GéOnto : Enrichissement d'une taxonomie de concepts topographiques

Sébastien Mustière, Nathalie Abadie, Nathalie Aussenac- Gilles, Marie-Noelle Bessagnet, Mouna Kamel, Eric Kergosien, Chantal Reynaud, Brigitte Safar

GéOnto : Enrichissement d'une taxonomie de concepts topographiques

Sébastien Mustière*, Nathalie Abadie*, Nathalie Aussenac-Gilles*, Marie-Noelle Bessagnet****, Mouna Kamel***, Eric Kergosien **** , Chantal Reynaud**, Brigitte Safar****

** IGN / Laboratoire COGIT*
2 av. Pasteur, 94160 Saint-Mandé
sebastien.mustiere@ign.fr, nathalie-f.abadie@ign.fr

*** LRI – Université Paris Sud 11, CNRS & INRIA Saclay Île-de-France*
Parc Orsay Université - 4 rue Jacques Monod- 91893 Orsay (France)
chantal.reynaud@lri.fr, safar@lri.fr

**** Université Paul Sabatier / IRIT / équipe IC3*
118 Route de Narbonne, 31062 TOULOUSE
kamel@irit.fr, aussenac@irit.fr

***** Université de Pau et des Pays de L'Adour / LIUPPA*
Avenue de l'Université, B.P. 1155, 64013 Pau Cedex
marie-noelle.bessagnet@univ-pau.fr, eric.kergosien@univ-pau.fr

RÉSUMÉ. *Dans cet article, nous présentons le projet GéOnto dont un des buts est de construire une ontologie de concepts topographiques. Cette ontologie est réalisée par enrichissement d'une première taxonomie de termes réalisée précédemment, et ce grâce à l'analyse de deux types de documents textuels : des spécifications techniques de bases de données et des récits de voyage. Cet enrichissement s'appuie sur des techniques automatiques de traitement du langage et d'alignement d'ontologies, ainsi que sur des connaissances externes comme des dictionnaires et des bases de toponymes.*

ABSTRACT. *In this paper we present the GéOnto project, aiming in particular to build an ontology of topographic concepts. This ontology is made by enrichment of a first taxonomy developed beforehand, through the analysis of two types of textual documents: technical database specifications and description of journeys. This work relies on natural language processing and ontology alignment techniques, as well as external knowledge resources such as dictionaries and gazetteers.*

MOTS-CLÉS : *ontologie, taxonomie, topographie, spécifications de bases de données, traitement automatique du langage, alignement d'ontologies, indexation spatiale, GéOnto.*

KEYWORDS: *ontology, taxonomy, topography, database specifications, natural language processing, ontology matching, spatial indexing, GéoOnto.*

1. Introduction

Avec d'un coté l'essor des techniques de l'information et, d'un autre coté, le développement des techniques de localisation spatiale, les données géographiques sont de plus en plus nombreuses et diverses. La gestion de cette diversité est un problème important qui se révèle en particulier à travers deux initiatives récentes et d'ampleur.

Au niveau national français, la direction générale de la modernisation de l'Etat (DGME) a lancé un projet de portail de l'information géographique publique qui a pour objectif de " constituer un point d'entrée le plus large possible pour rechercher les principales données géographiques de l'Etat, de ses établissements publics et des collectivités territoriales, en connaître leurs caractéristiques et les moyens d'y accéder et de les visualiser et les co-visualiser". Ce portail se donne pour but d'être " ouvert et interopérable, permettant ainsi la fédération des données "[1].

Au niveau européen, la commission chargée de l'Environnement a initié la directive INSPIRE adoptée en 2007 qui demande à mettre en place une infrastructure distribuée de données spatiales permettant " qu'il soit aisé de rechercher les données géographiques disponibles, d'évaluer leur adéquation au but poursuivi et de connaître les conditions applicables à leur utilisation [et] qu'il soit possible de combiner de manière cohérente des données géographiques tirées de différentes sources dans la Communauté et de les partager entre plusieurs utilisateurs et applications "[2].

Ces deux initiatives illustrent les besoins relatifs à la description et l'intégration cohérente de données géographiques, ce qui se révèle difficile en raison de la grande diversité de ces données, autant du point de vue de leur but que de leur niveau de détail.

Dans ce contexte, une approche de plus en plus privilégiée pour intégrer des données diverses, autant dans le monde des bases de données que dans celui de la recherche d'information et des systèmes d'information géographiques, est d'appuyer l'intégration des données sur une ontologie du domaine concerné (Gruber, 1993), (Guarino, 1998). Les ontologies jouent un rôle clé en intégration de sources d'information multiples et hétérogènes. Une ontologie est un modèle structuré des objets d'un domaine d'application, une vue sur ce domaine, une conceptualisation définissant des concepts, des propriétés, des relations. Son rôle est double. D'une part, elle précise le sens des concepts d'un domaine en étant le reflet d'un certain consensus au sein d'une communauté. D'autre part, elle fournit une sémantique formelle. Les concepts ne sont pas vus uniquement comme des notions sémantiques. Ils vérifient des propriétés qui ont une définition formelle. Le langage de

---

1 Charte du portail de l'information géographique publique, 21 juin 2006. www.geoportail.fr

2 Directive 2007/2/CE du Parlement européen et du Conseil du 14 mars 2007 établissant une infrastructure d'information géographique dans la Communauté européenne (INSPIRE)

.

représentation des connaissances utilisé doit permettre des traitements automatiques. Dans le contexte de l'intégration, ces représentations peuvent aider à comprendre et interpréter des descriptions hétérogènes de contenus relatifs à un même domaine pour ensuite pouvoir plus facilement les mettre en relation.

Divers travaux ont été réalisés dans le cadre particulier des ontologies dans le domaine géographique. Ils mettent en avant la nécessité de ces ontologies (Uitermark, 2001, Brodeur 2004), mais peu d'ontologies ont été réalisées en pratique (Lemmens, 2006) ou alors celles-ci décrivent des domaines ciblés, comme Towntology dans le domaine de l'aménagement et de l'urbanisme (Roussey et al., 2004) ou FoDoMuSt dans le domaine du traitement d'images (Brisson et al., 2007).

Dans cet article, nous présentons le projet GéOnto (ANR-07-MDCO-005) dont un des buts est de construire une ontologie de concepts topographiques, à partir de l'analyse de documents textuels francophones divers. Ce projet rapproche quatre équipes spécialistes de la construction et de l'alignement d'ontologies mais aussi de l'analyse des données et documents géographiques. La trame générale suivie par le projet pour constituer cette ontologie est décrite dans la partie 2. Les outils mis en œuvre et les premiers résultats obtenus sont ensuite détaillés dans la partie 3. Enfin, avant de conclure, quelques applications prévues de cette ontologie sont présentées dans la partie 4.

2. D'une taxonomie vers une ontologie : démarche globale

Nos recherches précédentes nous ont permis de définir une première taxonomie de termes décrivant des concepts spatialisés utilisés dans le domaine de la topographie (Abadie et al., 2008). Celle-ci prend la forme d'une hiérarchie d'environ 700 termes. Elle a été réalisée à partir des termes utilisés dans les spécifications de l'IGN, par analyse semi-automatique de ces documents puis réorganisée interactivement. Elle est en premier lieu francophone, mais à chaque terme est associée une traduction anglaise. Cette taxonomie est un premier pas vers la création d'une ontologie topographique plus riche. Pour être plus intensément exploitée, cette taxonomie mérite en effet d'être enrichie, à la fois de nouveaux concepts, mais aussi de définitions et propriétés formelles décrivant ces concepts ainsi que leurs relations. Réaliser cet enrichissement est un des objectifs du projet GéOnto. Cette partie présente la démarche globale suivie pour réaliser cet objectif (Figure 1).

L'enrichissement de la taxonomie précédemment construite (Figure 1, a) est réalisé à partir de deux sources de connaissances principales : d'une part des spécifications de bases de données topographiques de l'IGN (Figure 1, b), déjà utilisées pour créer la première taxonomie mais cette fois exploitées plus en profondeur et, d'autre part, des récits de voyage de la médiathèque de Pau (Figure 1, c), afin de prendre en compte un autre point de vue sur la topographie.

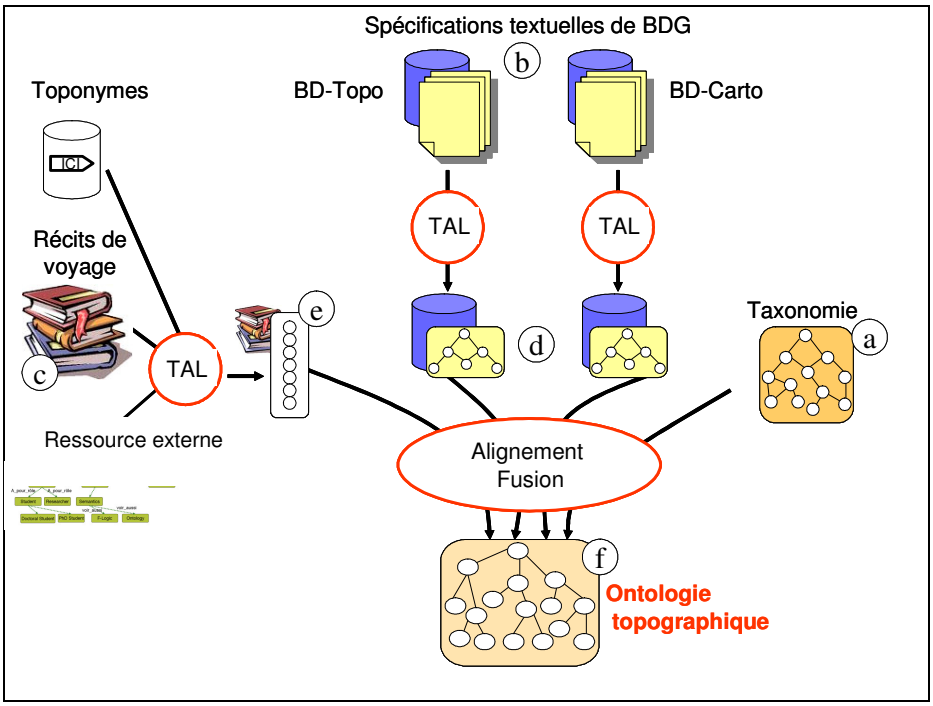

Figure 1. *Approche générale suivie pour la constitution de l'ontologie*

Les spécifications sont analysées par des techniques de traitement automatique du langage naturel pour être formalisées sous la forme d'ontologies, c'est-à-dire en explicitant le contenu des spécifications à travers des concepts, des relations entre ces concepts, des propriétés et des axiomes sur ces concepts (cf. partie 3.1). Ces ontologies reflètent le contenu des spécifications et de ce fait reflètent le point de vue et le vocabulaire utilisé dans chacune des bases de données qu'elles décrivent (en l'occurrence, la BDTOPO de l'IGN exploitée dans un premier temps). Ces spécifications ont un intérêt en elles-mêmes (Gesbert et al., 2004), (Abadie, 2009b). Elles seront exploitées pour une des applications visées dans le projet GéOnto, à savoir l'intégration de bases de données géographiques (cf. partie 4.2). Dans le contexte de constitution d'une ontologie topographique présenté ici, elles sont exploitées pour enrichir la taxonomie originelle de relations et concepts issus, en particulier, de l'analyse des définitions qu'elles contiennent.

Les récits de voyage ont une structure très différente et sont exploités d'une autre manière. On y recherche le vocabulaire utilisé pour décrire des concepts topographiques (rivière, ru, gave, ville, etc.), en s'appuyant sur l'idée que ceux-ci peuvent se trouver rattachés dans le texte à des toponymes (« le gave de Pau », « la ville de Pau », etc.). Une base de toponymes est donc utilisée pour repérer les

toponymes (en l'occurrence la BDNYME de l'IGN), et des techniques de traitement automatique du langage naturel sont ensuite appliquées pour identifier les termes rattachés à ces toponymes. Le traitement est affiné grâce à l'utilisation d'un vocabulaire contrôlé externe (thesaurus, taxonomie, etc.) pour mieux interpréter les termes identifiés (cf. partie 3.2). L'une des caractéristiques d'un vocabulaire contrôlé est que chaque terme le constituant a un seul sens et que ce sens est représenté par un seul terme.

Une fois ces traitements effectués, nous disposons donc d'une taxonomie qui a l'avantage d'être relativement structurée (Figure 1, a), des ontologies de spécifications qui ont l'avantage de représenter de manière formelle des propriétés et relations entre concepts (Figure 1, d), et d'une liste de termes issus des récits de voyage qui a l'avantage de contenir de nombreux termes et de refléter un vocabulaire plus grand public que celui des spécifications techniques (Figure 1, e). Afin de combiner ces trois avantages dans une seule ontologie topographique, il est nécessaire d'identifier les parties communes de ces ressources, ce qui est le rôle des techniques d'alignement d'ontologies (cf. partie 3.3), pour, ensuite, les fusionner en une ontologie topographique large (Figure 1, f).

Notons que cette approche permet, en plus de créer une ontologie topographique riche, de conserver les liens entre cette ontologie et les ontologies issues des spécifications textuelles des bases de données. Ceci permettra d'interroger les bases de données en bénéficiant de toute la richesse du vocabulaire de l'ontologie topographique. Plus concrètement, cette approche doit permettre de faire le lien entre le vocabulaire utilisé dans des récits de voyage quelconques (par exemple le terme "ville") et les éléments des bases de données qui y sont liés (par exemple la classe "commune" de la BD TOPO), rendant possible à terme une indexation spatiale fine des récits de voyage (cf. partie 4.1).

3. Outils mis en œuvre et premiers résultats

3.1 Traitement automatique des spécifications textuelles

Les méthodes de construction d'ontologies à partir de textes privilégient souvent l'analyse du texte proprement dit (Maedche, 2002) , (Buitelaar et al., 2005). Or la structure d'un document (titres, énumérations, définitions, etc.) donne forme et sens au contenu (Jacques, 2005). Nous proposons donc ici une approche qui s'appuie à la fois sur la structure matérielle du texte pour créer un premier noyau d'ontologie, et sur son contenu pour ensuite enrichir ce noyau. Cette approche est particulièrement adaptée aux documents de spécification de bases de données qui contiennent des descriptions d'objets, des relations existant entre eux, des contraintes et des définitions exprimées à la fois par la structure matérielle du document et par le langage naturel. Ces documents sont conformes à un schéma XML inspiré des normes ISO sur les données géographiques (avec des balises « definition »,

« class », « className », etc.), ce qui confère, par ailleurs, un certain degré de généricité à notre approche dans ce cadre d'application, et offre un repérage aisé de la structure.

3.1.1 Analyse de la structure du document

Les relations entre concepts des spécifications se traduisent en particulier par leurs positions relatives dans les documents. Nous considérons que lorsque chaque syntagme marqué par des balises réfère à un seul concept, alors la hiérarchie des balises traduit des relations sémantiques. Nous basons notre analyse sur la règle suivante :

Lorsque - A et B sont des balises sous la portée d'une même balise O *- C_1 et C_2 sont des concepts respectivement étiquetés par les unités textuelles marquées par A et B* *Alors une relation sémantique existe entre C_1 et C_2.*

Une étude systématique des balises du document XML et de leur imbrication a permis de déterminer comment identifier concepts, relations conceptuelles et propriétés. Cette analyse ne présente pas de difficulté majeure car les balises véhiculent elles-mêmes leur sémantique et les relations découlent de la connaissance du domaine. L'analyse du corpus à partir des différentes instanciations de cette règle fournit le noyau d'ontologie. Une description détaillée est donnée dans (Kamel et al.,2009).

3.1.2 Analyse du texte libre

Nous avons choisi d'utiliser des patrons lexico-syntaxiques pour repérer des relations sémantiques dans les parties du document en texte libre (Auger et al., 2008). Un patron lexico-syntaxique décrit une expression régulière, formée de mots, de catégories grammaticales ou sémantiques, et de symboles, visant à identifier des fragments de texte répondant à ce format. Le document de spécification de la base de données BDTOPO contient des champs définition riches en terme de relations. Pour analyser ce champ, une procédure d'annotation permet de caractériser les différents éléments de la définition par la séquence suivante : « *{Concept} ({M_PartieDe}) ? {Terme} {Propriete}* »,* où *Concept* est un concept de l'ontologie, éventuellement suivi d'un marqueur linguistique spécifique de la relation de méronymie (*M_PartieDe*), suivi d'un terme (*Terme*) obtenu à l'aide d'un extracteur de termes et de zéro ou plusieurs propriétés (*Propriété*) caractérisées par un adjectif ou un complément du nom. L'enrichissement du noyau de l'ontologie est réalisé conformément à l'algorithme suivant :

T : Terme, C: Concept, P : Propriété, ML : M_PartieDe

1. si seuls C et T sont présents, T devient un terme associé à C
2. si seuls C, T et P sont présents, T est considéré comme un concept plus générique que C, C est relié à T par la relation *est-un*, et les propriétés sont associées à C.
   a. si T existe déjà dans l'ontologie, il n'y a pas de nouvelles relations créées
   b. si T n'existe pas dans l'ontologie, nous recherchons un terme plus spécifique ST inclus dans T (au sens lexical) qui serait un concept
      i. si ST existe, T est relié à ST par la relation *est-un*
      ii. sinon T est créé comme un concept fils du concept Top
3. si C, T, ML et P sont présents, T est considéré comme un concept (reprendre l'algorithme à l'étape 2.) et C est relié à T par la relation *partie-de*

L'exemple ci-dessous illustre l'analyse du document de spécifications :

```
<pakage>    <packageName> Voies de communication routière </packageName>
<class>     <className> Tronçon de route </className>
              <definition> Portion de voie de communication
                                    destinée aux automobiles  </definition>
</class>  </package>
```

L'analyse de la structure du document établit une relation d'hyponymie entre les concepts *Tronçon de route* et *Voies de communication routière*. L'analyse du texte permet de créer le concept *Voie de communication* à partir du terme "voie de communication" comme un fils de *Top* et père de *Voie de communication routière* (car plus générique). Les concepts *Tronçon de route* et *Voie de communication* sont reliés par la relation *partie-de (*terme *portion de*), la propriété *destinée aux automobiles* est associée au concept *Tronçon de route*.

L'ontologie obtenue est plus riche que la taxonomie initiale essentiellement en termes de relations : les propriétés, des relations lexicales autres que *est-un*, d'autres relations sémantiques sont prises en compte. Par ailleurs, une fois les règles d'extraction écrites, le processus est totalement non supervisé.

3.2 Traitement automatique des récits de voyage

L'objectif ici est d'exploiter les termes rencontrés dans les récits de voyage et associés à des toponymes connus, pour déterminer le vocabulaire utilisé dans ces documents pour décrire la topographie.

Nous avons analysé 14 livres (récits de voyage dans les Pyrénées) à l'aide d'une chaîne de traitements qui permet l'annotation des informations spatiales détectées dans un document textuel (Lesbegueries et al. 2006). Cette annotation s'appuie sur des ressources géographiques diverses (bases de données de toponymie comme la BDNYME de l'IGN ou des « gazetteers »[3] contributives) pour l'identification des

[3] Dictionnaire ou index listant dans l'ordre alphabétique des noms de lieux ainsi que leur(s) représentation(s) géométrique(s).

entités nommées (EN) spatiales. La chaîne de traitements que nous proposons est présentée en Figure 2 (certaines étapes finales sont utiles à l'indexation spatiale des documents, évoquées en partie 4.1).

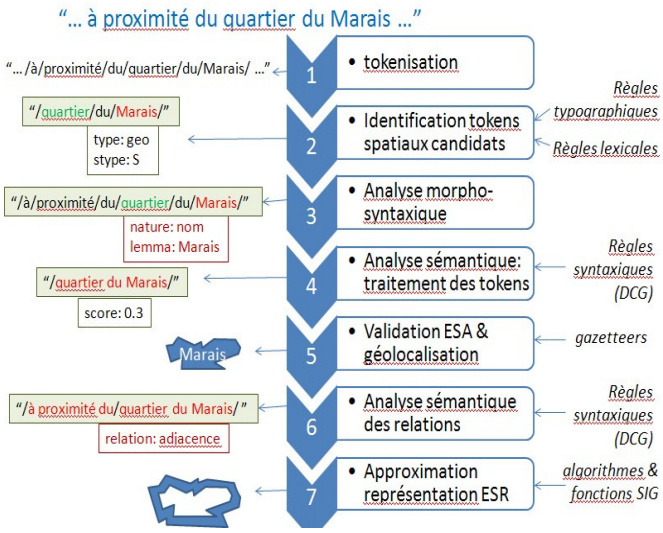

Figure 2. *Chaîne de traitements d'information spatiale dans des documents textuels*

L'étape (1) s'appuie sur une « tokenisation » classique. Nous adoptons ensuite une démarche de lecture « active », qui consiste à marquer rapidement des EN spatiales candidates puis à appliquer les étapes suivantes de l'analyse à ces EN uniquement. Un marqueur de *token* spatial candidat (2) utilise des règles lexicales (lexiques d'introducteurs d'information spatiale) et typographiques (majuscule en début de *token*). Puis, un analyseur morpho-syntaxique (3) associe un lemme et une nature à chaque *token* spatial candidat (i.e. « Marais », nom). Un analyseur sémantique (4) et (6) associé à des règles de grammaire DCG (Definite Clause Grammar) exprimées en Prolog qualifie des entités spatiales absolues (ESA) et des entités spatiales relatives (ESR). Une ESA est une entité nommée simple : « le quartier du Marais », « le pic d'Ossau », « la vallée d'Ossau », etc.Une ESR est une EN complexe définie à partir d'une autre EN : « au cœur du quartier du Marais », « au sud de la vallée d'Ossau », etc.

Nous obtenons ainsi un ensemble de termes associés à des EN. D'après notre analyse quantitative, nous obtenons des termes présents dans la taxonomie topographique initiale (130 termes distincts) ou non (1396 termes distincts). Parmi ces autres termes, certains ont un caractère topographique et pourraient enrichir la taxonomie (comme « gave » trouvé dans « le gave de Pau »), contrairement à d'autres (comme « maire » trouvé dans « le maire de Pau »). L'analyse des termes

associés aux EN permet de montrer l'intérêt de l'approche. 5% des termes identifiés (15% du nombre total d'occurrences dans les textes) se retrouvent dans les noms des classes ou les valeurs d'attribut des bases de données de l'IGN (concepts de niveau « nature » dans la Figure 3), alors que ce taux passe à 10% (33% du nombre total d'occurrences) si on utilise l'ontologie issue des spécifications liées à BDTOPO (concepts de niveau « sous-nature » dans la Figure 3), ce qui montre l'intérêt de cette ontologie.

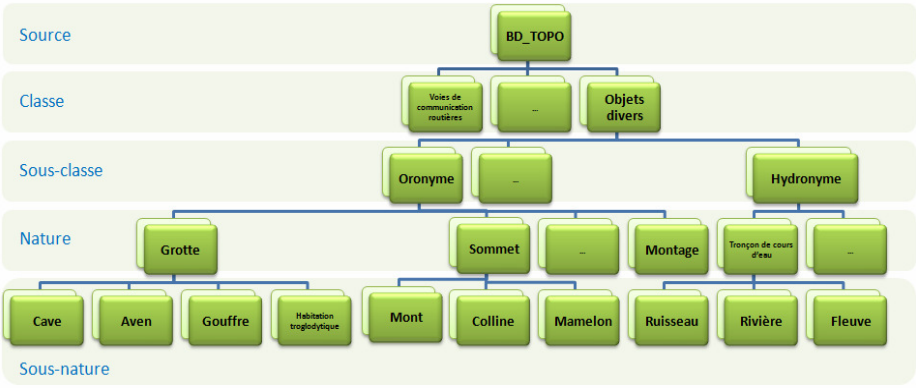

Figure 3. *Extrait du contenu des spécifications formalisées de la BDTOPO*

La liste des termes associées à des EN est, dans un second temps, complétée à l'aide du thésaurus RAMEAU[4] (Répertoire d'autorité-matière encyclopédique et alphabétique unifié). RAMEAU est utilisé, en France, par la Bibliothèque Nationale de France, les bibliothèques universitaires, de nombreuses bibliothèques de lecture publiques ou de recherche ainsi que plusieurs organismes privés. C'est une source de connaissance généraliste (des termes traitent des loisirs, des arts, etc.) et riche car elle comporte plus de 400.000 termes. Elle se compose d'un ensemble de termes reliés entre eux et d'une syntaxe indiquant les règles de construction des vedettes matière à l'indexation afin d'en assurer le bon usage.

La liste des termes associés à des EN ainsi complétée a été utilisée pour enrichir la taxonomie initiale. Nous proposons une méthodologie d'enrichissement dans laquelle nous vérifions s'il est possible d'identifier des proximités sémantiques entre ces termes et les concepts de la taxonomie à enrichir. Si on prend l'exemple du toponyme « Crabioules », identifié dans les récits de voyage par notre chaîne de traitements (Figure 2), ce dernier est qualifié, dans l'échantillon de textes analysés, par divers termes dont « col », « mont » et « abîme ». La taxonomie initiale permet de connaître la sémantique de concepts dénotés par certains termes, pour au final identifier une représentation spatiale dans les données géographiques pour le 'Col de Crabioules' et le 'Mont de Crabioules'. *Abîme* n'est pas représenté dans la

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

taxonomie mais a pour terme vedette *Grotte* dans RAMEAU (Figure 4). *Grotte* est présent à la fois dans la taxonomie et dans RAMEAU et, dans les deux cas, *Aven* et *Gouffre* sont représentés comme des éléments fils de *Grotte*. Cela nous permet de proposer la création d'un nouveau concept *Abîme*, fils du concept *Grotte* dans la taxonomie initiale. Parmi les 1396 termes distincts identifiés non présents dans la taxonomie, 1046 termes sont présents dans RAMEAU. Ils sont donc candidats à son enrichissement. Nos travaux actuels portent sur l'affinement du processus d'enrichissement afin de détecter les termes de RAMEAU porteur d'un sens géographique.


Grottes [+ subd. géogr.]

Vedette matière nom commun . S'emploie en tête de vedette

<Employé pour :
Abîmes
Antres
Avens
Cavernes *Ancienne vedette*
Cavernes préhistoriques
Cavités souterraines
Gouffres
Grottes ornées
Grottes préhistoriques
Préhistoire -- Grottes
Spélonques

<<Terme(s) générique(s) :
Habitat préhistorique
Relief (géographie)
Zones souterraines

Figure 4. *Exemple de notice descriptive RAMEAU décrivant le terme Grottes*

Nous présentons, dans la section suivante, les outils d'alignement d'ontologies utilisés pour identifier les parties communes des ressources construites en 3.1 et 3.2.

3.3 Alignement d'ontologies

L'enrichissement visé dans le projet GéOnto passe par l'identification de mises en correspondance ou mappings entre la taxonomie initiale et chacune des ressources servant à l'enrichir. Dans ce projet, l'alignement est réalisé à l'aide de *TaxoMap*, développé au LRI (Reynaud et al., 2007), (Hamdi et al., 2008). *TaxoMap* est particulièrement bien adapté à l'alignement de taxonomies comportant des descriptions très fines de domaines d'application, comme les taxonomies topographiques : (1) des concepts appartenant à un même domaine circonscrit, (2) des labels correspondant à des expressions composées de plusieurs mots, (3) des labels de concepts généraux inclus dans les labels de concepts plus spécifiques. *TaxoMap* a été utilisé pour aligner la taxonomie initiale et l'ontologie construite à

partir des spécifications de la BDTOPO obtenue comme décrit en partie 3.1, dans le but d'enrichir la première avec la seconde. Les deux premières sous-sections sont consacrées à la présentation de *TaxoMap*. Nous énonçons ensuite quelques pistes pour enrichir la taxonomie initiale à partir de résultats d'alignement.

3.3.1. Description de TaxoMap

TaxoMap a été conçu pour découvrir des alignements entre des taxonomies où les concepts sont seulement définis par leurs labels et les relations de subsomption qu'ils entretiennent avec les autres concepts. Le processus d'alignement est un processus orienté qui cherche à relier chaque concept d'une taxonomie source à un unique concept de la taxonomie cible. Il génère des relations de mise en correspondance, appelées mappings, qui sont des relations d'équivalence (*isEq*), de subsomption (*isA*) ou de proximité (*isClose*), auxquelles sont associées des mesures de similarité. La mesure de similarité est appliquée aux labels des concepts vus comme des ensembles de tri-grammes (Lin, 1998) et s'appuie sur l'utilisation d'un analyseur morpho-syntaxique *TreeTagger* (Schmid, 1994). Etant donné un concept C_S de l'ontologie source O_S, la mesure de similarité permet d'identifier l'ensemble des concepts de l'ontologie cible O_T, candidats au mappings avec C_S. Les techniques d'alignement mises en œuvre dans *TaxoMap* permettent de sélectionner le concept le plus pertinent, parmi l'ensemble de ces candidats. La découverte de mappings repose sur des techniques variées, terminologiques ou structurelles, appliquées séquentiellement de façon à rendre le processus de génération de mappings le plus efficace possible. Une proposition de mapping résulte de l'application d'une technique donnée et d'une seule. Chaque concept de O_S ne peut être aligné qu'avec au plus un concept de O_T. En revanche les concepts de O_T peuvent intervenir dans plusieurs propositions d'alignement.

3.3.2. Techniques d'alignement mises en œuvre dans TaxoMap

Soient C_S le concept de la source O_S pour lequel on recherche une correspondance et C_{Tmax}, C_{Tmax2} et C_{Tmax3} les trois concepts de la cible O_T qui ont les meilleures similarités avec C_S. Nous présentons ci-dessous les techniques d'alignement mises en œuvre dans *TaxoMap*.

- Technique de recherche de relations d'équivalence. Un mapping d'équivalence (C_S *isEq* C_{Tmax}) est proposé lorsque la similarité d'un des labels de C_S avec un des labels de C_{Tmax} est supérieure ou égale à un certain seuil.

- Techniques de recherche d'inclusion entre les mots des labels des concepts de O_S et ceux du label du concept ayant la plus forte similarité (C_{Tmax}). Trois techniques (T_2,T_3, T_4) sont appliquées séquentiellement. Selon T_2, un mapping (C_S *isA* C_{Tmax}) est généré si tous les mots d'un label de C_{Tmax} sont inclus dans ceux d'un des labels de C_S sans apparaître derrière un déterminant (Figure 5). T_3 et T_4 génèrent des mappings du type (C_S *isClose* C_{Tmax}) lorsque l'inclusion est inversée (T_3) ou lorsque le(s) mot(s) inclus se trouve(nt) derrière un déterminant quel que soit le sens de la relation d'inclusion (T_4).

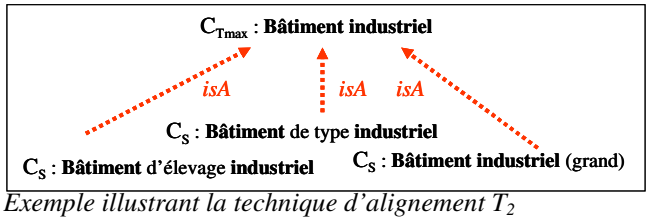

Figure 5. *Exemple illustrant la technique d'alignement T_2*

- Techniques basées sur la similarité relative. Ces techniques s'appliquent quand il n'existe pas d'inclusion de labels entre C_S et le concept ayant la plus forte similarité (C_{Tmax}) et lorsque la mesure de similarité de C_{Tmax} est significativement plus élevée que celle de C_{Tmax2}. Les techniques (T_5, T_6, T_7) de cette catégorie génèrent des mappings *isA* ou *isClose* selon les cas.

- Techniques basées sur la structure (T_8, T_9): T_8 est exécutée sur C_S si C_{Tmax}, C_{T2} et C_{T3} ont un père commun dans O_T partagé par au moins deux concepts. Dans ce cas, le mapping (C_S *isA PèreCommun*) est généré. T_9 s'appuie sur les mappings d'équivalence précédemment trouvés entre *Cs* et C_{Tmax}, et génère des mappings *isA* entre tous les spécialisants de C_S et C_{Tmax} (Figure 6).

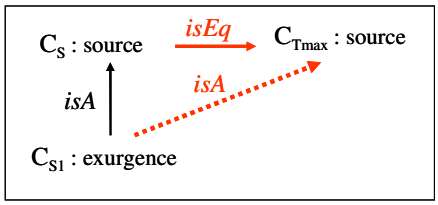

Figure 6. *Exemple illustrant la technique d'alignement T_9*

3.3.3. Enrichissement de TopoCarto-Cogit à partir de résultats d'alignement

L'ontologie construite à partir des spécifications textuelles associées à la base de données BD TOPO a été alignée en tant qu'ontologie source (O_S) à l'ontologie cible qu'est la taxonomie initiale (O_T). Les résultats d'alignement obtenus montrent que leur exploitation peut permettre d'enrichir O_T. Les mappings du type *isEq* ou *isA* entre un concept *Cs* de O_S et un concept C_T de O_T peuvent donner naissance à des relations du même type, représentées directement dans O_T. Si les concepts *Cs* sont munis de propriétés, celles-ci peuvent venir enrichir O_T.

Les mappings peuvent aussi donner lieu à des affinements. Ainsi, des mappings *isA* peuvent donner naissance à des relations de subsomption (*bâtiment d'élevage industriel subClassOf bâtiment industriel*) s'ils sont repris à l'identique mais également à des relations d'équivalence (*bâtiment de type industriel EquivalentClass bâtiment industriel*) ou donner lieu à la représentation de propriétés

(*Taille*, propriété de *bâtiment industriel,* pour pouvoir représenter le concept *bâtiment industriel (grand)*).

Les mappings *isClose* issus de *TaxoMap* traduisent une proximité entre concepts sans que l'on soit capable de clairement identifier leur sémantique. Deux interprétations sont possibles selon la façon dont ces mappings ont été générés. Ils correspondent soit à des concepts presque équivalents (voire issus d'erreurs comme *pépinnière* / *pépinière*) soit à des concepts C_S plus généraux que les concepts C_T (*monument* / *monument commémoratif*). Dans ce dernier cas, l'enrichissement consiste, non seulement à introduire une relation de subsomption entre les concepts C_T et C_S (*monument commémoratif subClassOf monument)*, mais également à positionner le nouveau concept dans O_T en identifiant son ou ses concept(s) père(s).

Ces premiers travaux montrent que l'alignement est un préalable à l'enrichissement, qu'il produit des résultats riches mais devant faire l'objet de traitements complémentaires spécifiques, en interaction avec l'expert. Une perspective de notre travail consiste à offrir un environnement d'aide à la spécification de tels traitements.

4. Utilisations prévues de l'ontologie réalisée

4.1 Indexation spatiale de documents

Au delà du besoin croissant de partage d'informations sur le Web qui passe par la structuration des ressources mises à disposition, le problème traité ici correspond également aux nouveaux besoins de valorisation des fonds documentaires patrimoniaux suscités par l'importante politique de numérisation mise en oeuvre par les différentes instances de conservation des collections documentaires territorialisées (archives régionales, musées, médiathèques...). Une part non négligeable de l'information contenue dans ces documents numériques fait référence de manière plus ou moins explicite à des entités géographiques. Or la plupart des systèmes permettant la gestion et la consultation de documents en ligne propose une indexation reposant sur l'exploitation de métadonnées produites manuellement, combinées à des méthodes de fouille plein texte basées essentiellement sur des méthodes statistiques.

Dans ce contexte, une des applications visées de l'ontologie topographique construite dans GéOnto est l'indexation spatiale fine des récits de voyage, c'est-à-dire l'affectation de géométries aux lieux mentionnés dans les documents, permettant de les localiser et de les interroger par des requêtes spatiales. La chaîne de traitements spatiale utilisée pour cette application (Figure 2, étapes 5 à 7) est détaillée dans (Gaio et al., 2008) et (Loustau et al., 2008). Plus précisément, nous espérons que l'ontologie produite dans le projet GéOnto va enrichir cette tâche d'indexation qui a déjà montré son utilité (Sallaberry et al, 2007), en affinant l'analyse sémantique des termes rencontrés dans les textes. Par exemple, l'entité

nommée « Artouste » aura une sémantique et une représentation spatiale différente selon la nature de l'élément géographique désigné. Qu'il s'agisse du lac, du pic ou de la vallée, les ressources invoquées et les stratégies de parcours de ces ressources seront différentes. La règle appliquée jusque là correspond à sélectionner la première représentation spatiale identifiée dans la ressource type *gazetteers*, pouvant ainsi amener à des confusions lors de l'indexation et de la recherche. L'ontologie produite sera d'autant plus utile qu'elle est ciblée sur le domaine traité, la topographie, et reliée à des bases de données topographiques. Prenons un exemple simple pour illustrer cela : si le texte fait référence à la « ville de Pau », l'ontologie nous permettra de faire le lien entre les notions de 'ville' et de 'commune', les spécifications formelles nous permettent de savoir que les communes sont stockées dans la base de données géographiques BDTOPO dans la classe 'commune' du thème 'administratif', et donc ainsi de retrouver la géométrie de la commune de Pau, qui servira d'approximation spatiale de la notion de ville de Pau évoquée dans le texte.

4.2 Intégration de données topographiques

Une autre application de l'ontologie créée vise l'interopérabilité de bases de données géographiques. Il s'agit de permettre l'intégration de bases de données géographiques hétérogènes, afin de disposer d'un ensemble de données cohérent. La détection et la résolution de l'hétérogénéité sémantique (Bishr, 1998) constitue, aujourd'hui encore, l'un des principaux obstacles à cette intégration. L'utilisation d'ontologies en tant qu'outils permettant de spécifier sans ambiguïté la sémantique de termes au sein d'une communauté fait actuellement consensus dans le domaine de l'intégration d'informations (Partridge, 2002) , (Hakimpour et Timpf 2001).

Ainsi, une première application permettant l'appariement de schémas, c'est-à-dire la détection des classes de deux schémas de bases de données qui représentent les mêmes types d'entités géographiques du monde réel, a été développée (Abadie, 2009a). Celle-ci repose sur le constat de l'existence, au sein des classes de bases de données géographiques, d'attributs dont le seul but est de préciser la nature exacte des instances de la classe. Leurs valeurs possibles, qui se réfèrent directement à des labels de concepts géographiques, constituent donc une information importante sur la sémantique de la classe concernée. Ainsi, à partir de chacun des schémas à apparier, on génère automatiquement l'ontologie sous-jacente de la base, en prenant en compte ces labels de concepts de géographiques dissimulés dans des valeurs d'attributs. Puis, afin de détecter les entités de schémas à apparier qui sont sémantiquement reliées, on procède à l'alignement des ontologies ainsi produites. C'est à cette étape qu'intervient l'ontologie topographique du domaine produite au sein du projet GéOnto. En effet, celle-ci est mise à profit, comme source de connaissances externe, pour l'appariement des schémas.

Afin de parfaire les résultats d'appariement de schémas obtenus, une approche complémentaire est envisagée, qui consiste, une fois les schémas appariés, à

apparier les instances des classes elles-mêmes. Les résultats de cet appariement d'objets géographiques, basé sur des techniques multi-critères (Saïs, 2007), (Olteanu 2008), pourraient venir parfaire ceux de l'appariement de schémas dans le cadre d'une approche basée sur une boucle de rétroaction.

Une autre approche consiste à réaliser l'intégration des bases de données géographiques en s'appuyant non seulement sur l'ontologie topographique, mais également sur les spécifications des bases. En effet, celles-ci constituent une source de connaissance extrêmement riche quant à la sémantique des bases qu'elles décrivent. Les spécifications sont des textes en langage naturel qui ne sont donc pas directement exploitables automatiquement sans une étape de formalisation (Gesbert et al., 2004). Des travaux antérieurs (Picard, 2007), ainsi que les travaux de GéOnto présentés précédemment, ont montré qu'il était possible d'utiliser des techniques de traitement automatique du langage naturel pour traduire les spécifications textuelles dans un langage formel. L'approche envisagée ici consiste donc à comparer deux jeux de spécifications formelles afin de détecter automatiquement divers types d'hétérogénéités (hétérogénéité sémantique, géométrique, différence de niveaux de détail, etc.) entre les bases de données à intégrer (Abadie, 2009b), et d'en déduire les liens d'appariements complexes existant entre leurs schémas respectifs.

5. Conclusion

Cet article a présenté le projet GéOnto, et en particulier une des ses tâches qui consiste à créer une ontologie relativement riche de concepts topographiques. Le projet est en cours, et les premiers résultats exposés ici tendent à montrer la faisabilité de l'approche. Une fois réalisée, cette ontologie sera mise à disposition pour, nous l'espérons, être exploitée dans d'autres applications que celles présentées ici.

Le projet GéOnto ne se limite néanmoins pas à cette constitution d'ontologies. Ses autres objectifs sont méthodologiques : ils visent à mettre au point des méthodes les plus génériques possibles d'analyse automatique des spécifications textuelles, d'alignement d'ontologie, de comparaison globale d'ontologies, d'indexation spatiale de texte, d'appariement de données géographiques, et enfin d'intégration de données géographiques.

Remerciements

Cette recherche est en partie financée par l'Agence Nationale de la Recherche à travers le projet GéOnto (ANR-O7-MDCO-005, http://geonto.lri.fr/).

Bibliographie

Abadie N. 2009a. Schema Matching Based on Attribute Values and Background Ontology, Proceedings of 12th AGILE International Conference on Geographic Information Science, 2-5 June, Hanover (Germany)

Abadie N., 2009b, Formal specifications to automatically identify heterogeneities, 12th AGILE International Conference on Geographic Information Science, Pre-Conference Workshop "Challenges in Spatial Data Harmonisation", 2 June 2009, Hannover (Germany).

Abadie N., Mustière S. 2008. Constitution d'une taxonomie géographique à partir des spécifications de bases de données. Actes de la conférence SAGEO, juin 2008, Montpellier.

Auger, A., Barriere, C., 2008. Pattern based approaches to semantic relation extraction: a state-of-the-art. Terminology, John Benjamins, 14-1,1-19. Academic Publishing.

Bishr, Y. 1998. Overcoming the Semantic and Other Barriers to GIS Interoperability. *Int. Journal of Geographical Information Science*, vol 12, n° 4, pp 299-314, 1998.

Brisson R., Boussaïd O., Gançarski P., Puissant A., Durant N., Navigation et appariement d'objets géographiques dans une ontologie. In *EGC '07*., p. 391-396. 2007.

Brodeur, J., 2004. Interopérabilité des données géospatiales: Élaboration du concept de proximité géosémantique. Thèse de doctorat, Université Laval, Québec.

Buitelaar, P., Cimiano, P., Magnini, B., 2005. Ontology Learning From Text: Methods, Evaluation and Applications. IOS Press.

Gaio M, Sallaberry C, Etcheverry P, Marquesuzaa C, Lesbegueries J , 2008. A global process to access documents' contents from a geographical point of view, Journal of Visual Languages and Computing. Vol. 19., Orlando, USA, Academic Press, Inc. 3-23

Gesbert N., Libourel T. et Mustière S., Apport des spécifications pour les modèles de bases de données géographiques. *Revue internationale de Géomatique*, vol 14, n° 2, pp. 239-257, Lavoisier, 2004.

Gruber T.R., Toward principles for the design of ontologies used for knowledge sharing. Formal ontology in conceptual analysis and knowledge representation. N. Guarino et R. Poli (dir.). Dordrecht: Kluwer academic, 1993.

Guarino N., Formal ontology and information systems. Formal ontology in information systems: proceedings of FOIS'98, Trento, Italy, 6-8 Juin 1998. N. Guarino (dir.) Amsterdam: IOS Press, pages 3-15, 1998.

Hakimpour, F., Timpf, S., Using Ontologies for Resolution of Semantic Heterogeneity in GIS. *Proceedings of 4th AGILE Conference on Geographic Information Science*, Brno, Czech Republic, 2001, p. 385-395.

Hamdi F., Zargayouna H., Safar B., Reynaud C., 2008. TaxoMap in the OAEI 2008 alignment contest, Ontology Alignment Evaluation Initiative (OAEI) 2008 Campaign - Int. Workshop on Ontology Matching, 2008.

Jacques M.P., 2005. Structure matérielle et contenu sémantique du texte écrit. CORELA - Cognition, Représentation, Langage - ISSN 1638-5748

Kamel M., Aussenac-Gilles N., 2009. Construction d'ontologies à partir de spécifications de bases de données. IC 2009, Hammamet, Tunisie.

Laurens F. 2006. Création d'une ontologie à partir de textes en langage naturel. Internship report, Master 1 Linguistique-Informatique, University Paris 7.

Lesbegueries J., 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, ACM SIGIR 2006.

Lin D. 1998. An Information-Theoretic Definition of Similarity, in proc. of the International Conference on Machine Learning – ICML-98, Madison, pp. 296-304.

Loustau P, 2008. Interprétation automatique d'itinéraires dans des récits de voyages. D'une information géographique du syntagme `a une information géographique du discours Thèse de doctorat, soutenue à l'Université de Pau et des Pays de l'Adour

Maedche, A. Ontology Learning for the Semantic Web, vol. 665. Kluwer Academic Pub. 2002.

Olteanu, A.-M., Appariement de données spatiales par prise en compte de connaissances imprécises, Thèse de doctorat, Université de Marne-La-Vallée, 2008

Partridge C., The role of ontology in integrating semantically heterogeneous databases. Rapport technique 05/02 LADSEB-CNR, Padoue, 2002.

Picard V., Instanciation automatique des liens entre ontologies et schémas de bases de données géographiques à partir des spécifications en langage naturel. Stage de Master 2 Document Electroniques et Flux d'Information, Université Paris 10 – Nanterre, 2007.

Reynaud C., Safar B. 2007. Techniques structurelles d'alignement pour portails Web, Revue RNTI W-3, Fouille du Web, ISBN : 978.2.85428.793.6, Cépaduès.

Roussey C., Laurini R., Beaulieu C., Tardy Y. et Zimmermann M., Le projet Towntology : Un retour d'expérience pour la construction d'une ontologie urbaine. Revue internationale de Géomatique, vol 14, n° 2, pp. 217-237, Lavoisier, 2004.

Saïs, F., Intégration sémantique de données guidée par un ontologie Thèse de doctorat, Université Paris-Sud 11. December 2007.

Sallaberry C., Baziz M., Lesbegueries J., Gaio M. 2007. Towards an IE and IR System Dealing with Spatial Information in Digital Libraries - Evaluation Case Study. ICEIS (5) 190-197

Schmid H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Int. Conference on New methods in language Processing.

Uitermark H. 2001. Ontology-Based Geographic Data Set Integration. PhD thesis, Universiteit Twente, the Netherlands, 2001.