Construction automatique d'ontologies à partir de spécifications de bases de données

Mouna Kamel, Nathalie Aussenac-Gilles

Laboratoire IRIT, Université Paul Sabatier de Toulouse
{kamel, aussenac}@irit.fr

Résumé : Les méthodes classiques de construction automatiques d'ontologies à partir de textes exploitent le texte proprement dit. Nous étendons ces approches en prenant en compte la structure du texte, élément porteur d'information. Pour cela, nous nous basons sur des documents de spécifications de bases de données au format XML, pour lesquels le découpage structurel du texte correspond à une caractérisation sémantique de son contenu. L'idée est de tirer profit à la fois de la structure du texte et du texte rédigé. La méthode proposée consiste à utiliser la sémantique des balises et à caractériser leurs relations pour définir des règles de création de concepts et de relations sémantiques. Un noyau d'ontologie a été ainsi construit automatiquement à l'aide de ces règles, noyau ensuite enrichi par l'exploitation du texte en langage naturel à l'aide de patrons lexico-syntaxiques définis. Règles et patrons ont été implémentés sous Gate.

Mots-clés : Ingénierie des connaissances, extraction de relations, ontologie, documents structurés, traitement automatique de textes.
Communication appliquée

1 Introduction

Les méthodes de construction d'ontologies à partir de textes privilégient souvent l'analyse du texte proprement dit, que ce soit selon une approche statistique ou linguistique (Nédellec et Nazarenko, 2003), (Aussenac et al., 2008), (Maedche, 2002), (Buitelaar et al., 2005). La plupart de ces travaux montrent la nécessité d'intégrer différents outils de TAL, et soulignent la complémentarité entre identification de concepts et extraction de relations. Notre objectif est d'étendre ces approches classiques en prenant en compte la structure explicite des textes, lorsque cette structure caractérise la sémantique d'unités textuelles repérées et leur hiérarchisation. C'est le cas notamment des documents de spécifications de bases de données, riches en descriptions de concepts. Ce type de document nous permet alors d'identifier i) les concepts spécifiés, ii) les relations entre ces concepts à l'aide de la structure hiérarchique du document et iii) de nouveaux concepts et/ou de nouvelles relations en exploitant le texte rédigé. Le format XML est un bon format de représentation de tels documents : il intègre les notions de description sémantique de la structure d'un document (le même fichier situe le texte en langage naturel au sein de cette structure) et de structure hiérarchique. En exploitant donc la sémantique véhiculée par les balises, la structure hiérarchique du document et l'information contenue entre les

balises, la complémentarité entre identification de concepts et extraction de relations est assurée, à l'instar des approches classiques.
Comme XML est un langage de balisage non prédéfini, nous nous plaçons dans une perspective d'Ingénierie des Connaissances. Nous nous limitons à un domaine spécifique et nous tenons compte à la fois de la sémantique associée aux balises et à leurs relations, et de connaissances d'arrière-plan. Une correspondance peut alors être établie entre les fragments de textes balisés et des éléments d'ontologie.
La méthode que nous présentons consiste à extraire en priorité et à l'aide de cette correspondance (exprimée par des règles) les concepts et les relations sémantiques, étape essentielle pour définir un noyau de l'ontologie, puis à enrichir ce noyau en exploitant le texte en langage naturel présent dans le document. Nous rappelons d'abord les différentes recherches sur l'identification des relations sémantiques à partir du contenu du texte ou de leur structure (partie 2). La partie 3 décrit comment notre méthode conjugue certaines de ces approches pour extraire des connaissances à partir de la structure du texte et de son contenu. La partie 4 présente la mise en œuvre de notre méthode au sein du projet GEONTO, dont l'objet est de construire automatiquement des ontologies à partir de spécifications de bases de données géographiques. La 5ème partie donne une évaluation de notre méthode dans ce contexte. Nous dressons enfin le bilan actuel de nos travaux au sein du projet (partie 6) et présentons les perspectives pour les améliorer.

2 Méthodes d'identification de relations sémantiques

L'identification de relations est utile pour construire automatiquement une ontologie ou pour l'enrichir par des relations entre instances. Deux familles de techniques extraient des relations sémantiques à partir de textes : les approches statistiques et les approches linguistiques. Les approches statistiques consistent à étudier les termes co-occurrents et la similarité entre leurs contextes syntaxiques (Hindle, 1990), (Grefenstette, 1994), à prédire les relations à l'aide de réseaux bayésiens (Weissenbacher et Nazarenko, 2007) ou de techniques de Text Mining (Grcar et al., 2007), ou encore à inférer des connaissances à l'aide d'algorithmes d'apprentissage (Guiliano et al., 2006). Ces méthodes sont efficaces, mais n'identifient pas toujours la sémantique de la relation. L'approche linguistique fait appel à des analyses syntaxiques ou des calculs de dépendance pour identifier les relations argumentatives (sujet, verbe, objet) (Jacquemin, 1997), (Bourigault, 2002), ou définit des patrons lexico-syntaxiques pour reconnaître les marques linguistiques des relations sémantiques (Aussenac et Seguela, 2000). Ainsi la sémantique des relations est bien identifiée, mais la variabilité de leur sémantique et de leur expression en corpus oblige à multiplier les patrons et rend l'approche coûteuse.
Ces techniques s'appliquent au niveau interne de la phrase, alors que d'autres études ont pour niveau d'analyse le texte lui-même. L'objectif est alors assez différent : il ne s'agit plus de trouver des relations entre concepts, mais des relations sémantiques plus diffuses entre les différentes unités textuelles repérées. Ces liens peuvent être décelés soit à l'aide de marqueurs linguistiques (Asher et al., 2001), soit en exploitant la

structure matérielle du texte, (Virbel & Luc, 2001) ayant montré que la matérialité d'un texte participe à son sens, soit en combinant les marqueurs linguistiques et la structure du texte (Charolles, 1997).
Les traitements statistiques et linguistiques sont largement utilisés pour la construction automatique d'ontologies (Buitelaar et Ciminao, 2005), (Maedche, 2002), alors qu'il n'existe à notre connaissance que très peu de travaux qui aient exploité les relations du discours ou la mise en forme matérielle d'un texte dans ce but (Laurens, 2006).
L'idée que nous développons ici est de combiner une approche basée sur la structure du document à une approche linguistique pour tirer profit à la fois de la structure du texte et du texte lui-même. Notre objectif est d'élaborer des ontologies plus riches que ne pourraient fournir ces approches prises indépendamment. Nous présentons dans le paragraphe suivant notre méthode pour l'analyse d'un document de type spécifications de bases de données (au format XML) en vue de la conception d'une ontologie.

3 Méthodologie

La méthode proposée pour la construction d'ontologies consiste à exploiter un document de spécifications de bases de données au format XML en associant une analyse de la structure du texte (sémantique des balises et de leur hiérarchie) à une analyse linguistique du texte présent entre les balises (exploitation du langage naturel, de signes typographiques). Ces deux traitements sont réalisés indépendamment, sauf dans quelques règles où la détection de concepts ou de relations découle à la fois de la sémantique d'une balise et du texte qu'elle marque.

3 .1 Exploitation des balises

Le langage de balises fournit à la fois une description du texte et, par le processus d'imbrication des balises, des liens existant entre les unités textuelles balisées. Dans le cas où ces unités textuelles sont simples et ne renvoient chacune qu'à un seul concept, l'analyse des balises et de leur imbrication hiérarchique permet d'identifier des concepts et des relations sémantiques. La règle générique mise en œuvre pour l'identification des relations est la suivante *(">>" est l'opérateur de recouvrement)* :

si	*A, B et C sont des balises, B et C introduisant des concepts*
	A >> B
	A >> C
alors	*Il existe une relation conceptuelle R_{BC} entre B et C.*

Une étude préalable de la nature des balises et de leur hiérarchisation a pour but de déterminer les balises B et C qui introduisent des concepts et la sémantique des relations conceptuelles R_{BC} (hiérarchiques, méronymiques, et autres) associées aux relations entre balises B et C. Cette étude permet de définir autant de règles du format de la règle générique que nécessaire. L'analyse automatisée du corpus à l'aide de ces règles fournit un noyau d'ontologie.

3 .2 Exploitation du texte en langage naturel

Le corps du document XML correspond au texte en langage naturel et peut contenir de l'information intéressante à exploiter pour enrichir l'ontologie obtenue à l'issue de l'exploitation de sa structure (section 3.1). Selon C. Barrière (Barrière et Agbado, 2006) et Hearst (Hearst, 1992), un des moyens de qualifier "des contextes riches en connaissances" est qu'ils contiennent des marques linguistiques de relations sémantiques. Nous avons choisi d'utiliser des patrons lexico-syntaxiques pour repérer des relations sémantiques (Auger et Barrière, 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. Dans le cas particulier de la recherche de relations, le patron caractérise un ensemble de formes linguistiques dont l'interprétation est relativement stable et correspond à une relation sémantique entre termes (Rebeyrolle & Tanguy, 2000). L'application de tels patrons nécessite de traiter préalablement le texte en appliquant différents outils du TAL (tokenizer, lemmatiseur, analyseur syntaxique, etc.). Les patrons exploitent les étiquettes morpho-syntaxiques ou sémantiques attribuées par ces logiciels. Ainsi, la forme des patrons dépend à la fois du logiciel de définition et de projection des patrons, et des analyses et étiquetages effectués sur les textes.
Notre approche complète l'exploitation de la structure des documents (§ 3.1) en identifiant, à l'aide de patrons préalablement définis, de nouveaux concepts, des relations, voire des propriétés dans les parties rédigées du document. Elle peut être reproduite sur tout document de ce genre, moyennant la définition de règles d'interprétation des balises et des patrons.

4 Cadre d'application : projet GEONTO

Au sein du projet GEONTO[1], un des partenaires dispose de bases de données géographiques hétérogènes et a pour objectif l'interopérabilité de ces bases. Pour cela, le projet prévoit de fournir une ontologie par base de données, et d'aligner les ontologies obtenues en vue de produire une ontologie de référence. Les ontologies produites seront issues des spécifications de ces bases de données, et non de leurs schémas comme dans les travaux de (Tirmizi et al., 2008), (Gardarin et al., 2008)

4.1 Description des données

L'expérimentation décrite ici porte sur la base de données BDTopo qui sert de référence pour la localisation de l'information relative aux problématiques d'aménagement, d'environnement ou d'urbanisme. Les spécifications de cette base de données, disponibles au format WORD utilisant un style pour chaque type d'information, ont été automatiquement traduites en XML : c'est ce document qui

[1] Projet ANR-07-MDCO-005, http://www.lri.fr/geonto : collaboration entre le COGIT, le LRI (Université paris Sud), le LIUPPA (Université de Pau) et l'IRIT (Université de Toulouse)

servira à la construction de l'ontologie. Un extrait du document de spécification de la base BDTopo (classe Tronçon de chemin) est présenté figure 1, et la portion du fichier XML correspondant à ces spécifications est donnée figure 2.

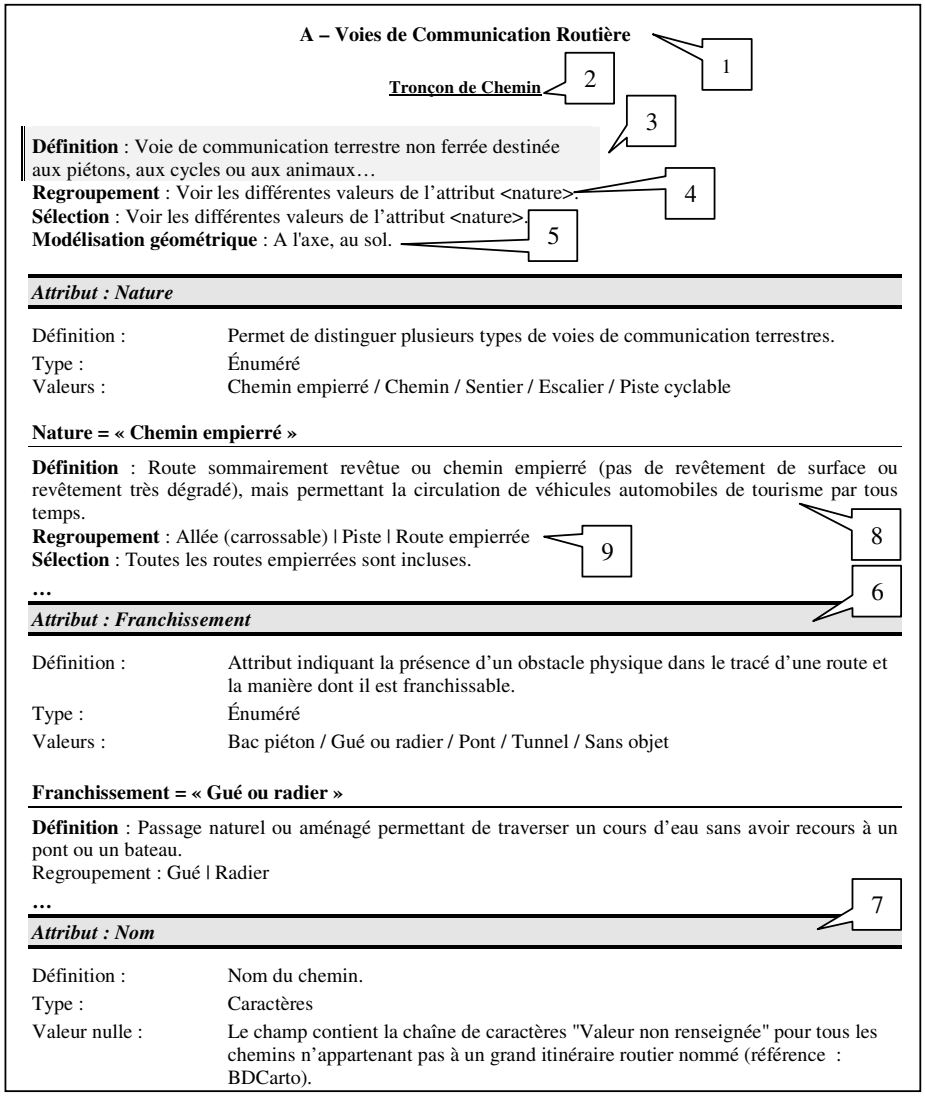


Figure 1 : extrait des spécifications concernant la classe **Tronçon de chemin**.

Sur la figure 1, les classes d'objets (2) sont réparties dans 9 domaines d'information (1). Les objets d'une même classe mentionnés dans le champ *Regroupement* (4) partagent une même définition (3), un même type de géométrie (5) et une même liste

d'attributs. Ces attributs supportent des informations à caractère qualitatif (liste d'objets) (6) ou quantitatif (attribut de type non énuméré) (7). Chaque valeur d'attribut a sa propre définition (8) et peut représenter une liste d'objets (9).

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<document> <domaine>
  <nom_domaine>A - Voies de communication routière</nom_domaine>
  <classe>
    <nom_classe>Tronçon de chemin</nom_classe>
    <definition>Voie de communication terrestre … </definition>
          …
    <regroupement> Voir les différentes valeurs de l'attribut &lt;nature&gt;
    </regroupement>
          …
    <attributs>
    <attribut>
      <nom_attribut> Nature </nom_attribut>
      <definition> Permet de distinguer plusieurs… </definition>
      <type> Énuméré</type>
      <valeurs> Chemin empierré/Chemin/Sentier/Escalier … </valeurs>
      <attribut_valeur>
        <valeur> Chemin empierré </valeur>
        <definition> Route sommairement revêtue … </definition>
        <regroupement> Allée (carrossable)|Piste|Route empierrée
        </regroupement>
      </attribut_valeur>
          …
    </attribut>
  </classe></domaine> </document>
```

Figure 2 : extrait du document XML correspondant à la classe **Tronçon de chemin**

4.2 Traitement du document XML

4.2.1 Exploitation des balises et de leur imbrication

Une étude systématique des balises du document XML et de leur imbrication a permis de déterminer comment identifier concepts, relations conceptuelles et propriétés.
Concepts : chacun des termes existant entre les balises <*nom_domaine*>, <*nom_classe*>, <*regroupement*>, <*nom_attribut*> (lorsque l'attribut est qualitatif) et <*valeur*> donne lieu à la définition d'un concept dont le label correspond à ce terme
Relations hiérarchiques : émanent des relations entre les balises identifiées (à savoir <*nom_domaine*> et < *nom_classe*>, < *nom_classe*> et <*regroupement*>, <*nom_attribut*> et <*valeur*>, <*valeur*> et <*regroupement*> et des concepts associés.
Propriétés : portées par les termes encadrés par les balises <*attribut*> lorsque ceux-ci sont quantitatifs, et associées aux concepts encadrés par les balises <*nom_classe*>
Relations conceptuelles autres que hiérarchiques : portées par les termes encadrés par les balises <*attribut*> lorsqu'ils sont qualitatifs, et associées aux concepts dont les noms sont encadrés par les balises <*nom_classe*>
La sémantique des propriétés et des relations autres que hiérarchiques ne peut être déterminée que suite à une analyse de la balise <type> et du texte qu'elle marque.

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.

4.2.2 Exploitation du texte à l'aide de patrons lexico-syntaxiques

Le document de spécification de la base de données BDTopo contient des textes très courts et très synthétiques, donc assez pauvres en matière d'expression de relations entre concepts. Le champ *définition* renferme néanmoins quelques expressions de la relation de méronymie ou de définition de propriétés. Soit l'extrait suivant :

<classe>
<nom_classe> Tronçon de route </nom_classe>
<définition> **Portion de** voie de communication destinée aux automobiles >/definition>
</classe>

Figure 3. Exemple de définition de la classe/concept **Tronçon de route**

a) La relation de méronymie peut être caractérisée par les termes *partie de*, *portion de, constitué de, formé de, composé de,* etc. Le patron permettant d'identifier cette relation et écrit selon le formalisme JAPE est le suivant :

(({Token.lemme== "portion"}|{Token.lemme== "partie"}|...)
({Token.lemme== "de"}) ({Terme}) :annot
) - - > annot.ANNOT = {kind="Partie", rule="Rule1"}

Ce patron recherche un des mots *partie*, *portion*, *composer*, etc. suivi du mot *de*, suivi d'un *Terme* (obtenu à l'aide d'un extracteur de termes, et annoté comme tel). Ce terme réannoté *Partie* (partie droite de la règle) à l'aide de la règle *Rule1*, sera relié à la classe qui recouvre le champ <définition> où s'applique le patron, par la relation *partie-de*. Dans notre exemple, le concept *Tronçon de route* sera relié au concept *Voie de communication* par la relation *partie-de.*

b) Une propriété peut être identifiée par une forme lexico-syntaxique composée d'un terme désignant un concept, suivi d'un participe passé (PP) puis d'une préposition (PREP). Le patron suivant, appliqué à l'exemple de la figure 3 permet d'identifier une nouvelle propriété *destiné aux automobiles* pour le concept *Tronçon de route* :

({Concept}
({Token.category==PP} {Token.category==PREP} {Terme}) : annot
) - - > annot.ANNOT = {kind="Propriete", rule="Rule2"}

4.2.3 Mise en œuvre

Ces traitements ont été réalisés à l'aide de la plate-forme GATE[2] dont le principe est d'appliquer successivement sous forme de pipeline des ressources linguistiques et/ou des ressources de traitement sur un corpus. Le résultat est un corpus annoté, et ces annotations peuvent faire l'objet de divers traitements à l'aide de règles écrites en JAPE ou en Java. Comme GATE dispose d'une *Ontology API,* une ontologie peut être construite à partir de l'exploitation de ces annotations. Par ailleurs, GATE considère une balise XML du document original comme une annotation, ce qui permet :

1) d'exploiter les balises XML comme des annotations, l'imbrication des balises comme des recouvrements d'annotations, pour construire une première ontologie

[2] General Architecture for Text Engineering : plate-forme d'ingénierie linguistique développée à l'Université de Sheffield (http://gate.ac.uk)

2) d'appliquer des outils du TAL pour produire des étiquettes morpho-syntaxiques, puis de les exploiter par des patrons pour enrichir l'ontologie

Nous avons défini 6 règles exploitant les balises pour réaliser le noyau de l'ontologie et 5 règles correspondant aux patrons pour enrichir ce noyau. Nous donnons en figure 5, un extrait de l'ontologie obtenue, qui correspond aux spécifications de la figure 1.

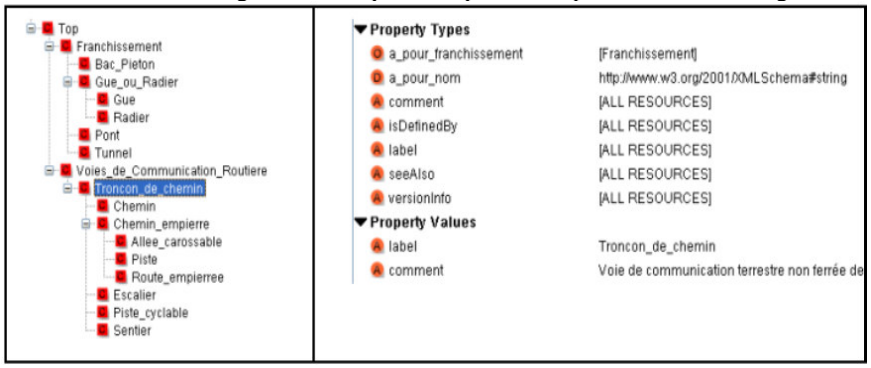


Figure 5. Extrait de l'ontologie obtenue avec GATE, conforme aux spécifications de la figure1.

Le concept *Tronçon de chemin* (<nom_classe>) est un fils du concept *Voies de communication routière* (<nom_domaine>) et a pour fils *Chemin*, *Escalier*, *Piste cyclable*, *Sentier* et *Chemin empierré* (<valeur>), qui à son tour a pour fils *Allée carrossable*, *Piste* et *Route empierrée* (<regroupement>). *Tronçon de chemin* est lié au type *String* par la relation *a-pour-nom* de type DataProperty, et au concept *Franchissement* par la relation *a-pour-franchissement* de type ObjectProperty.

5 Evaluation

Nous avons évalué l'apport de notre méthode sur ce type de corpus, par rapport aux approches connues. Au vu des spécifications, une approche statistique ne peut fournir de résultats significatifs (les textes étant très courts et très succincts, les concepts ne sont cités que peu de fois). De même, l'approche linguistique ne peut être satisfaisante car il y a relativement peu de texte rédigé (la plupart des expressions contenues entre les balises sont des noms de concepts). Par contre, une approche basée uniquement sur la structure visuelle du texte pour construire automatiquement des ontologies à partir de ces mêmes spécifications existe (Laurens, 2006). Nous avons donc comparé les deux méthodes à travers les ontologies obtenues.

5.1 Ontologie basée sur la structure visuelle

Cette étude (Laurens, 2006) se base sur un document HTML généré automatiquement à partir du document Word correspondant aux spécifications d'une base de données géographique, et exploite uniquement la structure visuelle du texte (style, caractères gras, soulignement, encadrement). Le document HTML est converti au format XML

en imbriquant les éléments du texte selon leur structure visuelle. Parallèlement, les groupes nominaux présents dans le texte sont extraits par des transducteurs et soumis à un expert pour validation comme termes du domaine : les termes retenus sont projetés sur le document XML. Le résultat, une taxonomie de concepts, correspond alors à la hiérarchie des termes associés dans le document XML.

5.2 Comparaison des deux ontologies obtenues

Les deux ontologies Onto_SV et Onto_ST obtenues respectivement par l'approche de F. Laurens et notre approche sont issues des mêmes spécifications (base BDTopo). Dans les deux cas, la mise au point des outils d'analyse du document s'appuie sur une interprétation approfondie de la sémantique des balises et de leur imbrication. La construction de Onto_SV a nécessité une intervention humaine à différents stades de sa conception (sélection/validation des expressions géographiques, nettoyage manuel de la hiérarchie XML obtenue, nettoyage/réorganisation manuel de la taxonomie OWL), alors que l'ontologue n'intervient dans Onto_ST que pour corriger, une fois l'ontologie construite, les incohérences dues à des erreurs de spécification (voir §5.3). Nous donnons figure 6 un état comparatif de ces deux ontologies.

	Onto_SV	Onto_ST
Nombre de concepts	615	1251
Profondeur	6	6
Relation hiérarchique "est-un"	oui	oui
Propriétés	non	oui
Relation de méronymie	non	oui
Relations conceptuelles autres	non	oui
Mode de construction	Supervisé	Non supervisé

Figure 6 : tableau comparatif des deux ontologies

Onto_ST est construite de manière automatique et est plus riche que Onto_SV en termes de concepts (notre méthode différencie les concepts portant un même label, lorsque ceux-ci se différencient par leurs propriétés) et de relations (relations autres que hiérarchiques).
Onto_ST n'est certainement pas la meilleure ontologie du domaine que l'on puisse obtenir, mais c'est la plus proche des spécifications.

5.3 Limites et intérêts de notre approche

La qualité de l'ontologie obtenue dépend entièrement de la qualité des spécifications : lorsque des incohérences existent au niveau des spécifications, une intervention humaine s'impose pour corriger l'ontologie. Et c'est là un des intérêts de la formalisation : aider à repérer des informations trop peu précises ou des incohérences au sein de documents a priori très structurés comme des spécifications. Or les variations de sens (au niveau du lexique, de la structure ou de la mise en forme) sont une des caractéristiques des textes. L'analyse du document a soulevé des cas pour

lesquels soit la relation identifiée n'avait pas la sémantique attendue, soit un des éléments d'une énumération avait un statut différent des autres, etc. Nous pointons ici quelques unes de ces anomalies.

5.3.1 Problèmes au niveau de la hiérarchie des concepts

Prenons pour exemple l'attribut "Autre classement" qui qualifie "route" :

Classement = « Autre classement »

Définition : Route qui ne fait partie ni du réseau autoroutier, ni du réseau national, ni du réseau départemental (voir ci-dessus).
Regroupement : Voies goudronnées (voies communales, chemins ruraux ou voies privées) | Rues | Rues piétonnes

Le champ *Regroupement* (qui désigne des concepts fils de ce type de route) met au même niveau *Rues* et *Rues piétonnes*, alors qu'il serait naturel de caractériser *Rues piétonnes* comme une spécialisation de *Rues*. De nombreux autres cas de ce type ont été rencontrés dans les énumérations, qui devraient ne pas être considérées comme contenant des concepts systématiquement de même niveau.

5.3.2 Incohérence au niveau des relations conceptuelles

Tronçon de route

Définition : Portion de voie de communication destinée aux automobiles, homogène pour l'ensemble des attributs et des relations qui la concernent. Représente uniquement la chaussée, délimitée par les bas-côtés ou les trottoirs.

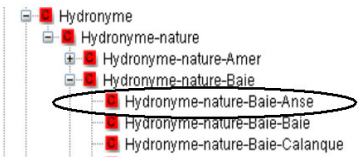
Géométrie : Linéaire

Attributs
- Identifiant (1)
- Source géométrique des données (1)
- Nature
- Classement
- Département gestionnaire
- Fictif

Le concept *Tronçon de Route* est défini à partir d'une <classe> du <domaine> *Voies de communication routière.* La règle d'interprétation des balises conduit à définir deux concepts liés par la relation hiérarchique "est-un". Or un patron de méronymie projeté sur le champ *définition* montre que *Tronçon de route* est une *partie-de voie de communication*. Ce genre d'incohérence (deux relations hiérarchiques entre les mêmes concepts) ne peut être levé que par intervention humaine.

5.3.3 Présence d'un même concept à différents niveaux de la hiérarchie

Les spécifications font qu'un même terme peut se retrouver à différents niveaux.



On retrouve ainsi le terme *Anse* comme désignant un concept fils de *Tronçon de laisse* et comme un concept fils de *Baie*, lui-même concept fils de *Hydronyme*. Une première solution est de créer un seul concept *Anse* et de lui associer plusieurs concepts pères. Or les spécifications donnent des définitions et des propriétés différentes dans chaque cas. Pour respecter la structure du document, nous avons choisi de concaténer le nom du concept courant à celui de ses concepts pères. Ceci permet de différencier les deux concepts *Anse*, et de fournir par ailleurs une traçabilité de l'ontologie vers le document de spécification qui a servi à la construire.

6 Conclusion et perspectives

Nous avons montré que, dans le cas favorable où des textes sont structurés à l'aide de balises dont la sémantique est claire, et dont la hiérarchisation porte aussi une sémantique précise, il est possible de définir une chaîne de traitements efficace pour construire automatiquement une ontologie. Ces traitements s'appuient sur des règles exploitant à la fois la structure des documents, le texte en langage naturel et en partie la mise en forme matérielle. Ils étendent donc les informations habituellement exploitées pour l'extraction de relations à partir de texte. L'ontologie ainsi obtenue à partir de textes s'avère riche en concepts et relations, et un lien précis est assuré entre éléments d'ontologie et textes. Nous sommes conscients que l'ontologie contient des incohérences qu'il faudra corriger manuellement. Dans le cadre de GEONTO, ces ajustements se feront après une étape d'alignement entre les différentes ontologies obtenues. Notre approche a été implémentée avec la plate-forme GATE.
L'évolution majeure envisagée pour notre méthode sera d'en assurer la portabilité à tout document de spécification de bases de données. Pour cela, nous pensons paramétrer les règles génériques en fonction des types de balises. Pour le moment, notre objectif sera d'être capable d'analyser tous les documents de spécification à traiter dans le projet GEONTO. Par ailleurs, nous comptons enrichir l'ontologie obtenue, d'une part en analysant plus systématiquement la mise en forme matérielle du texte (interprétation d'autres types d'énumération (Luc, 2001), des parenthèses, etc.), et d'autre part à l'aide de concepts, de relations ou de termes provenant de ressources externes. Dans le cas de GEONTO, il s'agira de concepts, relations et termes tirés de textes grand public, étude réalisée par le LIUPPA.

Références

ASHER N., BUSQUET J. ET VIEU L. (2001), La SDRT: une approche de la cohérence du discours dans la tradition de la sémantique dynamique. *Verbum* 23, 73-101.

AUGER A., BARRIERE C. (2008), Pattern based approaches to semantic relation extraction : a state-of-the-art. *Terminology,* John Benjamins*, 14-1,1-19.*

AUSSENAC-GILLES N., SEGUELA P. (2000), Les relations sémantiques : du linguistique au formel. *Cahiers de grammaire*. Numéro spécial linguistique de corpus. A. Condamines (Ed.). Toulouse : Presse de l'UTM. 25, 175-198.

AUSSENAC-GILLES N., DESPRES S., SZULMAN S. (2008), The TERMINAE Method and Platform for Ontology Engineering from texts. *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*. P. Buitelaar, P. Cimiano (Eds.), IOS Press, p. 199-223.

BARRIÈRE C., AGBADO A. (2006), TerminoWeb : a software environment for term study in rich contexts. *International Conference on Terminology, Standardization and Technology Transfert (TSTT 2006), Beijing (China)*, p. 103-113.

BOURIGAULT D. (2002), UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *TALN 2002,* Nancy*, 24-27* juin *2002*

BUITELAAR P., CIMIANO P., MAGNINI B. (2005), Ontology Learning From Text: Methods, Evaluation and Applications. *IOS Press.*

CHAROLLES M. (1997), L'encadrement du discours : Univers, Champs, Domaines et Espaces. *Cahier de Recherche Linguistique, LANDISCO, URA-CNRS 1035, Univ. Nancy 2, n°6, 1-73.*

GARDARIN G., BEDINI I., NGUYEN B. (2008), B2B Automatic Taxonomy Construction, ICES (3-2) 2008 : 325-330

GIULIANO C., LAVELLI A., ROMANO L. (2006), Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. *In Proc. EACL 2006*.

GRCAR M., KLEIN E., NOVAK B. (2007) Using Term-Matching Algorithms for the Annotation of Geo-services. Post-proceedings of the ECML-PKDD 2007 Workshops, Springer, Berlin – Heidelberg – New York.*Boston, MA : Kluwer Academic Plubisher* .

GREFENSTETTE G. (1994), Explorations in Automatic Thesaurus Discovery. *Boston, MA : Kluwer Academic Plubisher.*

HEARST M.A. (1992), Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th conference on Computational linguistics*, Morristown, NJ, USA, ACL, 539–545.

HINDLE D. (1990), Noun classification from predicate argument structures. *In Actes, 28th Annual Meeting of the Association for Computational Linguistics (ACL'90), Berkeley USA.*

JACQUEMIN C. (1997), Présentation des travaux en analyse automatique pour la reconnaissance et l'acquisition terminologique. *In Séminaire du LIPN,* Université Paris 13, Villetaneuse.

LAURENS F. (2006), Construction d'une Ontologie à partir de Textes en Langage Naturel. *Rapport de Stage Master 1 en linguistique-Informatique, Septembre 2006*

LUC C. (2001), Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte*. TALN2001*, *Université de Tours*, juillet 2001, p. 263-272.

MAEDCHE A. (2002), *Ontology learning for the Semantic Web*, vol 665. Kluwer Academic Pub.

NÉDELLEC C., NAZARENKO A. (2003). Ontology and Information Extraction. in S. Staab & R. Studer (eds.) *Handbook on Ontologies in Information Systems*, Springer.

REBEYROLLE J., TANGUY L. (2000), Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, 25, 153-174

TIRMIZI S., SEQUEDA S., MIRANKER J.F (2008), Translating SQL Applications to the Semantic Web. Dexa 2008, Turin , Italie, 450-464

VIRBEL J., LUC C. (2001), Le modèle d'architecture textuelle : fondements et expérimentation. *Verbum, Vol. XXIII, N. 1, p. 103-123.*

WEISSENBACHER D., NAZARENKO A. (2007), Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. *TALN 2007*, Toulouse, Juin 2007.