 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval - GIR (Geographic Information Retrieval) Workshop, pp. 40-43, www.geo.unizh.ch/~rsp/gir06/accepted.html, ACM SIGIR 2006.

(Maedche & Staab, 2001) A. Maedche, S. Staab, Ontology Learning for the Semantic Web, *IEEE Intelligent Systems and Their Applications*, Vol. 16, No. 2. pp. 72-79, 2001.

(Nieszkowska, 2003) E. Nieszkowska, Quelle indexation pour une bibliothèque spécialisée ?, Mémoire d'étude, Sous la direction de Max Naudi, BNF, janvier 2003.

(Pepper & Moore, 2001) S. Pepper, G. Moore, «XML Topic Maps (XTM) 1.0 Specification», TopicMaps.Org, Aug. 2001. Available at http://www.topicmaps.org/XTM/1.0.

(Scheibling, 1994) J. Scheibling. *Qu'est-ce que la géographie?* Éd. Hachette, Paris, 1994, 199p.

(Tricot et al, 2006) C. Tricot, C. Roche, C.E. Foveau, S Reguigui. «Cartographie sémantique de fonds numériques scientifiques et techniques» Revue Document Numérique, Visualisation pour les bibliothèques numériques, Volume 9 2006/2, pp 13-35.

(Van der Sluijs et al, 2008) K. van der Sluijs, G.-J. Houben in: ERCIM News 72 - Special: The Future Web, published online at: http://ercim-news.ercim.org/content/view/336/536/

(Velardi et al, 2001) P. Velardi, P. Fabriani, M. Missikoff, Using text processing techniques to automatically enrich a domain ontology, In Proceedings of the ACM Conference on Formal Ontologies and Information Systems, pp 270-284, 2001.