Analyses linguistiques et techniques d'alignement pour créer et enrichir une ontologie topographique

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

Analyses linguistiques et techniques d'alignement pour créer et enrichir une ontologie topographique

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

** IGN / Laboratoire COGIT*
173 av. de Paris, 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
**** IRIT, Université de Toulouse*
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
christian.sallaberry@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. One of the goals of the GéOnto project is to build an ontology of topographic concepts. This ontology results from the 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 d'ampleur. Au niveau national, a été mis en place le géoportail, 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 sources d'information multiples et hétérogènes est d'appuyer l'intégration des données sur une ontologie du domaine concerné (Gruber, 1993), (Guarino, 1998). 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 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), mais peu d'ontologies ont été réalisées en pratique (Laurens,

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)

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. Leur originalité est d'exploiter, en plus d'éléments lexicaux ou syntaxiques, des indices jusqu'ici peu utilisés pour l'extraction de concepts et de relations : les entités nommées et la structure des textes. 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 Mustière 2010). Celle-ci prend la forme d'une hiérarchie d'environ 700 termes (voir un extrait Figure 1). Elle a été réalisée à partir des termes utilisés dans les spécifications de l'IGN, par analyse semi-automatique de ces documents (de leur vocabulaire et de leur structure) puis réorganisée interactivement. Cette taxonomie est un premier pas vers une ontologie topographique plus riche. Pour être plus intensément exploitée, cette taxonomie mérite 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, dont nous présentons ici la démarche globale (Figure 2).

Figure 1. *Extrait de la taxonomie initiale de concepts topographiques*

L'enrichissement de la taxonomie précédemment construite (Figure 2, 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 2, 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 2, c), afin de prendre en compte un autre point de vue sur la topographie.

Figure 2. *Approche générale suivie pour la constitution de l'ontologie*

Les ontologies mentionnées par d sur la Figure 2 reflètent le contenu des spécifications et, de ce fait, le point de vue et le vocabulaire utilisés dans chacune des bases de données qu'elles décrivent (en l'occurrence, la BDTOPO® de l'IGN exploitée dans un premier temps). Ces spécifications seront exploitées pour une des applications visées dans le projet GéOnto, à savoir l'intégration de bases de données géographiques (Gesbert *et al.* ,2004), (Abadie, 2009b) (cf. partie 4.2). Dans le contexte de constitution d'une ontologie topographique présenté ici, elles sont exploitées pour enrichir la taxonomie originelle de nouvelles relations, de propriétés et de concepts issus, en particulier, de l'analyse des définitions qu'elles contiennent. La démarche originale mise en œuvre, décrite en partie 3.1, consiste à exploiter non seulement le langage présent dans les spécifications, mais aussi leur structure (titres et sous-titres, énumérations, etc.) visible à travers leur mise en forme dans l'extrait présenté en Figure 3. L'analyse de cet extrait permet de spécifier que *Cap*, *Carrière, Grotte*, etc. sont des *éléments de l'orographie,* que *Aven*, *Cave*, *Gouffre* sont des *Grottes,* etc.

Figure 3. *Extrait des spécifications de la BDTOPO sur l'orographie*

Une représentation ontologique de ces informations est donnée en Figure 3. C'est ce type de résultat que nous souhaitons obtenir automatiquement et intégrer au sein d'une ontologie propre à chaque document de spécification (Figure 1, d).

Figure 4. *Représentation ontologique des informations contenues dans l'extrait du document décrit à la Figure 3.*

Les récits de voyage ont une structure très différente (voir un extrait Figure 5) et sont exploités d'une autre manière. On y recherche le vocabulaire utilisé pour nommer des concepts topographiques (rivière, rue, 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 pour mieux interpréter les termes identifiés (cf. partie 3.2). Dans un vocabulaire contrôlé (thesaurus, taxonomie, etc.), chaque terme a un seul sens et ce sens est représenté par un seul terme. Notons que dans le cadre du projet, il n'était pas possible de considérer de manière exhaustive tous les points de vue et, par exemple, traiter des documents tels que des cahiers des charges de topographie ou encore des documents cadastraux. Notre objectif, dans le projet GéOnto, est de tester prioritairement une chaîne de traitement afin d'enrichir une ontologie à partir de textes dont sont extraits des termes (associés à des entités nommées) rapprochés d'un thesaurus généraliste. La chaîne de traitement est néanmoins utilisable sur d'autres textes en y appliquant les mêmes patrons.

```
13 h 50 LA TUSSE DE MAUPAS (3.110 m)
```

```
   Les abîmes du Crabioules, tombant au N. sur un glacier qui se
déchire en cascades, sont terrifiants ; le Perdighero s'impose à
l'admiration, magnifique et dominateur ; les Monts-Maudits portent
fièrement leur royauté. Sur la France, toujours la mer de nuages,
flottante et moelleuse, emplie de lumière. De l'extrémité E. de la
crête qui forme le sommet, le Boum paraît, insignifiant, avec son
petit glacier, et le cirque où s'enveloppent de brouillard les lacs
Vert et Bleu.
```

Figure 5. *Extrait du récit de voyage « Au pays des Isards » des cinq frères Cadier, aux éditions Monhélios*

Une fois ces traitements effectués, nous disposons donc d'une taxonomie qui a l'avantage d'être relativement structurée (Figure 2, a), des ontologies de spécifications qui ont l'avantage de représenter de manière formelle des propriétés et relations entre concepts (Figure 2, d), et d'une liste de termes issus des récits de voyage qui a l'avantage de refléter un vocabulaire riche et plus grand public que celui des spécifications techniques (Figure 2, 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 enrichie (Figure 2, 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 BDTOPO®), rendant possible dans le futur une indexation spatiale fine des récits de voyage (cf. partie 4.1).

3. Outils mis en œuvre et premiers résultats

Nous détaillons dans cette partie des outils développés durant le projet GéOnto et concourants au développement de l'ontologie topographique enrichie.

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 aussi forme et sens au contenu (Jacques, 2005). Pour exploiter les documents de spécification

(Figure 2, b et Figure 3) dont la forme est particulièrement structurée et produire une ontologie propre à chacun d'eux (Figure 2, d), nous proposons ici une approche une approche qui s'appuie à la fois sur la structure matérielle du texte pour créer un premier noyau d'ontologie, et sur l'exploitation conjointe du contenu et de la structure 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.

3.1.1 Analyse de la structure du document

Les documents de spécification des bases de données du COGIT utilisent un style spécifique à chaque type d'information. Ces documents ont pu alors être transcrits en XML (Abadie et Mustière 2010), une balise spécifique étant associée à chaque style. La Figure 6 correspond à l'extrait du document présenté Figure 3.

```
<class name="Oronyme">
 <className>Oronyme</className>
 <geometryType>Ponctuelle bidimensionnelle</geometryType>
 <description type="definition">Détail du relief portant un nom.    </description>
 <description type="extensionalDefinition">Voir les différentes valeurs de l'attribut <nature></description>
 <description type="selectionPrinciple">Tous les oronymes figurant sur la carte au 1 : 25 000 en service.</description>
 <description type="geometryDescription">Centre du lieu nommé.    </description>
  <attributes>
        <attribute name="Nom">
            <attributeName>Nom</attributeName>
            <valueType>Caractères</valueType>
           <description type="definition">Nom du détail ou du relief.</description>
            <enumeratedValues />
         </attribute>
        <attribute name="Nature">
           <attributeName>Nature</attributeName>
           <valueType>Énuméré</valueType>
           <description type="definition">Attribut donnant plus précisément la nature du lieu nommé.</description>
           <enumeratedValues>
                 <value name="Cap">
                    <valueName>Cap</valueName>
                    <description type="definition">Proéminence dans le contour d'une côte.</description>
                    <description type="extensionalDefinition">Cap | Pointe | Promontoire</description>
                 </value>
                    <value name="Grotte">
                        <valueName>Grotte</valueName>
                        <description type="definition">Grotte naturelle ou excavation.</description>
                        <description type="extensionalDefinition">Aven | Cave | Gouffre | Habitation troglodytique</description>
                    </value>
                    <…  >
```

Figure 6. *Version XML de l'extrait du document présenté Figure 3.*

Ces documents semi-structurés traduisent différents niveaux de hiérarchie : par la mise en forme matérielle du texte, et par l'imbrication des balises. Le principe retenu ici consiste alors à définir comme labels de concepts tous les syntagmes nominaux isolés présents dans des éléments de structure (dans le cas des spécifications, il s'agit des titres et sous-tires, qui sont marqués par des balises spécifiques) ; la position relative de ces syntagmes balisés dans les documents permet d'identifier des relations sémantiques (souvent hiérarchiques : est-un ou partie-de) entre ces concepts. Nous basons donc notre analyse sur la règle générale suivante :

Si	*- Tag1 et Tag2 sont des balises sous la portée d'une même balise Tag*

- C_1 et C_2 sont des concepts respectivement étiquetés par les unités textuelles marquées par Tag1 et Tag2
Alors une relation sémantique existe entre C_1 et C_2.

Figure 7. *Règle d'analyse des balises pour créer concepts et relations sémantiques*

Une étude du schéma XML des documents a permis de déterminer les diverses manières d'identifier concepts, relations conceptuelles et propriétés, et donc de définir des règles spécifiques selon le modèle de la règle de la Figure 7. Par exemple, une relation sémantique existe entre les syntagmes nominaux marqués par les balises <className> et <attributeName> (relation entre une classe et un de ses attributs) lorsque ceux-ci sont sous la portée de la même balise <class>. La nature de la relation découle de la connaissance du domaine présentée dans la grille de lecture fournie dans les documents de spécification. L'analyse automatique du corpus à partir des différentes règles spécifiques fournit un noyau d'ontologie. Ce processus est décrit en détail dans (Kamel *et al.*, 2009). Le noyau d'ontologie obtenu est composé de 728 concepts et de 41 relations sémantiques.

3.1.2 Analyse du texte libre

Nous avons aussi choisi d'utiliser des patrons lexico-syntaxiques pour repérer de nouveaux concepts et 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 relations. Pour analyser ce champ, une première procédure d'annotation permet de caractériser les différents fragments de texte pour y reconnaître des concepts du noyau de l'ontologie construite, ou des termes identifiés par un extracteur de termes. Afin de mieux gérer la mise en relation des termes marqués par les balises et ceux présents dans les relations, ces patrons prennent également en compte la structure et le balisage. Les patrons lexico-syntaxiques permettent ensuite d'annoter certaines parties des définitions comme des marques de relations.

Les définitions des documents de spécification que nous avons exploitées correspondent au schéma "*terme défini : énoncé définitoire*", où le terme défini correspond au label du concept à définir, et l'énoncé définitoire (introduit par la balise <description type="definition">) représente une seule caractérisation du terme défini. Dans l'exemple de la Figure 6, un des termes définis est *Oronyme*, son énoncé définitoire est *Détail du relief portant un nom*.

L'exploitation de ce type de définition est basée sur la structure syntaxique de l'énoncé définitoire. Outre l'annotation par les concpets reconnus, ces définitions ont fait l'objet d'annotations réalisées par différentes ressources linguistiques (extracteur de termes, patrons lexico-syntaxiques, analyseur morpho-syntaxique, etc.). *Terme* désigne alors un terme, *Propriete* une propriété caractérisée grammaticalement par

un adjectif ou un complément du nom, et *Marqueur* un marqueur linguistique de relation lexicale. Nous avons identifié les 3 cas suivants :

cas 1 : Un terme est défini par un seul autre terme (cas de quasi synonymie). Les termes défini et définitoire sont alors associés au même concept.
Exemple : <valueName> Cascade </valueName>
< description type="definition"> [Chute d'eau] $_{\text{Terme}}$ </description>

cas 2 : Un terme est défini par un autre terme auquel sont associées des propriétés. Le terme de l'énoncé définitoire est alors considéré comme un terme plus générique que le terme à définir. La ou les propriétés sont associées au concept relatif au terme à définir.
Exemple : <valueName>Terrain de sport</valueName>
<description type="definition"> [Equipement sportif] $_{\text{Terme}}$ [de plein air] $_{\text{Propriété}}$ </description>

cas 3 : Un terme est défini par une relation lexicale le reliant à un autre terme (une ou plusieurs propriétés pouvant éventuellement être associées à ce dernier). La relation lexicale correspondante est établie entre les concepts relatifs aux termes défini et définitoire. La ou les propriétés sont associées au concept relatif au terme à définir.
Exemple : <valueName> Tronçon de route </valueName>
< description type="definition"> [Portion de] $_{\text{Marqueur}}$ [voie de communication] $_{\text{Terme}}$ [destinée aux automobilistes] $_{\text{Propriété}}$ </description>

Nous disposons actuellement d'une base de 27 patrons de ce type, intégrant lexique, syntaxique et structure, permettant de traiter les cas 1, 2 et certaines configurations de la relation de méronymie (« X est une partie de Y », « Y est composé de X », « X est un tronçon de Y », etc.) pour le cas 3. Nous donnons ci-dessous un exemple de patron traitant une définition correspondant au cas 1.

{type:class}
{type:className} <Concept> {type:SN} </Concept> {type:className}
{type:geometryType} {type:sentence}{type:geometryType}
{type:description type="definition"}<Terme> {type:SN} </Terme> {type:description}
→ {*a-pour-terme* (Concept, Terme)}

Ce patron s'applique lorsque les chaînes lexicales marquées respectivement par les balises <className> et <description type="definition"> sont des syntagmes nominaux (SN), et lorsque ces balises sont sous la portée de la balise <class> et séparées par la balise <geometryType>. Ces SN sont respectivement annotés *Concept* et *Terme*. Une relation *a-pour-terme* est alors établie entre le concept et le terme.

L'enrichissement du noyau de l'ontologie obtenu à l'étape précédente est réalisé conformément à l'algorithme suivant :

L'analyse des définitions a permis d'enrichir le noyau d'ontologie de 253 concepts, de la relation de méronymie, et d'associer de nouveaux termes et de nouvelles propriétés aux concepts.

D'autres patrons lexico-syntaxiques correspondant à d'autres types de relations (essentiellement hyperonymie, fonction et artefact) ont été mis au point à partir d'une analyse linguistique de la totalité des textes des définitions. Ces nouvelles relations sont en cours de validation afin d'être intégrées à l'ontologie.

3.2 Traitement automatique des récits de voyage

L'objectif traité ici est d'exploiter les termes rencontrés dans les récits de voyage (Figure 2, c et Figure 5) et associés à des toponymes connus, pour déterminer le vocabulaire utilisé dans ces documents pour décrire la topographie. Ce vocabulaire (Figure 2, e) servira à enrichir l'ontologie obtenue à l'issue de l'étape précédente.

Nous avons analysé 14 livres (récits de voyage dans les Pyrénées) à l'aide d'une chaîne de traitement qui permet l'annotation des informations spatiales détectées dans un document textuel (Lesbegueries *et al.*, 2006), (Gaio, 2008), (Loustau, 2008). Cette annotation s'appuie sur des ressources géographiques diverses (bases de données de toponymie comme la BDNYME de l'IGN ou des gazetteers contributives comme Geonames) pour l'identification des entités nommées spatiales (ENs). La chaîne de traitement que nous proposons est présentée en Figure 8 (certaines étapes finales sont utiles à l'indexation spatiale des documents, évoquée en partie 4.1).

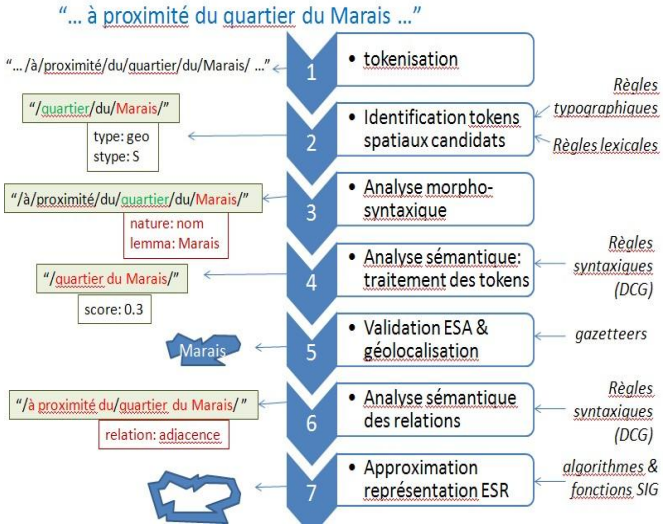

Figure 8. *Chaîne de traitement d'information spatiale dans des documents textuels*

L'étape (1) s'appuie sur une « tokenisation » classique. Nous adoptons ensuite une démarche de lecture « active », qui consiste à marquer rapidement des entités nommées spatiales (ENs) candidates puis à appliquer les étapes suivantes de l'analyse à ces ENs 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 », … 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 », … A titre d'exemple, la Figure 9 illustre les entités nommées et les syntagmes nominaux associés détectés par application de la chaine de traitement appliquée au texte de la Figure 5. Ainsi, le toponyme « Crabioules » est identifié et le syntagme associé « abîme » devient candidat à l'enrichissement de l'ontologie (Fig.1-e).

[p2]De LA TUSSE DE MAUPAS (3.110 m). Les [abîmes du Crabioules], tombant au N. sur un glacier qui se déchire en cascades, sont terrifiants ; le Perdighero s'impose à l'admiration, magnifique et dominateur; les Monts-Maudits portent fièrement leur royauté. Sur la France, toujours la mer de nuages, flottante et moelleuse, emplie de lumière. De l'extrémité E. de la crête qui forme le sommet, le Boum paraît, insignifiant, avec son petit glacier, et le cirque où s'enveloppent de brouillard les lacs Vert et Bleu.[/p]

Figure 9. *Résultat du traitement du texte de la Figure 5*

Nous obtenons ainsi un ensemble de termes associés à des ENs. 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 » dans « le gave de Pau »), contrairement à d'autres (comme « maire » dans « le maire de Pau »). L'analyse des termes associés aux ENs 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 (i.e. on ne connaît alors que les concepts au niveau de « grotte » et « crête » dans Figure 1, Figure 3, et Figure 4). Ce taux passe à 10% (33% du nombre total d'occurrences) si on utilise l'ontologie issue des spécifications liées à BDTOPO® (on connaît alors les concepts de niveau « aven », « gouffre » dans Figure 1, Figure 3, et Figure 4), ce qui montre l'intérêt de cette ontologie.

La liste des termes associés à des ENs est, dans un second temps, complétée à l'aide du thésaurus RAMEAU[3] (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

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

publique 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.

Figure 10. *Exemple de notice descriptive RAMEAU décrivant le terme Grottes*

La liste des termes associés à des ENs ainsi complétée a été utilisée pour enrichir la taxonomie initiale. Notre méthodologie d'enrichissement vérifie s'il est possible d'identifier des proximités sémantiques entre ces termes et les concepts de la taxonomie à enrichir. Si l'on prend l'exemple précédent du toponyme « Crabioules », identifié dans les récits de voyage par la chaîne de traitement de la Figure 4, 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, ici « col » et « mont ». *Abîme* n'est quant à lui pas représenté dans la taxonomie mais a cependant pour terme vedette *Grotte* dans RAMEAU (Figure 10). *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. Nous affinons actuellement ce processus d'enrichissement afin de détecter les termes de RAMEAU porteur d'un sens géographique.

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, comme décrit dans les parties précédentes 3.1 et 3.2 (Alignement/Fusion Figure 2). Ce processus de découverte de mappings, appelé *alignement d'ontologies,* est une fonction *f* qui s'applique sur deux ontologies *O* et *O'*, avec un ensemble de paramètres *p* (poids, seuils, etc.) et un ensemble de ressources externes *r*, et produit un ensemble de mises en correspondance *A=f(O,O,p,r)*. L'alignement comprend plusieurs étapes (Ehrig 2007) : l'extraction des données à rapprocher, la sélection des couples d'éléments à comparer, le calcul d'une similarité pour chaque couple sélectionné, la déduction de l'alignement à partir des mesures de similarités préalablement calculées. Chaque mode de calcul d'une mesure de similarité correspond à l'exécution d'une technique d'alignement particulière. Plusieurs classifications de ces techniques ont été proposées dans la littérature (Rahm *et al.,* 2001), (Klafoglou *et al,*. 2003) (Euzenat *et al.,* 2007). De manière générale, on distingue les techniques terminologiques (exploitant les labels des différents composants des ontologies), les techniques structurelles (exploitant la structure des ontologies ou de leurs composants), les techniques extensionnelles (exploitant les données peuplant les ontologies) et les techniques sémantiques (comparant les interprétations, ou plus exactement, les modèles des entités considérées). Les résultats des techniques peuvent nécessiter d'être agrégées avant d'être interprétés.

De nombreux systèmes d'alignement mettent en œuvre ce processus d'alignement en exploitant des techniques variées. Dans le projet GéOnto, l'alignement est réalisé à l'aide du système *TaxoMap* (Reynaud *et al*., 2007), (Hamdi *et al.*, 2009). Ce système exécute différentes techniques d'alignement séquentiellement mais contrairement aux systèmes d'alignement qui mettent en œuvre cette stratégie dans le but d'améliorer leur connaissance sur l'existence d'une similarité entre les éléments rapprochés, tels Cupid (Madhavan et al. 2001) ou Falcon-AO (Jian et al. 2005), les mappings proposés ici ne résultent de l'application que d'une technique et d'une seule. Une première technique est appliquée sur l'ensemble des couples d'éléments étudiés. Une seconde technique est appliquée sur l'ensemble des couples pour lesquels un alignement n'a pas été trouvé à l'aide de la première technique, etc. Les techniques appliquées et l'ordre de leur exécution sont dans ce cas très importants. Dans *TaxoMap*, il s'agit d'éléments paramétrables. Ce paramétrage possible associé à l'existence d'un ensemble relativement exhaustif de techniques terminologiques et structurelles particulièrement bien adaptées à l'alignement de taxonomies comportant des descriptions très fines de domaines d'application ont fait de *TaxoMap* un système bien adapté à l'alignement des taxonomies géographiques dans le projet GéOnto. 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 découverte de mappings repose sur des techniques variées, terminologiques ou structurelles, qui s'appuient sur l'utilisation de l'analyseur morpho-syntaxique *TreeTagger* (Schmid, 1994). Celui-ci permet une lemmatisation et une catégorisation des mots qui composent un label. A la sortie de l'analyseur, les mots du label sont répartis en deux classes, *mots pleins* et *mots complémentaires*, suivant leur catégorie et leur position dans le label. Cette répartition est utilisée ensuite par une mesure de similarité appliquée aux labels des concepts vus comme des ensembles de tri-grammes (Lin, 1998) en donnant plus de poids aux tri-grammes communs appartenant à des mots pleins dans les labels comparés. 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 de *TaxoMap* permettent de sélectionner le concept le plus pertinent, parmi ces candidats.

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*.

- **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.

- **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 les mots pleins d'un des labels de C_S (cf. Figure 11). 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 sont considérés comme complémentaires quel que soit le sens de la relation d'inclusion (T_4).

Figure 11. *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 C_S et C_{Tmax}, et génère des mappings *isA* entre tous les spécialisants de C_S et C_{Tmax} (cf. Figure 12).

Figure 12. *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 BDTOPO® (Figure 2, d) a été alignée en tant qu'ontologie source (O_S) à l'ontologie cible qu'est la taxonomie initiale (Figure 2, a) (O_T). Les résultats d'alignement obtenus montrent que leur exploitation peut permettre d'enrichir O_T. L'exploitation s'appuie sur le type du mapping et la technique ayant permis de l'obtenir. Ainsi, les mappings du type *isEq* entre un concept C_S 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 C_S sont munis de propriétés, celles-ci peuvent venir enrichir O_T.

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).

Les mappings *isA*, les plus nombreux, vont donner lieu à des traitements différenciés suivant la technique ayant permis de les obtenir. Par exemple, pour les mappings issus de la technique T_2 identifiant une inclusion d'un label de C_T dans un des labels de C_S, une analyse manuelle des mappings montre qu'il est intéressant d'essayer de qualifier le contenu de la partie du label de C_S non présente dans C_T que nous appellerons dans la suite la *partie restante*. Ainsi, dans le mapping (*Place ou carrefour isA carrefour*), la partie restante est la chaîne de caractères « *Place ou* » et dans le mapping (*Gué ou radier isA gué*) la partie restante est la chaîne « ou radier». Dans ces deux cas, la partie restante contient un marqueur de conjonction comme « ou, et, / » et une chaîne de caractères qui correspond au label d'un concept de O_T comme « place » ou un concept de O_S comme « radier». Après discussion avec l'expert du domaine, celui-ci souhaite que le mapping correspondant au label d'un concept de C_S, contenant un marqueur de conjonction et faisant référence à deux concepts dont les labels exacts sont déjà présents dans l'ontologie cible O_T, ne soit pas considéré dans la phase d'enrichissement. Ainsi, si « *Place ou carrefour »* est un label d'un concept de *Cs* avec « place » et « carrefour » correspondant exactement au label d'un concept de O_T, le mapping correspondant ne produira aucun enrichissement. En revanche, si la partie restante de la conjonction est un concept de O_S qui n'existe pas dans O_T, comme « radier » dans l'exemple précédent, ce concept doit participer à l'enrichissement en étant introduit comme un frère du concept auquel il est relié, i.e. « radier » sera un frère de « gué ». De même, si la partie restante contient le label d'un autre concept de O_T sans en être le label exact, c'est la partie restante qu'il est intéressant d'essayer d'exploiter pour l'enrichissement. Ainsi, dans le mapping (*chemin ou sentier côtier isA chemin*), la partie restante « sentier côtier » contient le label du concept « sentier » de O_T et l'adjectif « côtier ». L'expression « sentier côtier » pourrait être retenue comme le label d'un nouveau concept spécialisant le concept « sentier ».

Si la partie restante ne contient pas de conjonction et se limite à des mots identifiés par l'expert comme des mots vides pour le domaine, par exemple les mots « de type » ou « zone », comme dans (*bâtiment de type industriel isA bâtiment industriel*), le mapping donnera plutôt lieu à une relation d'équivalence (*bâtiment de type industriel EquivalentClass bâtiment industriel*). Si la partie restante ne contient qu'un adjectif entre parenthèses le mapping pourrait donner lieu à la représentation de propriétés (*Taille*, propriété de *bâtiment industriel,* pour pouvoir représenter le concept *bâtiment industriel (grand)*).

Dans tous les autres cas, les mappings pourraient correspondre à des relations de subsomption (*bâtiment d'élevage industriel subClassOf bâtiment industriel*) s'ils sont repris à l'identique.

TaxoMap a été par ailleurs utilisé pour identifier les proximités sémantiques entre les termes extraits des récits de voyage complétés par le thesaurus RAMEAU (cf. section 3.2 et Figure 2, e) et l'ontologie initiale. Les parties extraites de RAMEAU ont été considérées comme des ontologies sources O_S à part entière ; elles ont été alignées avec l'ontologie cible O_T, TopoCarto_COGIT. L'alignement a

ensuite été complété par une phase de vérification de la compatibilité du contexte des ontologies alignées, avant d'envisager l'enrichissement proprement dit. Cette étape permet de distinguer parmi les termes des entités nommées considérées comme candidats pour l'enrichissement, ceux qui sont des termes topographiques comme *gave* de ceux qui n'en sont pas comme *maire*. Elle consiste à vérifier l'existence d'au moins deux mappings entre les deux ontologies, un premier mapping d'équivalence entre un c_S et un c_T, et un autre mapping jugé *fiable* entre un autre concept c_{S2} avec un concept c_{T2} distinct de c_{T1} mais qui lui est relié. La notion de mapping *fiable* recouvre les mappings d'équivalence, d'inclusion par la technique T_2 ou de très grande proximité lexicale (T_5). On qualifie de *reliés* les concepts d'une même ontologie qui sont en relation mutuelle de père, fils ou frères. Lorsque le test est vérifié, les domaines des deux ontologies sont valides. Si on ne trouve pas au moins deux mappings, l'un d'équivalence, l'autre *fiable*, l'ontologie source est rejetée, les domaines étant jugés incompatibles. Si un mapping d'équivalence et un mapping fiable existent mais que les deux concepts de O_T ne sont pas reliés, le domaine de O_S doit être validé par l'expert. La Figure 13 présente dans les encadrés les concepts identifiés comme équivalents lors de l'alignement. Il existe au moins deux mappings fiables tels que les concepts considérés dans les mappings soient en relation deux à deux (ici, quatre mappings d'équivalence relatifs à quatre concepts tous reliés entre eux), le domaine de la partie extraite de RAMEAU a donc été jugé compatible avec celui de l'ontologie à enrichir.

Les travaux décrits montrent que l'alignement est un préalable à l'enrichissement, qu'il produit des résultats riches mais devant faire l'objet de validation et de traitements complémentaires spécifiques, en interaction avec l'expert. Les traitements ont été implémentés dans l'environnement TaxoMap Framework (Hamdi *et al.*, 2010b) à l'aide de patterns (Hamdi *et al.,* 2011).

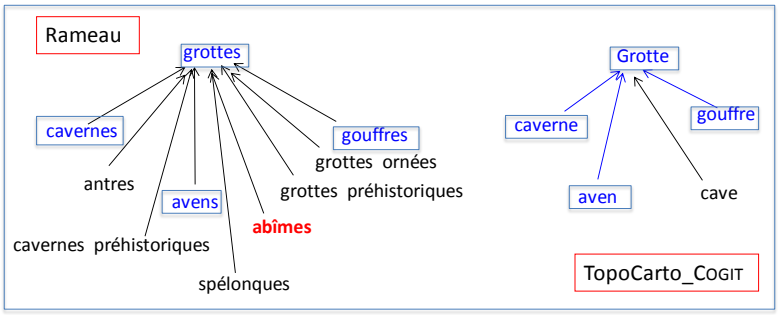

Figure 13. *Exemple d'extrait de RAMEAU dont le domaine a été jugé valide*

1. Utilisations prévues de l'ontologie réalisée

Une ontologie riche telle que celle visée dans GéOnto permet d'affiner deux traitements typiques présentés ici: l'indexation spatiale de textes grand public, et l'intégration de bases de données géographiques.

4.1 Indexation spatiale de documents

Le problème traité ici correspond aux nouveaux besoins de valorisation des fonds documentaires patrimoniaux suscités par l'importante politique de numérisation mise en œuvre par les différentes instances de conservation des collections documentaires territorialisées (archives régionales, musées, médiathèques, etc.). Une part non négligeable de l'information contenue dans ces documents numériques fait référence plus ou moins explicitement à des entités géographiques. Or, la plupart des systèmes de gestion et de consultation de documents en ligne propose une indexation reposant sur l'exploitation de métadonnées produites manuellement, combinées à des méthodes de recherche d'information 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 de récits de voyage : elle consiste à associer une représentation symbolique aux entités nommées (ENs) détectées (Figure 8, étapes 4 et 6) puis, à calculer une représentation numérique de ces entités (Figure 8, étapes 5 et 7). Ces géométries, affectées aux ENs, sont stockées dans des index qui supporteront des scénarios d'interrogation et d'exploitation de tels fonds documentaires à partir de critères spatiaux.

Ainsi, la construction de la représentation symbolique associée à une EN passe par l'analyse du groupe nominal constituant le toponyme (désambiguisation). Par exemple, l'entité nommée « Artouste » aura une représentation symbolique différente selon la sémantique de l'élément topographique désigné. Qu'il s'agisse du lac, du pic ou de la vallée, la nature du toponyme est différente. Par voie de conséquence, les ressources invoquées pour la recherche et le calcul de la géométrie correspondante ainsi que les stratégies de parcours de ces ressources seront spécifiques. Nous analysons donc chaque EN afin d'en extraire le syntagme nominal (e.g. « pic ») associé au nom toponymique (e.g. « Artouste »). Nous proposons ensuite un module de parcours d'ontologie dont les paramètres en entrée sont le syntagme nominal (e.g. « pic »), l'ontologie et la propriété à consulter (e.g. « GéOnto », « label »), la fonction de calcul de similarité à utiliser (e.g. « QGrams »). Ce module construit dynamiquement une requête SPARQL, l'exécute puis retourne une liste de couples

concepts-score de similarité extraits de l'ontologie. Le concept de meilleur score sert d'étiquette sémantique à l'EN (e.g. « sommet » étiquettera « pic d'Artouste »).

La construction de la représentation numérique associée à l'EN « pic d'Artouste » consiste désormais à exploiter l'étiquetage sémantique pour accéder directement à la bonne ressource (e.g. éléments de nature « sommet » dans la table « Oronyme » de la BD-NYME de l'IGN). De même, la construction de la représentation numérique associée à des ENs complexes (e.g. « près du pic d'Artouste ») sera guidée par l'étiquette sémantique associée. La relation spatiale de proximité d'une voie de communication ou d'un cours d'eau (représentés par une polyligne) est appréhendée différemment de celle d'une vallée ou d'une commune (représentées par un polygone).

Comme nous l'avons exposé, l'analyse des termes associés aux ENs et l'utilisation des données des BD de l'IGN permet d'étiqueter automatiquement 5% des informations spatiales annotées (15% du nombre total d'occurrences). Ce taux passe à 50% de typage automatique des informations spatiales (75% du nombre total d'occurrences) si l'on utilise l'ontologie topographique GéOnto enrichie à partir d'une analyse combinée d'échantillons de textes grand public (récits de voyage) et du thésaurus RAMEAU (Kergosien *et al.*, 2009). Les premières expérimentations montrent un gain en précision des géométries associées aux ENs et en temps de calcul intéressant ; une évaluation est actuellement en cours.

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 cohérent de données. 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 al.,* 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 éléments de schémas à apparier qui sont

sémantiquement reliés, 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 ontologie de support (Aleksovski *et al.,* 2006), pour l'appariement des schémas. Tout d'abord, chaque ontologie générée à partir d'un schéma de base de données est alignée avec cette ontologie de support, à l'aide de techniques d'alignements d'ontologies utilisées dans l'outil *TaxoMap* (Hamdi *et al.,* 2009). Puis, les alignements entre ontologies issues de schémas de bases de données sont déduits à partir de ceux détectés entre ces dernières et l'ontologie de support. Enfin, les liens d'appariement entre schémas peuvent être dérivés des résultats d'alignement entre leurs ontologies sous-jacentes. La mise en évidence des concepts topographiques compris dans des valeurs d'attributs au niveau des ontologies sous-jacentes de chacune des bases à apparier permet d'augmenter le niveau de granularité de leurs schémas conceptuels respectifs et ainsi de faire émerger des appariements qui n'auraient pas pu être détectés en considérant les seuls noms de classes. Ainsi, les classes "Oronyme" de la BDTOPO® et "Site et curiosité touristique" de la BDCARTO®, dont les noms sont lexicalement et sémantiquement différents ont pu être appariées via leur valeur d'attribut commune "grotte".

De plus, l'utilisation de l'ontologie de support permet d'introduire, dans le processus d'appariement de schémas, des connaissances externes à celles issues des schémas de bases de données seuls. Certaines relations sémantiques impossibles à détecter en l'absence de connaissances externes ont ainsi pu être détectées. A titre d'exemple, l'attribut « Nature » de la classe « Point d'activité ou d'intérêt » de la BDTOPO® a comme valeur possible "Habitation troglodytique" et désigne donc des grottes ou des cavernes aménagées. Or, au niveau du schéma de la base, seule la valeur "Habitation troglodytique" apparaît explicitement. Dans la BDCARTO®, en revanche, les grottes aménagées sont représentées au sein de la classe « Site et curiosité touristique» sous la valeur « Grotte, gouffre aménagés » de l'attribut « Nature ». L'ontologie de support désignant le concept d'habitation troglodytique comme une spécialisation du concept de grotte, un appariement entre ces deux valeurs d'attributs pourtant lexicalement différentes a tout de même pu être détecté (voir Figure 14).

Une approche plus détaillée 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 connaissances extrêmement riche quant à la sémantique des bases qu'elles décrivent. Ainsi, on y trouvera des indications sur les critères de sélection que les entités du monde réel doivent remplir pour figurer dans la base : « les bâtiments de plus de $20m^2$ sont représentés ». Sont fournies également, des règles de saisie pour la géométrie des objets géographiques : « Les tronçons de route sont saisis à l'axe, au sol ».. 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). L'approche envisagée ici consiste donc à formaliser les spécifications

de chaque jeu de données à intégrer sous la forme d'ontologies d'application (Abadie et al., 2010) et à comparer ces ontologies 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. Ainsi, si les spécifications d'une base précisent que « seuls les bâtiments de plus de 20 m^2 sont saisis dans la classe « bâtiments »», tandis que celles d'une seconde base stipulent que « les bâtiments de plus de 50 m^2 sont inclus dans la classe « zones bâties » », alors l'appariement détecté entre ces deux classes devra spécifier que seules les instances de la classe « bâtiments » de la première base ayant une surface supérieure à 50 m^2 pourront correspondre à des instances de la classe « zones bâties » de la seconde base.

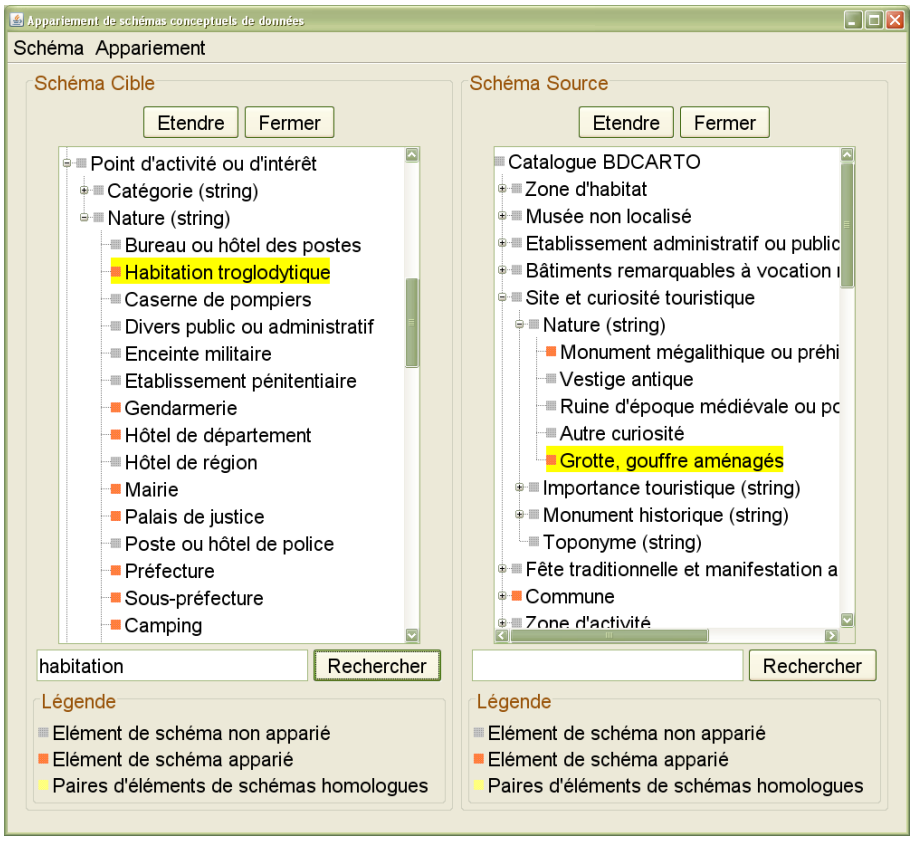

Figure 14. *Visualisation de liens d'appariement entre les schémas de la BDTOPO® et de la BDCARTO® illustrant l'intérêt de l'ontologie de support.*

5. Conclusion

Cet article offre un panorama des objectifs et réalisations du projet GéOnto, et en particulier une de 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.

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

Abadie N., 2009, « Formal specifications to automatically identify heterogeneities », 12th AGILE Int. Conf. on Geographic Information Science, *Pre-Conference Workshop on Challenges in Spatial Data Harmonisation,* 2 June 2009, Hanover (Germany), 2009b.

Abadie N., Mechouche A., Mustière S., 2010, OWL-based formalisation of geographic databases specifications, *17th International Conference on Knowledge Engineering and Knowledge Management (EKAW'10)*, 11-15 Octobre 2010, Lisbon (Portugal). 2010.

Abadie N., Mustière S. 2010. Constitution et exploitation d'une taxonomie géographique à partir des spécifications de bases de données. *Revue Internationale de Géomatique*, vol.20, n.2 , pp.145-174.

Aleksovski, Z., ten Kate, W., van Harmelen, F., « Exploiting the structure of background knowledge used in ontology matching ». *Proc. of the Ontology Matching Workshop at 5th International Semantic Web Conference*, Athens (Georgia) USA, pp. 13-24, 2006.

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

Bishr, Y., « 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*, pp. 391-396. 2007.

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

Ehrig M., *Ontology alignment: Bridging the semantic gap*, vol. 4 of Semantic Web and beyond Computing for Human Experience, Springer, 2007.

Euzenat J, Shvaiko P., *Ontology Matching*. Heidelberg (DE): Springer-Verlag, 2007.

Gaio M, Sallaberry C, Etcheverry P, Marquesuzaa C, Lesbegueries J , « 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., pp.3-23, 2008

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 ». In *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, pp. 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, pp. 385-395.

Hamdi F., Safar B., Niraula N., Reynaud C., « TaxoMap in the OAEI 2008 alignment contest », *Ontology Alignment Evaluation Initiative (OAEI) 2009 Campaign - Int. Workshop on Ontology Matching*, 2009.

Hamdi F., Reynaud C., Safar B., "L'approche TaxoMap FrameWork et son application au raffinement de mappings », *Actes de la 17ème conférence en Reconnaissance des Formes et Intelligence Artificielle, RFIA 2010*, Caen. 2010a.

Hamdi F., Reynaud C., Safar B., « Pattern-based Mapping Refinement », *EKAW 2010, 17th International Conference on Knowledge Engineering and Knowledge Management*, Lisbon, Portugal, 2010b.

Hamdi F., SafarB., Reynaud C., « Utiliser des résultats d'alignement pour enrichir une ontologie », EGC 2011, Brest, 25-28 janvier, 2011

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

Kamel M., Aussenac-Gilles N., « Construction d'ontologies à partir de spécifications de bases de données ». *Actes des journées francophones d'ingénierie des connaissances IC 2009*, Hammamet, Tunisie. pp. 85-96, 2009.

Kalfoglou Y., Schorlemmer M., Ontology matching : the state of the art, *The Knowledge Engineering Review*, 18, p. 1-31, 2003.

Kergosien E, Kamel M, Sallaberry C, Bessagnet MN, Aussenac-Gilles N., Gaio M, « Construction automatique d'ontologie et enrichissement à partir de ressources externes », *Actes des Journées Francophones sur les Ontologies, JFO 2009*, ISBN 978 1 60558 842 1, pp. 11-20, 2009.

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

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

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

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

Rahm E., Bernstein P.A., A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4), p. 334-350, 2001.

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

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.

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

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the Int. Conference on New methods in Language Processing*. pp. 44-49, 1994.

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