# Utiliser la Structure du Document dans le Processus de Construction d'Ontologies

Mouna Kamel[1], Nathalie Aussenac-Gilles[1]

[1] Laboratoire IRIT, Université Paul Sabatier, Toulouse
{kamel,aussenac}@irit.fr

**Résumé** : Les méthodes classiques de construction d'ontologies à partir de textes exploitent le texte rédigé qu'ils contiennent. Nous étendons ces approches en y ajoutant l'analyse de la structure du texte, qui contribue à en caractériser la sémantique. Lorsque le document existe sous un format numérique de type SGML, les différents éléments de la structure deviennent facilement accessibles. Notre méthode de construction d'ontologie se déroule alors en deux étapes. Une première phase s'appuie sur le repérage d'éléments de structure tels que titres et énumérations pour fournir un premier noyau d'ontologie. Nous montrons que ces éléments sont pertinents car ils dénotent des relations d'ordre ontologique. La seconde phase consiste à enrichir ce noyau en analysant le texte rédigé, plus précisément les définitions, selon des techniques classiques de TAL.

**Abstract**: Most methods for ontology learning from text exploit the natural language they contain. We broaden these approaches by analysing the text structure which also conveys semantics. When documents exist in a SGML-like digital form, it becomes easy to get most of their structural features. Our ontology learning method follows two steps. In a first step, it identifies structural features such as titles or enumerations in order to produce a first ontology kernel. We show that such features are particularly relevant since they indicate ontological relations. In a second step, the text natural language, specifically definitions, are processed to enrich this ontology kernel.

**Mots-clés** : Construction d'ontologies, structure du document, patrons structurels, patrons lexico-syntaxiques, Ingénierie des connaissances.

# Utiliser la Structure du Document dans le Processus de Construction d'Ontologies

Mouna Kamel[1], Nathalie Aussenac-Gilles[1]

[1] Laboratoire IRIT, Université Paul Sabatier, Toulouse
{kamel,aussenac}@irit.fr

**Résumé** : Les méthodes classiques de construction d'ontologies à partir de textes exploitent le texte rédigé qu'ils contiennent. Nous étendons ces approches en y ajoutant l'analyse de la structure du texte, qui contribue à en caractériser la sémantique. Lorsque le document existe sous un format numérique de type SGML, les différents éléments de la structure deviennent facilement accessibles. Notre méthode de construction d'ontologie se déroule alors en deux étapes. Une première phase s'appuie sur le repérage d'éléments de structure tels que titres, énumérations et définitions pour fournir un premier noyau d'ontologie. Nous montrons que ces éléments sont pertinents car ils dénotent des relations d'ordre ontologique. La seconde phase consiste à enrichir ce noyau en analysant le texte rédigé selon des techniques classiques de TAL.

**Mots-clés** : Construction d'ontologies, structure du document, patrons structurels, patrons lexico-syntaxiques, Ingénierie des connaissances.

## 1 Introduction

Un texte est une production linguistique composée d'une suite de signes en une langue donnée (le texte proprement dit) mais aussi d'un dispositif de structuration de mise en page, de typographie (organisation dispositionnelle). Les méthodes de construction d'ontologies à partir de textes privilégient souvent l'analyse du texte brut, que ce soit selon une approche statistique ou linguistique (Buitelaar et al., 2005), (Aussenac et al., 2008). La plupart de ces travaux intègrent différents outils de TAL et soulignent la complémentarité entre identification de concepts et extraction de relations. Mais lorsque l'auteur du document utilise les éléments de structure (moyens de mise en forme) pour organiser, subdiviser et hiérarchiser le contenu de son document, la matérialité du texte joue un rôle important, car elle est également porteuse de sémantique : selon le Modèle d'Architecture Textuelle (MAT), toute mise en forme matérielle possède une formulation discursive (Luc & Virbel, 2001) et contribue à la sémantique du document.

Partant de ce constat, l'idée que nous défendons ici est que la structure hiérarchique du document est porteuse de sémantique, et que son exploitation contribue à améliorer le processus de construction d'ontologie à partir de texte. En effet, les éléments de structure traduisent des relations hiérarchiques entre unités textuelles, et leur analyse

permet souvent d'obtenir un premier noyau d'ontologie. Ce noyau est ensuite enrichi par l'analyse linguistique du texte. Notre approche consiste alors à combiner ces deux analyses (structurelle et linguistique) pour identifier les concepts et les relations sémantiques, ces dernières pouvant être exprimées de façon implicite (par le biais des éléments de structure) ou explicite (en langage naturel). Pour cela, nous définissons d'une part des patrons structurels qui prennent en compte la structure hiérarchique du document, et d'autre part des patrons lexico-syntaxiques qui prennent en compte les éléments lexicaux, syntaxiques et sémantiques du texte brut. L'originalité de ce travail réside dans la définition de patrons structurels : un patron structurel permet d'associer une propriété sémantique à un élément de structure. Le choix de cette propriété provient du contexte dans lequel chaque élément de structure apparaît. Néanmoins, il existe des classes de textes pour lesquelles on peut associer à un élément de structure particulier une sémantique précise. La mise en évidence de telles associations requiert l'analyse des textes par un expert.

Cet article est organisé de la façon suivante. A l'aide d'un exemple en partie 2, nous montrons comment la mise en forme d'un document textuel, et plus particulièrement les éléments de structure, sont porteurs de sémantique. Nous rappelons ensuite les différentes méthodes de construction d'ontologies à partir du contenu d'un texte ou de sa structure (partie 3). La partie 4 décrit notre approche qui combine l'analyse de la structure du document à l'analyse linguistique du contenu textuel, pour améliorer le processus de construction d'ontologies à partir de textes. La partie 5 présente les résultats obtenus en appliquant notre méthode sur des textes très structurés contenant des spécifications de bases de données (projet GEONTO). Nous dressons enfin le bilan actuel de nos travaux au sein du projet et présentons les perspectives pour les améliorer (partie 6).

## 2 Structure de document et sémantique

Le document présenté à la figure 1 montre que la mise en forme matérielle du texte est porteuse de sémantique : différents éléments de structure (titraille, énumération, liste, etc.) et éléments typographiques (taille des caractères, police, encadré, etc.) sont utilisés pour souligner l'importance de certains éléments du texte et les liens qui existent entre eux. Les éléments de structure sont particulièrement pertinents pour la construction d'ontologie dans la mesure où ils traduisent des relations hiérarchiques entre différentes unités textuelles. Les éléments de structure de la figure 1 notifient au lecteur de façon assez intuitive (bien qu'implicite) qu'un *Tronçon de route* est une *Voie de communication routière*, les différents éléments de l'énumération (*Autoroute*, *Départementale*, *Nationale*, etc.) sont des catégories de *Tronçon de route*, chaque élément de la liste (*Identifiant, Classement,* etc.) représente un attribut de *Tronçon de route*, etc.
Notre approche se base essentiellement sur l'analyse des titrailles, listes et énumérations car elles ont fait l'objet de nombreux travaux et sont fréquentes en corpus. Les structures énumératives sont composées d'une amorce (syntagme nominal

introducteur de l'énumération) et d'un ensemble d'items (éléments de l'énumération) (Luc, 2001). Lorsque les items ont la même structure discursive, une relation sémantique entre l'amorce et chacun des items peut être établie. (Jacquemin & Bush, 2000) sont les premiers à avoir utilisé la structure visuelle des énumérations pour une application de recherche d'Entités Nommées. L'emboîtement ou le parallélisme de titres de sous-sections d'une section donnée reflètent des relations de subordination ou de juxtaposition existant entre ces sections (Jacques, 2005). Là encore, lorsque les sous-titres ont la même structure discursive, une relation sémantique entre le titre et chacun des sous-titres, ou entre les sous-titres, peut être instaurée.

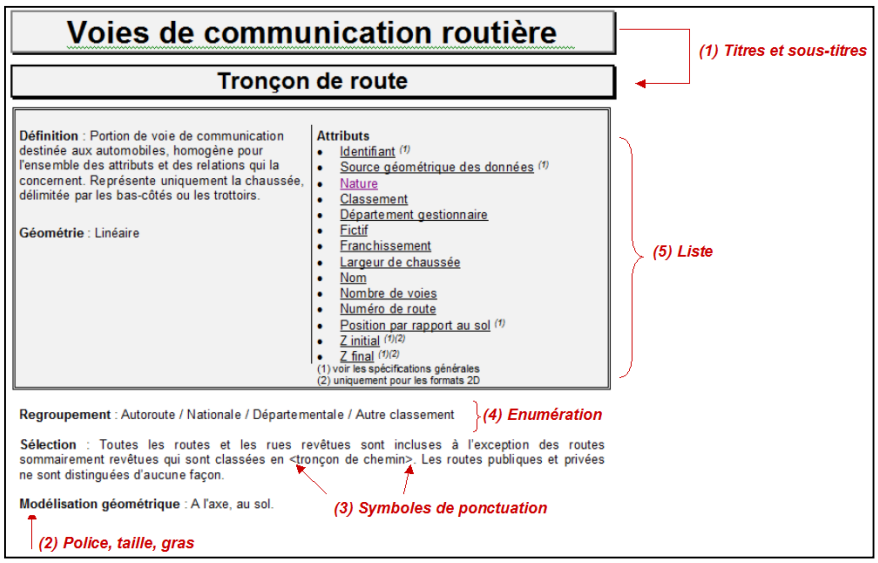


Figure 1. Exemple de document structuré

Le traitement automatique de la structure d'un document nécessite que les contraintes suivantes soient vérifiées :

1. Le document doit être accessible dans un format qui rende explicite sa structure et son caractère hiérarchique (format de type SGML).
2. Le document doit être analysé par un expert pour caractériser les manifestations matérielles des éléments de structure pertinents, et les propriétés sémantiques de ces manifestations (qui sont propres au document).
3. Les termes présents dans les structures discursives des éléments de structure doivent être des termes désignant des concepts.

La complémentarité entre identification de concepts et de relations peut être alors assurée en exploitant simultanément la sémantique véhiculée par l'intitulé des balises, la structure hiérarchique du document et l'information contenue entre les balises.

## 3 Construction d'ontologies à partir de textes

L'extraction et l'identification de relations est le complément indispensable de l'identification des concepts pour construire ou enrichir une ontologie. Deux familles de techniques extraient des relations sémantiques à partir de textes : les approches statistiques et les approches linguistiques. *Les approches statistiques* consistent, entre autres, à analyser les termes co-occurrents et la similarité entre leurs contextes syntaxiques (Hindle, 1990), (Grefenstette, 1994), à découvrir des indices de présence de relation à l'aide de techniques telles que les réseaux bayésiens (Weissenbacher & Nazarenko, 2007), les algorithmes d'apprentissage (Giuliano et al., 2006). Ces méthodes sont efficaces, mais identifient mal la sémantique de la relation. *Les approches linguistiques* font appel à des analyses syntaxiques ou des calculs de dépendance pour identifier les relations argumentatives (sujet, verbe, objet) (Jacquemin, 1997), (Bourigault, 2002), ou définissent des patrons lexico-syntaxiques pour reconnaître les marques linguistiques des relations sémantiques (Aussenac-Gilles & Seguela, 2000). De ce fait, la sémantique des relations est bien identifiée, mais la variabilité de leur expression en corpus oblige à multiplier les patrons et rend souvent l'approche coûteuse.
Ces techniques ne s'appliquent que pour la reconnaissance des relations intraphrastiques, alors que d'autres études ont pour objectif d'identifier des relations interphrastiques. Les liens entre différents paragraphes du texte peuvent exprimés soit par des relations du discours repérables à l'aide de marqueurs linguistiques (Charolles, 1997) (Asher et al., 2001), soit par la structure matérielle du texte (Luc & Virbel, 2001). La structure du texte joue un rôle important car elle donne "forme et sens au contenu" (Jacques, 2005) et, du fait de la numérisation des documents, son contenu devient plus facilement accessible grâce à des formats de représentation de la structure. Pourtant, à notre connaissance, très peu de travaux la mettent à profit pour construire des ressources sémantiques. Citons cependant la construction de taxonomies à partir de la structure visuelle du texte (style, caractères gras, soulignement, encadrement) (Abadie & Mustière, 2008), et l'analyse d'objets textuels tels que l'énumération qui met en évidence le rôle des propriétés visuelles dans l'acquisition de relations sémantiques (Virbel et al., 2005).

## 4 Analyse du document

Notre approche associe une analyse de la structure du document pour obtenir un premier noyau d'ontologie, à une analyse linguistique du texte pour enrichir ce noyau.

### 4.1 Analyse de la structure

L'imbrication des composants des éléments de structure traduit des relations d'ordre ontologique entre ces composants. Par ailleurs, lorsque le document est accessible

```
<Document>
 <Domaine>
  <nomDomaine> Voies de communication. routière</nomDomaine>
  <Classe>
   <nomClasse> Tronçon de route <nomClasse>
   <nomClasse> Tronçon de chemin </nomClasse>
  </Classe>
 </Domaine>
</Document>
```

Figure 2.a. Document XML

**↓ Ré-annotation des balises**

```
<Tag name="Document" path="/">
 < Tag name=″Domaine″ path=″/Document″>
  < Tag name=″nomDomaine″ path=″/Document/Domaine″>
                                    Voies de communication. routière</Tag>
  <Tag name=″Classe″ path=″/Document/Domaine″>
   <Tag name=″nomClasse″ path=″/Document/Domaine/Classe″> Tronçon de route </Tag>
   <Tag name=″nomClasse″ path=″/Document/Domaine/Classe″> Tronçon de chemin </Tag>
  </Tag>
 </Tag>
</Tag>
```

Figure 2.b. Document XML ayant les balises ré-annotées

**↓ Annotation des termes et lexèmes**

```
<Tag name="Document" path="/">
 <Tag name=″Domaine″ path=″/Document″>
  <Tag name=″nomDomaine″ path=″/Document/Domaine″>
      <Terme> <Token> Voies </Token> <Token> de </Token>
              <Token> communication </Token>. <Token> Routière </Token>  </Terme>
  </Tag>
  <Tag name=″Classe″ path=″/Document/Domaine″>
   <Tag name=″nomClasse″ path=″/Document/Domaine/Classe″>
      <Terme> <Token> Tronçon </Token>  <Token> de </Token> <Token> route </Token>
      </Terme>
   </Tag>
   <Tag name=″nomClasse″ path=″/Document/Domaine/Classe″ >
    <Terme> <Token> Tronçon </Token> <Token> de </Token> <Token> chemin </Token>
    </Terme>
   </Tag>
  </Tag>  </Tag>
</Tag>
```

Figure 2.c  Document XML avec les balises, les termes et les lexèmes annotés

Figure 2. Processus de ré-annotation et d'annotation d'un document XML

dans un format de type SGML, les composants principaux et subordonnés sont marqués par des balises spécifiques. L'analyse de la structure consiste alors à :

1. associer une sémantique aux balises qui marquent les composants des éléments de structure et aux relations qui les lient, en fonction de connaissances d'arrière-plan
2. ré-annoter et annoter automatiquement le document pour mettre en évidence d'une part les propriétés des balises (nom, localisation, attributs), et d'autre part pour attribuer des catégories grammaticales ou sémantiques aux unités textuelles présentes dans le texte à l'aide d'outils du TAL.
3. définir des patrons structurels qui ont pour rôle d'expliciter la sémantique véhiculée par ces éléments de structure.
4. projeter ces patrons sur le document pour obtenir des fragments d'ontologies.

### 4.1.1 Ré-annotation et annotation de document

La ré-annotation d'un document consiste à expliciter les caractéristiques des balises (nom et localisation hiérarchique de la balise au sein du document). Chaque balise est traduite en une balise nommée *Tag* ayant pour attributs *name* (nom initial de la balise) et *path* (localisation de la balise). Les attributs présents dans la balise initiale sont conservés tels quels. L'annotation sert à caractériser les lexèmes et termes (annotés respectivement *Token* et *Term*) à l'aide d'un tokenizer et d'un extracteur de termes.
La figure 2 décrit ce processus appliqué à un document XML.

### 4.1.2 Définition des patrons structurels

Un patron structurel (PS) a pour rôle de caractériser la sémantique portée par un élément de structure et de produire un fragment d'ontologie. L'identification des balises marquant les composants de l'élément de structure et de la relation qui les lie reste à la charge de l'expert. Une liste de PS doit être définie pour chaque type de document, en fonction du jeu de balises et de leur sémantique.
Les PS sont définis à l'aide du langage JAPE (Java Annotation Patterns Engine), langage d'expression de règles qui portent sur les annotations (et non sur le texte). La partie gauche d'une règle Jape est composée d'un motif basé sur les annotations, la partie droite d'un ensemble d'instructions qui manipulent les annotationsou des éléments d'ontologie. Les instructions sont exprimées en langage Java.
Le PS de la figure 3 permet d'exploiter le document de la figure 2.

```
({Tag name="nomDomaine" path="/Document/Domaine"} ({Terme}) : T1
({Token})*
({Tag name="nomClasse" path="/Document/Domaine/Classe"} ({Terme}) : T2
{Tag.name="nomDomaine",Tag.path="/Docment/Domaine", Tag contains
                {Tag.name="nomClasse", Tag.path="/Document/Domaine/Classe" } }
→ O = { est-un (T1, T2)}
```

Figure 3. Un exemple de Patron Structurel

Ce patron est appliqué lorsque :

(1) un terme annoté T1 par le PS (resp. T2) est marqué par une balise *Tag* ayant pour attributs <*name="nomDomaine"*> et <*path="/Document/Domaine"*> (resp. <*name="nomClasse"*> et <*path="/Document/Domaine/Classe"*>).
(2) Il peut exister 0 ou plusieurs lexèmes entre T1 et T2
(3) La balise marquant T2 est sous la portée de la balise marquant T1

Pour chaque appariement de la partie gauche du PS, un fragment d'ontologie O est construit (instructions en code Java) : deux concepts ayant respectivement pour termes T1 et T2 sont reliés par la relation *est-un*.

### 4.1.3 Construction du noyau d'ontologie

Les fragments d'ontologie produits par les PS sont fusionnés pour former un noyau d'ontologie. Dans le cas de l'exemple de la figure 2, deux fragments d'ontologies sont produits par le PS de la figure 3, et fusionnées (figure 4).

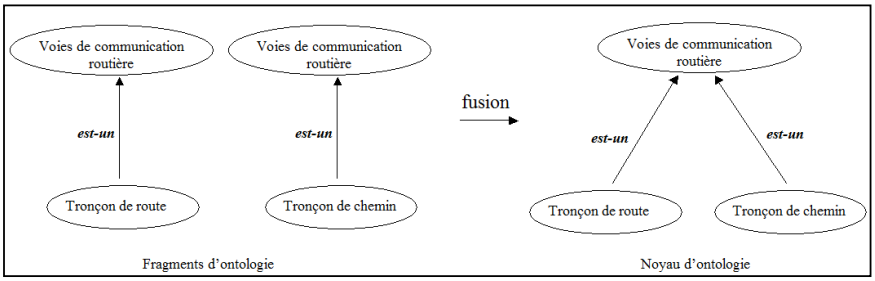


Figure 4. Fragments d'ontologie produits par le PS de la figure 3.

L'analyse linguistique du texte consiste alors à enrichir ce noyau en découvrant de nouveaux concepts, relations, termes associés aux concepts, propriétés, etc.

## 4.2 Analyse du texte rédigé

Les définitions sont "un lieu privilégié où s'actualisent les relations sémantiques que les mots entretiennent entre eux" (Rebeyrolle & Tanguy, 2000). Nous avons choisi d'exploiter les définitions directes (par opposition aux définitions indirectes, qui ont une structure discursive) à l'aide de patrons incluant des critères lexicaux, syntaxiques et typo-dispositionnels (Auger & Barrière, 2008). Ces patrons exploitent les étiquettes morpho-syntaxiques ou sémantiques attribuées par différents outils de TAL (tokenizer, lemmatiseur, analyseur syntaxique, etc.), pour caractériser les relations sémantiques existant entre les éléments de la structure. Seules les définitions conformes au schéma suivant ont été étudiées : "*Terme défini : énoncé définitoire*", où l'énoncé définitoire représente une seule caractérisation du terme défini.

Une étude des énoncés définitoires sur un corpus spécifique nous a conduit à définir les patrons lexico-syntaxiques suivants. Chacun d'eux correspond à un cas de définition de concept. Ces patrons exploitent les étiquettes morpho-syntaxiques *Term*,

*Concept*, *Property* (caractérisée grammaticalement par un adjectif ou un complément du nom) *et RMarker* (marqueur de relation lexicale). Ces étiquettes sont obtenues respectivement à partir d'un extracteur de termes, du noyau d'ontologie (partie 4.1), d'un étiqueteur grammatical, et de patrons lexico-syntaxiques définis par nos soins.

**{Concept} {Token.string=":"} {Term}**

Un terme est défini par un autre terme (cas de quasi synonymie). Ces deux termes (le défini et le définitoire) sont alors associés au même concept.

*Exemple* : [Cascade]$_{\text{Terme}}$ : [Chute d'eau] $_{\text{Terme}}$

**{Concept} {Token.string=":"} ({Property})* {Term} ({Property })***

Un terme est défini par un autre terme auquel sont associées des propriétés. Le terme 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* : [Terrain de sport] $_{\text{Terme}}$ : [Equipement sportif] $_{\text{Terme}}$ [de plein air] $_{\text{Propriété}}$

**{Concept} {Token.string=":"} {RMarker} {Term} ({Property })***

Un terme est défini par une relation lexicale le reliant à un autre terme. La relation lexicale correspondante est donc établie entre les concepts relatifs au terme défini et au terme définitoire.

*Exemple* : [Carburateur] $_{\text{Terme}}$ : [Partie de] $_{\text{Marqueur}}$ [moteur] $_{\text{Terme}}$

Si les termes *défini* et *définitoire* sont associés à des concepts présents dans l'ontologie, la relation sémantique entre ces concepts est établie. Pour les termes non présents, de nouveaux concepts sont créés. Pour les insérer dans l'ontologie, nous utilisons l'inclusion lexicale qui traduit souvent une relation hiérarchique (hyponymique ou méronymique) (Kleiber & Tamba, 1990). Lorsqu'aucune solution n'est trouvée, une interface de validation demandera à l'ontologue de placer ces nouveaux concepts au sein de l'ontologie.

## 5 Cadre d'application

Au sein du projet GEONTO[1], le COGIT[2] 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 avec une ontologie de référence. Plutôt que de construire une ontologie à partir du schéma de chacune des bases de données (Tirmizi et al., 2008), (Gardarin et al., 2008), elle est construite à partir d'un document de spécifications de cette base. En effet, ces documents sont intéressants à exploiter à plusieurs titres. Tout d'abord, ils sont sémantiquement riches : ils contiennent des descriptions de concepts, de relations, de contraintes, de définitions, etc. qui sont exprimées à la fois à travers leur

[1] Projet ANR-07-MDCO-005, http://www.lri.fr/geonto

[2] COGIT : Conception Objet et Généralisation de l'Information Topographique, laboratoire de l'IGN

mise en page (forte présence de titres, d'énumérations et de définitions), et à travers le langage naturel. Ensuite, ces documents sont fournis dans un format XML qui rend les éléments de la structure facilement accessibles. Enfin, ces documents sont conformes à un schéma XML inspiré de normes ISO pour les données géographiques.
Nous avons définis un ensemble de patrons structurels et lexico-syntaxiques qui s'appliquent à l'ensemble des documents de spécification disponibles au COGIT.

## 5.1 Construction de l'ontologie

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. Un extrait de ces spécifications au format XML est présenté figure 5.

```
<class name="Tronçon de chemin">
<className>Tronçon de chemin</className>
<description type="definition">Voie de communication terrestre non ferrée …</description>
<description type="extensionalDefinition"> <AttTermList>nature</AttTermList> </description>
<description type="selectionPrinciple">Voir les différentes valeurs de <Nature>… </description>
<attributes>
<attribute name="Nature">
<attributeName>Nature</attributeName>
<valueType>Énuméré</valueType>
<description type="definition">Permet de distinguer plusieurs types de voies de comm. … </description>
<enumeratedValues>
<value name="Sentier">
<valueName>Sentier</valueName>
<description type="definition">Chemin étroit ne permettant pas le passage de véhicules.</description>
<description type="extensionalDefinition">
<TermList>Allée piétonne (étroite)</TermList>
<TermList>Piste de cross</TermList>
<TermList>Ruelle étroite</TermList>
<TermList>Sentier</TermList>
</description>
<description type="selectionPrinciple">Seuls les principaux sentiers sont  inclus.</description>
 </value>
```

**Figure 5** : Extrait des spécifications de BDTopo au format XML

L'application de patrons structurels et lexico-syntaxiques que nous avons définis après une expertise du document a permis de construire une ontologie en deux temps : le noyau d'ontologie obtenu à partir de la structure, puis le noyau enrichi par l'analyse linguistique du texte. L'extrait de l'ontologie correspondant à l'extrait de la figure 5 est présenté figure 6.

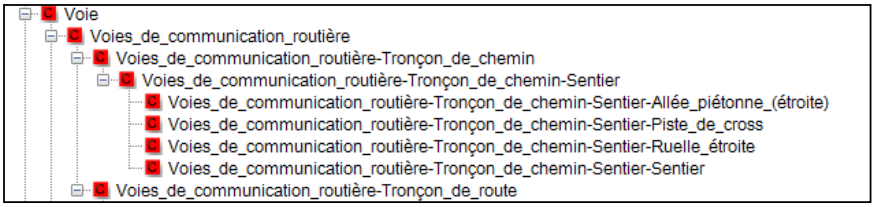


Figure 6. Extrait de l'ontologie relative à BDTopo

Comme un même terme peut se retrouver à différents niveaux dans le document et que les spécifications peuvent donner des définitions et des propriétés différentes dans chaque cas, nous avons choisi de concaténer le nom du concept courant à celui de ses concepts pères. Ainsi l'identifiant d'un concept précise sa position dans la hiérarchie du document.
La figure 7 montre les différentes propriétés associées au concept *Sentier*.

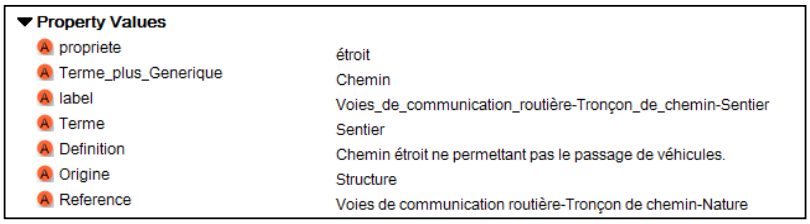


Figure 7. Propriétés du concept *Sentier*

L'analyse structurelle a permis d'établir les propriétés **label** (identifiant du concept), **Terme** (termes associés au concept), **Definition** (définition du concept), **Origine** (si le concept provient de l'analyse de la structure ou du texte) et **Reference** (référence à l'attribut dénotant une hiérarchisation). L'analyse linguistique permet d'enrichir l'ensemble de ces propriétés : les propriétés **Terme_plus_Generique** et **propriete** permettent respectivement de proposer *Chemin* comme terme plus générique que *Sentier* à l'ontologue, et d'associer *étroit* comme une propriété physique à *Sentier*.

Le tableau ci-dessous montre les caractéristiques obtenues à chaque étape.

| | Analyse de la Structure | Analyse du Langage |
|---|---|---|
| Nombre de concepts | 1183 | 53 |
| Profondeur | 6 | 8 |
| Concepts à proposer | 0 | 101 |
| Propriétés | 0 | 174 |
| Relation de méronymie | non | oui |
| Relations conceptuelles autres | 50 | 0 |
| Termes | 1183 | 142 |
| Mode de construction | Expertise puis Non supervisé | Non supervisé |

### 5.2 Avantages et limites de notre approche

Dans le contexte de GEONTO, notre approche a permis d'obtenir une ontologie riche dans le sens où de nombreuses connaissances présentes dans le document sont modélisées dans l'ontologie. Cette ontologie aide à la validation, à la traçabilité et à l'alignement car les concepts sont documentés par le biais de propriétés, au fur et à mesure de leur construction (paragraphe 5.1).
Mais la qualité de l'ontologie obtenue dépend entièrement de la qualité des spécifications : lorsque des incohérences existent au niveau des spécifications (par ex., un des éléments d'une énumération a un statut différent des autres), une intervention humaine s'impose pour corriger l'ontologie. L'étape de validation de l'ontologie permettra aussi de pointer les erreurs présentes dans les spécifications.

## 6 Conclusion

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 patrons exploitant à la fois la structure des documents et le texte en langage naturel. Ils étendent donc les informations habituellement exploitées pour l'extraction de relations à partir de texte. L'ontologie ainsi obtenue s'avère riche en concepts, relations, propriétés et termes. De plus, la documentation des concepts assure la traçabilité vers le texte dont est issue l'ontologie. Il est à noter que la qualité du document a des conséquences sur la qualité des résultats issus de la méthode présentée.
Deux évolutions majeures sont envisagées pour enrichir l'ontologie obtenue. La première amélioration consiste à s'appuyer sur les travaux existants pour identifier les structures linguistiques qui permettraient de traiter automatiquement les éléments de structure porteurs d'une relation paradigmatique. La seconde évolution concerne une analyse plus systématique du texte brut contenu dans les énoncés définitoires, pour identifier des relations du discours et des relations interphrastiques.

# 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.

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*.

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

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.*

JACQUEMIN C., BUSH C. (2000), Fouille du Web pour la collecte d'Entités Nommées. *In E. Wehrli (Ed.), TALN 2000, Lausanne.*.

JACQUES M-P. (2005), Structure matérielle et contenu sémantique du texte écrit. *Corela, Volume 3, Numéro 2.*

KLEIBER G., TAMBA I., (1990), L'hyperonymie revisitée : inclusion et hiérarchie. *Langages, 98 :7-32, juin.*

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, juillet 2001.

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

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., GARCIA-DEBANC C., BACCINO T., CARRIO L., DOMINGUEZ C., JACQUEMIN C., LUC C., MOJAHID M., PERY-WOODLEY M-P., SCHMIDS S. (2005). Approches cognitives de la spatialisation du langage. De la modélisation de structures spatiolinguistiques des textes à l'expérimentation psycholinguitique : le cas d'un objet textuel, l'énumération. *Agir dans l'espace*. Catherine Thinus-Blanc, Jean Bullier (Eds.), Editions de la Maison des sciences de l'homme, p.233-254, Cognitique.

WEISSENBACHER D., NAZARENKO A. (2007), Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. *TALN 2007*, Toulouse, Juin 2007, p47-56.