



Rapport de stage
Master 1 TAL, Dictionnaires, Corpus, Terminologie

Recherche et analyse de ressources terminologiques liées à la topographie

Mai – Juin 2008

Anne-Lyse MINARD

Encadrant IGN : Sébastien Mustière

Co-encadrant IGN : Nathalie Abadie

Tuteur universitaire : Antonio Balvet

Remerciement

Tout d'abord je remercie toutes les personnes avec qui j'ai travaillé dans le laboratoire COGIT de l'IGN pour la qualité de l'accueil et l'ambiance agréable.

Je remercie en particulier mon maître de stage Sébastien Mustière pour m'avoir permis de faire ce stage, pour avoir encadré mon stage et relu mon rapport. Les réunions « bilans, méthodes et perspectives » qu'il a organisé au cours de mon stage m'ont été profitables.

Je remercie également Nathalie Abadie, ma co-encadrante, pour m'avoir accompagné pour mes premiers pas en JAVA avec le logiciel Eclipse.

Je remercie Jean-François Gleyze et Adrien Paget pour m'avoir accueilli dans leur bureau, fait découvrir le monde de la recherche, et permis de passer des bons moments dans ce laboratoire.

Je remercie mes parents pour m'avoir toujours soutenu dans les choix que j'ai faits au cours de mes études.

Table des matières

Remerciement.....	2
Table des matières	3
Table des figures	4
Introduction	5
Introduction	5
Présentation de l'entreprise	7
1. Présentation du sujet.....	8
1.1. Pré-requis	8
1.1.1. Vocabulaire contrôlé	8
1.1.2. Ontologie.....	9
1.1.2.1. Contenu d'une ontologie	10
1.1.2.2. Ontologie d'objets géographiques.....	11
1.1.3. Science de l'information géographique.....	11
1.2. Contexte	12
1.2.1. Ontologie de l'IGN.....	12
1.2.2. Projet GEONTO	13
1.3. Objectifs du stage	14
2. Réalisation de l'état de l'art	15
2.1. Méthodes	15
2.1.1. Mode de recherche	15
2.1.2. Grille de renseignements	16
2.1.3. Outils	20
2.2. Résultats	21
2.2.1. Ontologies en français	21
2.2.2. Ontologies en anglais	26
2.2.3. Ontologies multilingues	29
2.2.4. Thesaurus	31
2.2.5. Bilan	32
3. D'un thesaurus à une ontologie.....	33
3.1. Thesaurus	33
3.2. Ontologie.....	34
3.3. Prétraitement du thesaurus	34
3.4. Structuration en XML	38
3.5. Création de l'ontologie en OWL.....	39
3.6. Résultat.....	40
Bilan personnel.....	41
Conclusion.....	42
Bibliographie.....	43

Table des figures

Figure 1 Visualisation de l'ontologie de l'IGN grâce à l'éditeur Protégé	13
Figure 2 Grille de renseignements type.....	17
Figure 3 Grille de renseignements d'une ontologie de l'Ordnance Survey	18
Figure 4 Grille de renseignements d'une ontologie appelée geo_swoogle.....	19
Figure 5 Extrait du tableau recensant les ontologies.....	20
Figure 6 Visualisation du graphe de l'ontologie <i>ville</i> avec les outils du projet Towntology ...	22
Figure 7 Visualisation des définitions et des relations du concept <i>ville</i> de l'ontologie du projet Towntology	22
Figure 8 Extrait de l'ontologie des objets urbains du projet Fodomust.....	24
Figure 9 Projet GIEA : présentation des données sur le sol.....	25
Figure 10 Visualisation de l'ontologie <i>hydrology</i> de l'Ordnance Survey avec Swoop.....	27
Figure 11 Visualisation de l'ontologie géographique du projet SUMO avec Swoop	28
Figure 12 Visualisation de l'ontologie WalkOnWeb sur Protégé	30
Figure 13 Extrait du thesaurus de l'université de Montréal.....	31
Figure 14 Description du terme <i>Cours d'eau</i> du thesaurus de l'université de Montréal	31
Figure 15 Extrait du thesaurus au format PDF.....	35
Figure 16 Extrait du thesaurus au format TXT.	35
Figure 17 Extrait du thesaurus obtenu après traitement.....	36
Figure 18 Extrait du fichier TXT de départ.....	37
Figure 19 Extrait de la figure 18 après traitement.....	38
Figure 20 Balisage XML de l'extrait de la figure 19.....	39
Figure 21 Extrait de l'ontologie obtenue.....	40

Introduction

Dans le cadre de ma formation universitaire j'ai effectué un stage en entreprise de deux mois qui se clôture par la rédaction de ce rapport. Il s'est déroulé au sein du laboratoire COGIT de l'IGN.

Les objectifs de ce stage de fin de première année de master sont l'élaboration d'un état de l'art des ontologies d'objets géographiques et l'écriture d'un programme en JAVA qui converti un thesaurus au format PDF en une ontologie au format OWL.

Une ontologie est une représentation du monde pour un domaine particulier, conçue comme un ensemble de concepts hiérarchisés. Les concepts peuvent être accompagnés de leurs définitions, de leurs interrelations, de leurs propriétés, etc. Un état de l'art des ontologies d'objets géographiques permet de répertorier toutes les ontologies décrivant des objets géographiques et ainsi d'avoir un aperçu de ce qui existe. Pour réaliser l'état de l'art il faut consulter les sites Web traitant de la géographie, de la géomatique, des sciences de l'information géographique, etc. L'état de l'art regroupe les ontologies d'objets géographiques trouvées sur le Web, leurs descriptifs et les descriptions des projets pour lesquelles elles ont été créées.

La deuxième partie du stage consiste en l'écriture d'un programme en java pour transformer automatiquement un thesaurus en une ontologie. L'avantage d'une ontologie par rapport à un thesaurus est qu'elle est écrite dans un langage XML, elle peut donc être facilement utilisée par une application informatique. Au contraire un thesaurus n'est pas structuré en XML, mais il est souvent en texte brut. Il permet de hiérarchiser une grande quantité de termes, mais ses informations sont peu accessibles par un programme informatique. Il paraît donc enrichissant de pouvoir récupérer les données d'un thesaurus et de les transformer pour qu'elles soient utilisables par une application informatique.

Ce travail me permettra de mettre en pratique les connaissances en terminologie que j'ai acquises au cours de ma formation. La mise en pratique passera par l'utilisation du logiciel d'édition d'ontologie Protégé. Je pourrai aussi me familiariser avec le langage JAVA, langage de programmation utilisé par tous les chercheurs du laboratoire. C'est un langage

orienté objet, c'est donc un complément au langage PERL que j'ai appris durant ma formation.

Les ontologies sont souvent utilisées pour la recherche d'information, domaine dans lequel je souhaiterais travailler à la fin de ma formation, ce stage est un premier aperçu pour moi des outils utilisés dans ce domaine.

Je commencerai ce rapport par une présentation de l'entreprise, ensuite je ferai un rappel des différents types de vocabulaires contrôlés et de ce que sont les Sciences de l'Information Géographique. Puis j'exposerai le contexte dans lequel s'inscrit mon stage et je présenterai les deux travaux que j'ai effectués. Je finirai par un bilan personnel du stage.

Présentation de l'entreprise

L'Institut Géographique National (IGN) est situé à Saint Mandé (94), il comporte quatre laboratoires de recherche dont le COGIT (Conception Objet et Généralisation de l'Information Géographique). Le laboratoire est dirigé par Anne Ruas. Vingt chercheurs y travaillent.

Le COGIT est un laboratoire de géomatique. C'est une discipline qui s'intéresse à la représentation numérique de l'espace et à son analyse. Les recherches en géomatique ont pour but d'améliorer la gestion des données spatiales (acquisition, stockage, traitement ou diffusion). Le laboratoire COGIT travaille sur les bases de données topographiques vectorielles. Ces bases de données « s'attachent à décrire la localisation et la nature des entités liées à l'activité humaine sur l'espace support »¹. Les problématiques du COGIT sont structurées en 5 actions de recherche :

- SISSI « traite de l'aide à l'expression de besoins et à la construction de ressources personnalisées (cartes, lots de données) en fonction des besoins et des ressources disponibles »,
- LUCIL « traite de l'analyse et de l'amélioration de légendes prédéfinies »,
- GIGA « traite de l'automatisation du processus de généralisation »,
- BDMUL « traite de l'intégration de données multisources et multi-résolutions. En particulier sont traités la description des spécifications des bases de données et la conception de méthodes d'appariements basées sur des connaissances imprécises »,
- XDOGS « traite de l'analyse de données topographiques dans les domaines des risques naturels et des phénomènes territoriaux liés aux réseaux et aux bâtiments ».

Mon stage s'inscrit dans l'action de recherche BDMUL, qui est dirigée par Sébastien Mustière.

¹ Cartographie des Activités de Recherche du Laboratoire Cogit 2000 – 2004.

1. Présentation du sujet

Pour travailler sur les ontologies d'objets géographiques il est important de définir les différents types de structuration de termes, en particulier l'ontologie, objet principal de mon stage. Je présente ensuite le domaine dans lequel s'inscrit mon stage : les Sciences de l'Information Géographique (SIG), ainsi que son contexte, c'est-à-dire l'ontologie de l'IGN attendant des améliorations et le projet GEONTO dans lequel s'inscrit mon travail. Et pour finir j'expose les objectifs de mon stage.

1.1. Pré-requis

1.1.1. Vocabulaire contrôlé

Un vocabulaire contrôlé est un ensemble de termes définis et reconnus par un groupe dans un domaine donné. Il est utilisé en particulier en recherche d'information pour indexer, analyser et rechercher l'information. Il permet aussi de désambiguïser les termes utilisés par les chercheurs d'un domaine.

Si plusieurs termes désignent un même concept, l'un d'entre eux est choisi comme *terme préféré*. Si les termes sont hiérarchisés grâce à des relations de spécialisation (c'est-à-dire qu'un concept général est relié à des concepts plus spécifiques) on parle de taxonomie.

Exemple² de relations de spécialisation :

Surface d'eau
 Rivière large
 Fleuve large
 Mare

Un thesaurus est un vocabulaire contrôlé dans lequel les termes sont reliés entre eux grâce à trois types de relations. On y trouve les relations de spécialisation (comme dans une taxonomie), les relations d'équivalence (par exemple entre des termes « quasi-synonymes ») et les relations d'association (par exemple entre des termes décrivant des sujets connexes).

² Exemple extrait de la taxonomie de l'IGN.

Exemple³ des relations entre les termes d'un thesaurus :

(TS = Terme Spécifique, TA = Terme Associé)

Cours d'eau

TS *Fleuve*

TA *Eau*

TA *Hydrographie*

TA *Lac*

1.1.2. Ontologie

Une ontologie est une représentation du monde, par rapport à un domaine, qui est conçue comme un ensemble de concepts avec leurs définitions, leurs interrelations, leurs propriétés, ...

Définition de Gruber (1993) :

« Une ontologie implique ou comprend une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts – e.g. entités, attributs, processus –, leurs définitions et leurs interrelations. On appelle cela une conceptualisation.

[...]

Une ontologie peut prendre différentes formes mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification.

[...]

Une ontologie est une spécification rendant partiellement compte d'une conceptualisation. »

On peut distinguer 3 types d'ontologies : de référence, de domaine et de tâche.

Une ontologie de référence décrit les concepts les plus généraux d'une ou de plusieurs disciplines, elle permet d'en fixer le vocabulaire. Les ontologies de référence les plus connues sont WordNet⁴, Dolce⁵ ou encore SUMO⁶.

³ Exemple extrait du thesaurus de l'université de Montréal portant sur les activités gouvernementales.

⁴ <http://wordnet.princeton.edu/>

⁵ <http://www.loa-cnr.it/DOLCE.html>

Une ontologie de domaine est spécifique à un domaine, elle présente les concepts et les termes utilisés par les experts d'une spécialité. Par exemple une ontologie décrivant les phénomènes physiques liés à l'environnement est une ontologie de domaine.

Une ontologie de tâche est une ontologie spécifique qui décrit les concepts utilisés pour parler, décrire, etc. une certaine activité. Elle peut présenter par exemple les termes utilisés pour interpréter des images satellites.

1.1.2.1. Contenu d'une ontologie

Une ontologie décrit des concepts et elle les hiérarchise. Ces concepts sont appelés classes. Ces classes possèdent des propriétés, qu'elles peuvent avoir héritées de leur classe supérieure. Elles sont soit nécessaires et suffisantes, soit nécessaires. Les propriétés sont parfois exprimées sous forme d'axiome, c'est-à-dire grâce à des relations logiques.

Une classe peut être reliée à une autre grâce à une relation, la principale est la relation de spécialisation (*est un*), il peut aussi y avoir des relations de méronymie (*est une partie de*), de composition (*est composé de*), etc.

Une classe peut posséder des individus (ou instances), c'est-à-dire des entités du monde. Pour appartenir à une classe, un individu doit vérifier les mêmes propriétés qu'elle. Certaines propriétés correspondent donc à des relations entre une classe et des individus.

Exemple :

Classe : *pays*

Propriétés nécessaires : *est un lieu, a un gouvernement, a au moins 1 ville*, etc.

Propriété nécessaire et suffisante : *a une capitale*

Individu : *Canada*

Dans cet exemple la classe *pays* est liée à la classe *capitale* par une relation de composition, un pays comporte une capitale. La classe *pays* possède plusieurs propriétés que l'individu doit vérifier, le Canada est bien un lieu, il possède une capitale, un gouvernement et plus d'une ville.

⁶ <http://www.ontologyportal.org/>

1.1.2.2. Ontologie d'objets géographiques

Les ontologies d'objets géographiques sur lesquelles je travaille sont des ontologies de domaine qui décrivent les concepts caractérisant l'espace ou les concepts des domaines géographiques.

Un objet géographique est un objet modélisant un phénomène du monde réel, notamment en décrivant un ou plusieurs lieux de la surface du globe terrestre. Un objet géographique peut être décrit par des données sémantiques (son nom, sa nature, etc.) et/ou par des données géométriques (latitude, longitude, etc.).

Exemple de concepts décrivant des objets géographiques : *faille, tronçon de route, cours d'eau, sommet, ...*

À partir de ces deux définitions, on peut dire qu'une ontologie d'objets géographiques est une représentation hiérarchisée des termes décrivant des objets géographiques.

1.1.3. Science de l'information géographique

Mon stage s'inscrit dans le domaine des SIG (Science de l'information géographique) et plus spécifiquement dans celui de la géomatique.

L'information géographique est l'ensemble de la description d'un objet et de sa position géographique. Les SIG regroupent les outils informatiques utilisés pour organiser des données spatialement référencées et ainsi produire des cartes et des plans. Leur rôle est de présenter l'environnement spatial de façon plus ou moins réaliste, grâce à des primitives graphiques (point, vecteur, ...). Ces primitives sont accompagnées d'informations telles que leur nature (voie ferrée, forêt, etc.).

Une partie des SIG est appelée géomatique, elle constitue l'ensemble des outils et méthodes qui permettent de représenter, d'analyser et d'intégrer des données géographiques.

1.2. Contexte

1.2.1. Ontologie de l'IGN

L'IGN possède plusieurs bases de données, dont la BDTOPO qui contient des données topographiques et la BDCARTO qui contient des données cartographiques. Elles sont toutes les deux accompagnées d'une spécification, c'est « l'expression des règles de passage des phénomènes du monde réel à des objets d'une base » [Abadie et Mustière 2008]. Grâce à ces spécifications deux taxonomies ont été formées, c'est-à-dire deux ensembles de termes hiérarchisés grâce à des relations de spécialisation, une taxonomie pour chaque base de données [Laurens 2006]. Le but était d'avoir une seule taxonomie décrivant les deux bases de données, un alignement des deux taxonomies a donc été fait.

Le but premier de disposer d'une ontologie géographique est de résoudre les problèmes liés à l'hétérogénéité sémantique des données géographiques. En effet les données géographiques sont nombreuses et diverses, elles reflètent chacune une conceptualisation du monde propre au domaine (par exemple topographie ou cartographie) et/ou au producteur. Une ontologie géographique peut permettre de comparer la nature des objets géographiques pour l'appariement de données (identification des correspondances entre des objets géographiques décrivant la même réalité mais provenant de données différentes). Elle peut permettre aussi qu'un utilisateur qui cherche un élément dans une base de données puisse exprimer sa requête en termes intuitifs, et bien d'autres applications.

Cette ontologie comporte 761 concepts hiérarchisés, grâce à la relation de spécialisation « is-a », sur 7 niveaux de profondeur. Les termes décrivant les concepts sont traduits en anglais, le langage de description reste le français. Tous les concepts sont regroupés en deux grandes classes : « entités topographiques artificielles » et « entités topographiques naturelles ». La figure 1 présente cette ontologie au travers du logiciel Protégé. A gauche de la fenêtre est présenté la hiérarchisation des concepts, avec un concept par ligne, à droite la hiérarchisation sous forme d'arbre.

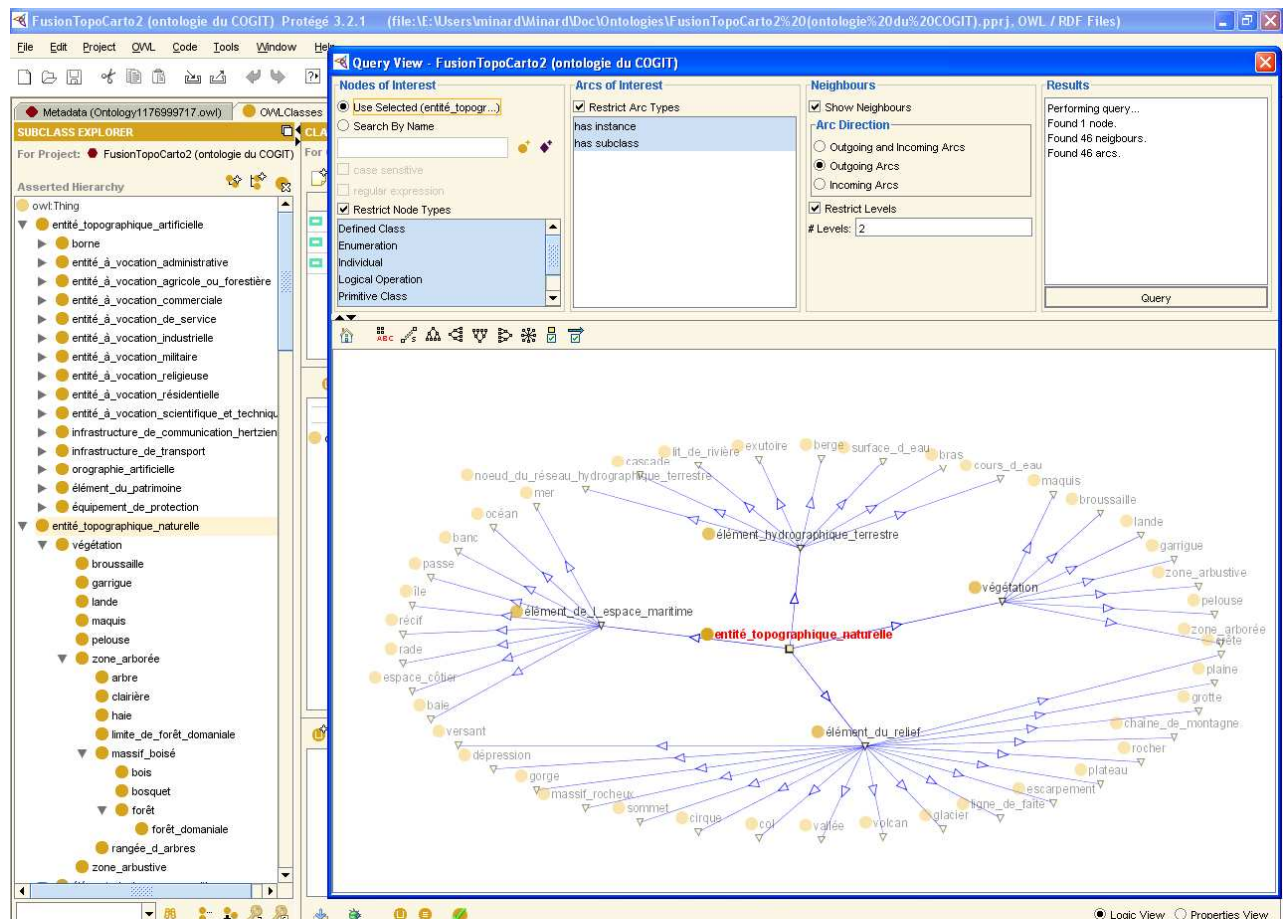


Figure 1 Visualisation de l'ontologie de l'IGN grâce à l'éditeur Protégé

1.2.2. Projet GEONTO

Ce projet lancé en 2007, a pour objectif la constitution, l'alignement, la comparaison et l'exploitation d'ontologies géographiques hétérogènes. Le projet porte sur l'interopérabilité⁷ de données diverses relatives à l'information géographique.

Objectifs généraux : - intégration de bases de données géographiques hétérogènes

- interrogation d'une collection importante de documents textuels plus variés et destinés à un plus large public.

Sous-objectifs : - construire des ontologies associées à des bases de données par exploitation de leurs spécifications ou associés à un corpus de documents géographiques

⁷ « Capacité qu'à un produit de fonctionner avec d'autres produits existants », définition extrait de l'article *interopérabilité* de Wikipédia.

- aligner les ontologies et étudier leurs différences
- apparier⁸ les schémas de bases de données via les ontologies et développer un moteur de recherche d'information dans une base de document via les ontologies.

Le rôle des ontologies en intégration de sources d'informations multiples et hétérogènes est de préciser le sens des concepts d'un domaine, de fournir une sémantique formelle, et d'aider à comprendre et interpréter des descriptions hétérogènes de contenus relatifs à un même domaine.

A l'heure actuelle des taxonomies de termes géographiques ont été créées à partir de récits de voyages et des spécifications des bases de données de l'IGN.

1.3. Objectifs du stage

L'objectif principal du stage est d'établir un état de l'art des ontologies d'objets géographiques. Il servira d'une part à mieux situer les ontologies produites par le projet GEONTO, et d'autre part à alimenter le projet GEONTO en ressources externes utiles à la constitution d'ontologies. En plus de répertorier les ontologies, j'ai été chargé de mettre en place une grille d'évaluation des ontologies. Elle prendra en compte le contenu mais aussi l'objectif de cette ontologie. Elle permettra de classer et regrouper les ontologies selon différents critères, ainsi que de relever des informations les concernant. A partir des grilles d'évaluation je proposerai un bilan sur les ontologies d'objets géographiques présentes sur le Web.

L'analyse des ontologies sera présentée dans l'état de l'art, elle permettra d'avoir un aperçu du contenu des ontologies, de la façon dont elles sont structurées, des relations entre les concepts, etc. et ainsi d'envisager des améliorations de l'ontologie de l'IGN.

Le deuxième objectif du stage est d'écrire un programme en Java permettant de convertir un thesaurus au format PDF en une ontologie au format OWL. L'ontologie produite comportera les données du thesaurus, structurées de manière à ce qu'une application informatique soit possible.

⁸ C'est-à-dire repérer que les schémas représentent les mêmes données, et ainsi pouvoir les faire correspondre.

2. Réalisation de l'état de l'art

Comme nous l'avons vu précédemment, le laboratoire COGIT possède une ontologie des concepts géographiques présents dans les spécifications de la BDTOPPO et de la BDCARTO. Celle-ci est restreinte et ne propose qu'une hiérarchisation de type spécialisation des objets géographiques. L'objectif du projet GEONTO est d'élaborer une ontologie plus complète, avec entre autre plus d'objets, leurs propriétés, des définitions, ...

La mission, qui m'a été confiée pendant le stage, était de chercher s'il existait des ontologies permettant de mieux situer les ontologies issues du projet GEONTO et de les alimenter en ressources externes. Pour montrer le résultat de mes recherches j'ai élaboré un état de l'art des principales ontologies que j'ai pu trouver.

2.1. Méthodes

L'objectif étant de parcourir le maximum de projets au cours duquel une ontologie d'objets géographiques ait été construite, il fallait réfléchir à une méthode pour ne pas perdre de temps. Dans les deux points qui suivent, je vais expliquer le mode de recherche que j'ai utilisé, ainsi que la façon dont j'ai répertorié les informations collectées pour chaque ontologie. Lorsque toutes les ontologies ont été répertoriées j'ai recherché des thesaurus.

2.1.1. Mode de recherche

Mes recherches ce sont axées sur le Web. Je n'ai pas pu l'explorer entièrement par manque de temps. Je suis allée chercher les ontologies proposées sur le Web en format numérique. Il est, en effet, très difficile de rechercher des ontologies et des projets d'ontologies au format papier, et je pense qu'il n'en existe pas beaucoup : les ontologies étant créées pour servir à des applications informatiques on peut imaginer que la majorité soit présentée sur le Web en accès libre ou non.

Pour rechercher des ontologies je me suis servie du moteur de recherche d'ontologies Swoogle⁹ et des mots clés suivants : *geography, space, topography, cartography, landscape, environment, territory, landuse, coast* et *relief* (j'ai fait des recherches avec les mêmes mots clés en français). Une fois tous les résultats fournis par Swoogle explorés, j'ai recherché sur

⁹ <http://swoogle.umbc.edu/>

Google les fichiers au format OWL, RDF, KIF, DAML et XML grâce à la syntaxe suivante : « filetype:owl » suivie des mots clés cités précédemment (en anglais et en français). Et pour finir j'ai analysé une liste d'ontologies créées avec l'éditeur Protégé¹⁰.

2.1.2. Grille de renseignements

Pour chaque ontologie d'objets géographiques trouvée je remplissais une grille de renseignements, qui m'a permis ensuite de comparer et de trier les ontologies. Cette grille contenait les mêmes champs pour toutes les ontologies (elle était différente pour les thesaurus), qui étaient renseignés ou non. Elles permettaient de récupérer le même type d'informations et ainsi de regrouper les ontologies selon certains critères. Par exemple dans ces grilles je renseignais le nombre de concepts que contenait l'ontologie, le sujet décrit dans celle-ci, etc. Certaines informations étaient difficiles à trouver, par exemple je n'ai pas pu accéder aux noms des auteurs à chaque fois.

La difficulté que j'ai rencontrée est que certains projets pour lesquels une ontologie avait été construite, était accompagné de documentation, et pour d'autres il n'y avait aucune information.

La figure 2 présente la grille de renseignements utilisée, accompagnée de l'explication de chaque champ. Les figures 3 et 4 présentent des exemples, la première concerne une ontologie pour laquelle j'ai trouvé beaucoup d'informations, et la deuxième présente une ontologie pour laquelle nous n'avions aucune information.

¹⁰ http://protegewiki.stanford.edu/index.php/Protege_Ontology_Library

Fiche de renseignements

Nom	<i>Nom du fichier contenant l'ontologie.</i>
Format	<i>OWL / XML / RDF / KIF / etc.</i>
Langue	<i>Langue dans laquelle sont écrits les noms des concepts.</i>
Auteur	<i>Information remplie quand l'ontologie a été créée par une seule personne.</i>
Organisme	<i>Nom de l'organisme à l'origine de l'ontologie. Adresse du site Web.</i>
Projet	<i>Nom du projet dans le cadre duquel l'ontologie a été créée.</i>
Type	<i>Ontologie de référence (ou de haut niveau) / Ontologie de domaine / Ontologie de tâche</i>
Domaine	<i>Domaine représenté par les concepts.</i>
Descriptif	<i>Descriptif du contenu de l'ontologie (par exemple le nom des grandes classes de concepts).</i>
Nb classes	<i>Nombre fourni par SWOOP¹¹.</i>
Nb individus	<i>Nombre fourni par SWOOP.</i>
Nb propriétés	<i>Nombre fourni par SWOOP.</i>
Nb types propriétés	<i>Nombre fourni par SWOOP.</i>
Nb relations	<i>C'est un nombre que je donne moi-même après l'analyse de l'ontologie. Correspond parfois aux nombres de propriétés.</i>
Définition	<i>Oui (formelle / informelle) / Non</i>
Traduction	<i>Oui (nom de la langue) / Non</i>
Commentaire Relation	<i>S'il n'y a pas trop de relations différentes je les répertorie ici.</i>
Niveau de complexité	<i>Lightweight (hiérarchie de concepts et de relations sémantiques non taxonomiques) / Heavyweight (+ des axiomes exprimés dans un langage logique adapté)</i>
Type de hiérarchie	<i>Hiérarchie de concepts / Hiérarchie de termes</i>
Profondeur	<i>Niveau de profondeur maximal, nombre fourni par SWOOP.</i>
Commentaire	
Divers	<i>Informations supplémentaires sur l'ontologie, par exemple les autres ontologies du projet.</i>

Figure 2 Grille de renseignements type

¹¹ Logiciel d'édition d'ontologies (cf. 2.1.3.).

Nom	Hydrology
Format	OWL (full)
Langue	Anglais
Auteur	Cathy Dolbear et Glen Hart
Organisme	Ordnance Survey (OS) http://www.ordnancesurvey.co.uk/oswebsite/ontology/
Projet	
Type	Ontologie de domaine
Domaine	Hydrologie
Descriptif	Caractéristiques topographiques impliquées dans la retenue et le transport d'eau intérieure superficielle.
Nb Classes	186
Nb individus	101
Nb propriétés	37
Nb type propriétés	0
Nb relations	37
Définition	Oui (formelles) + label
Traduction	Non
Commentaire Relation	
Niveau de complexité	Heavy weight
Type de hiérarchie	Hiérarchie de concepts
Profondeur	4
Commentaire	Les définitions ont l'air de traduire en langage naturel ce qui est exprimé grâce aux relations logiques.
Divers	Cette ontologie fait appel à d'autres ontologies de OS : LanguageRelations.owl ; SpatialRelations.owl ; PoliticalGeography.owl ; etc.

Figure 3 Grille de renseignements d'une ontologie de l'Ordnance Survey

Nom	Geo_swoogle2
Format	OWL
Langue	Anglais
Auteur	
Organisme	MobiLife Space Ontology
Type	Ontologie de domaine
Domaine	Concepts spatiaux
Descriptif	Propriétés géographiques Entités géopolitiques Lieux
Nb Classes	165
Nb individus	0
Nb propriétés	17
Nb type propriétés	4
Nb relations	1 (« is-a »)
Définition	Oui (informelle)
Traduction	Non
Commentaire Relation	
Niveau de complexité	Light weighth
Type de hiérarchie	Hiérarchie de concepts
Profondeur	6
Commentaire	
Divers	

Figure 4 Grille de renseignements d'une ontologie appelée geo_swoogle

A partir de ces grilles, j'ai créé un tableau recensant les différentes ontologies d'objets géographiques trouvées et les informations les concernant. J'ai fait apparaître dans ce tableau le type d'ontologie, c'est-à-dire si c'est une ontologie de référence, de domaine, de tâche ou un thesaurus, cette information m'a permis de faire le premier classement. Je ne peux pas trier la classe des thesaurus car je ne possède pas d'informations permettant de les regrouper. Dans les trois autres classes, j'ai regroupé les ontologies selon leur niveau de complexité (heavy weight ou light weight), ensuite selon la présence ou non de définitions pour les concepts et enfin selon la profondeur de la hiérarchisation. La figure 5 est un extrait du tableau obtenu, la totalité du tableau est proposé dans le document annexe.

Type	Niveau de complexité	Définition	Profondeur	Nom du fichier	Organisme	Domaine traité	Langue
Ontologie de domaine	Heavy weight	Oui	8	http://www.fao.org/aims/aos/cwr.owl	AGROVOC	Agricole	Multilingue
Ontologie de domaine	Heavy weight	Oui	8	http://wow.sfsu.edu/ontology/rich/EcologicalConcepts.owl	Rocky Mountain Biological Lab	Ecologie et biotique	Anglais
Ontologie de référence	Light weight	Non	4	http://iri.columbia.edu/%7EBenno/gcmd.owl	NASA EOS	Sciences de la Terre	Anglais
Ontologie de référence	Light weight	Non	5	geobrain.laits.gmu.edu/ontology/2004/11/gcmd-science.owl	NASA EOS	Sciences de la Terre	Anglais

Figure 5 Extrait du tableau recensant les ontologies

2.1.3. Outils

Pour visualiser les ontologies trouvées sur le Web, j'ai utilisé l'éditeur Protégé-2000¹². Il a été développé par l'université de Stanford et permet principalement de construire des ontologies. Nous pouvons également utiliser la librairie Java pour créer des applications à bases de connaissances. J'ai choisi d'utiliser cet éditeur car c'est un logiciel open source et il propose une visualisation des ontologies sous forme d'arbre.

J'ai aussi utilisé l'éditeur SWOOP¹³, c'est un outil pour créer, éditer et déboguer les ontologies en OWL. Il a été produit par un laboratoire MIND de l'université du Maryland. C'est également un logiciel open source. Cet éditeur m'a permis de visualiser les ontologies dont le code était erroné, celles dont l'extension du fichier était .rdf ou .xml, et celles faisant appel à des ontologies que je ne possédais pas. Il m'a donc servi à visualiser presque toutes les ontologies que je ne pouvais pas ouvrir avec Protégé. De plus cet éditeur fournissait des informations utiles à l'étude des ontologies, il donnait pour chacune d'elles le nombre de concepts, le nombre de types de propriétés, le nombre d'individus, le niveau de profondeur maximal, etc.

¹² <http://protege.stanford.edu/>

¹³ <http://www.mindswap.org/2004/SWOOP/>

2.2. Résultats

Je présente dans cette partie quelques ontologies que j'ai trouvées sur le Web. La totalité de l'état de l'art est présenté dans le document annexe joint au rapport.

2.2.1. Ontologies en français

- *Towntology*¹⁴ :

Le projet Towntology a pour but de « *définir une ontologie utilisée à la fois pour l'enseignement de l'urbanisme et proposer aussi aux experts un cadre de référence, pour l'indexation de leur documentation, l'aide à la recherche d'information ou la formation du personnel* »¹⁵. Il est issu de la collaboration du LIRIS (Laboratoire d'InfoRmatique en Image et Systèmes d'information) et de l'EDU (Equipe Développement Urbain). Dans ce cadre une suite d'outils de visualisation d'ontologies ainsi qu'une ontologie dans le domaine de l'aménagement et l'urbanisme ont été développées. L'objectif de l'ontologie est de permettre « *l'interopérabilité entre modèles de bases de données et systèmes de conception coopératifs différents, ainsi que la communication entre les différents intervenants dans l'aménagement du territoire* »¹⁶. Pour construire leur ontologie ils ont utilisé l'approche middle-out qui consiste à déterminer les notions les plus courantes dans le domaine, puis à les généraliser ou/et les spécialiser. Pour récupérer les termes du domaine et leur(s) définition(s) plusieurs ouvrages du domaine de l'urbanisme écrit par des chercheurs de l'EDU, ainsi qu'un dictionnaire de l'urbanisme et *Le grand dictionnaire terminologique* ont été utilisés.

L'ontologie que j'ai pu récupérer (ville.xml) comporte 98 classes avec une profondeur de 3 niveaux ainsi que 4 relations, chaque classe (ou concept) est accompagnée d'une définition. Mais il semblerait que cette ontologie ait été complétée depuis : elle utilise maintenant environ 21 relations sémantiques, et contient plus de 800 concepts. La figure 6 propose un extrait de l'ontologie sous forme de graphe, et la figure 7 présente les définitions du concept ville et les relations qu'il entretient avec d'autres concepts.

¹⁴ <http://liris.cnrs.fr/~townto/>

¹⁵ Abdel Kader Keita, Catherine Roussey et Robert Laurini, « Un outil d'aide à la construction d'ontologies pré-consensuelles : le projet Towntology ».

¹⁶ Roussey C., Laurini R. Beaulieu C., Tardy Y. et Zimmermann M., 2004, « Le projet Towntology, un retour d'expérience pour la construction d'une ontologie urbaine », in *Revue internationale de Géomatique*, n° 2/2004.

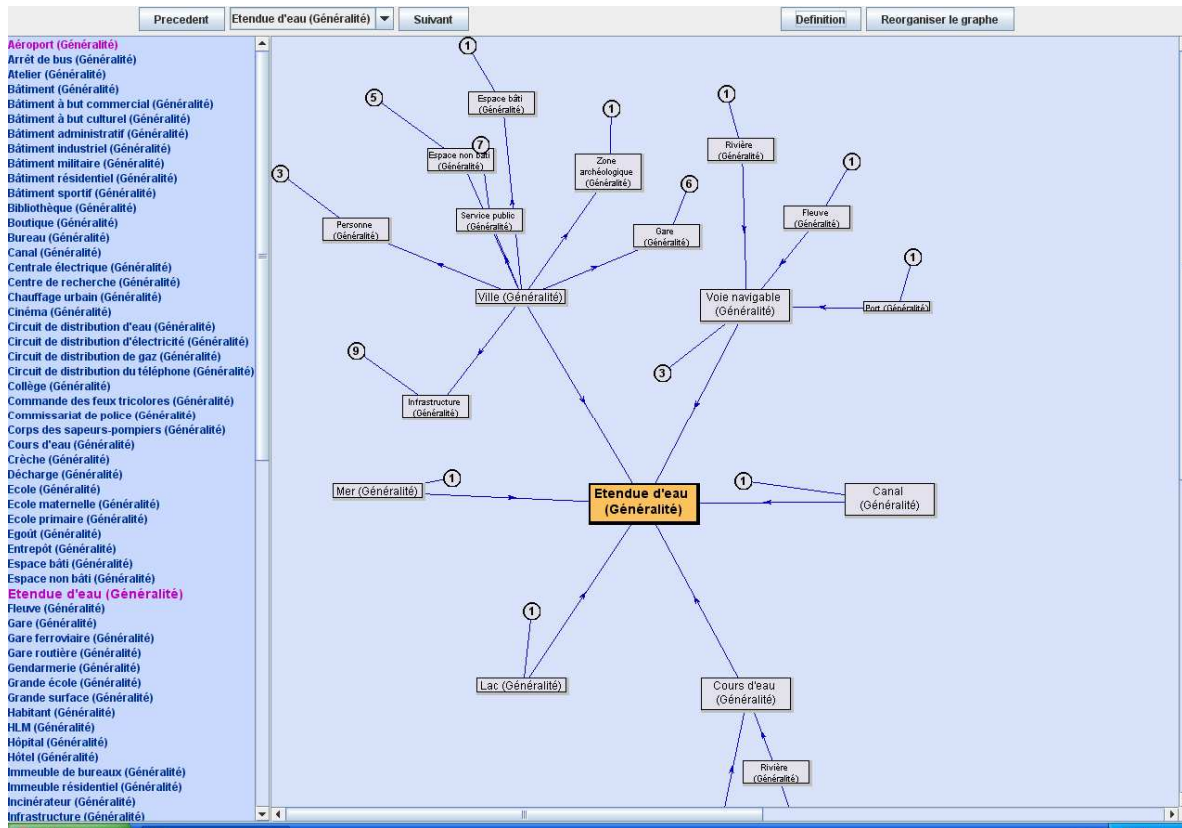


Figure 6 Visualisation du graphe de l'ontologie ville avec les outils du projet Towntology

Graphe	Ville
<p>Ontologie</p> <p>Titre : Ville Langue : français Organisme : EDU - Samuel GESCHE Dernière modification : 2004/09/23</p>	<p>Ville</p> <p>Définitions (Généralité)</p> <p>(entrée le 2004/09/21 par Samuel GESCHE)</p> <p>UNE COLLECTION DE DÉFINITIONS DU CONCEPT "VILLE"</p> <p>Les définitions de la ville ici présentes, ont fait l'objet d'un choix arbitraire, mais aussi contrasté que possible. Bien évidemment il s'agit d'une liste ouverte, à compléter. Il faut remarquer que la plupart des auteurs dissertant sur la ou les villes ne donne jamais de définitions de ce mot ou n'accepte peu ou prou de réduire leur pensée en une formule brève.</p> <p>ALBERTI L.B. (Milieu du XVème siècle) cité par P. LAVEDAN (1941)</p> <p>ville = commodité + volupé</p> <p>ALEXANDRE C. (1976)</p> <p>"La ville - le système urbain - système complexe animé de flux multiples et divers - d'interactions nombreuses souvent insoupçonnées d'interférences voulues ou spontanées - pourvu de phénomènes d'autorégulation - de mémoire - de facultés conversationnelles avec son environnement - terrain de rencontre du Politique et du Social - du Social et de l'Economique - de l'Economique et du Politique - du Spontané et du Programmé-PUZZLE aux pièces innombrables - MECANISME à démonter - à désarticuler pour être mieux compris".</p>
<p>Liste des relations</p> <p>Ville (Généralité) est composé de Personne (Généralité) Ville (Généralité) est composé de Infrastructure (Généralité) Ville (Généralité) est composé de Espace non bâti (Généralité) Ville (Généralité) est composé de Espace bâti (Généralité) Ville (Généralité) est composé de Service public (Généralité) Ville (Généralité) est composé de Etendue d'eau (Généralité) Ville (Généralité) est composé de Zone archéologique (Généralité) Ville (Généralité) est composé de Gare (Généralité)</p>	<p>BAILLY A.S. (1975)</p> <p>"Le système urbain, tel qu'il est dégagé par les théories et modèles des hiérarchies urbaines est conçu comme formé d'un ensemble de centres de niveaux différents, liés entre eux par des flux".</p> <p>CASTELLS M. et GODARD F. (1974)</p> <p>"Un système urbain est l'ensemble d'articulations sociales spécifiques qui composent une unité de reproduction collective de la force de travail. Il est donc formé d'éléments (composés intérieurement) et de la loi de combinaison, d'interdépendance et de hiérarchie entre ces éléments". donc : Les éléments définissant, dans leurs rapports, le système urbain sont - l'élément Consommation (C), expression spécifique de la reproduction de la force de travail, - l'élément Production (P), expression spécifique de la reproduction des moyens de production, - l'élément Echange (E), résultant des transferts internes entre P et C, à l'intérieur de P et à l'intérieur de C, - l'élément Symbolique, expression spécifique de l'Instance idéologique, - l'élément Gestion ou ensemble des interventions du politico-institutionnel relatives aux éléments urbains".</p> <p>CHOAY F. (1975 p. 78)</p> <p>"La ville n'est seulement un objet ou un instrument, le moyen d'accomplir certaines fonctions vitales ; elle est également un cadre de relations inter-conscientielles, le lieu d'une activité qui consomme des systèmes de signes autrement (plus) complexes". L'urbanisme a méconnu cette réalité, méconnaissant par là même la nature de la ville.</p>

Figure 7 Visualisation des définitions et des relations du concept ville de l'ontologie du projet Towntology

Le projet COST UCE (Urban Civil Engineering) Action C21¹⁷ englobe le projet towntology, il a pour objectif de produire une ontologie dans le domaine de l'aménagement et de l'urbanisme, avec une présentation textuelle et visuelle. Cette ontologie est accompagnée d'un éditeur qui permet d'intégrer et de mettre à jour les concepts, définitions, photos, etc. Les objectifs parallèles à la constitution d'une ontologie sont de développer une série de recommandations pour la construction d'une ontologie multilingue dans le domaine UCE, ontologie basée sur des cas pratiques, et d'analyser le rôle des ontologies comme un outil pour améliorer la communication entre les intervenants du domaine UCE. L'ontologie créée a pour rôle de faciliter la communication entre les systèmes d'information, les intervenants et les spécialistes du domaine UCE au niveau européen.

- ***Fodomust***¹⁸ :

Fodomust est « *un projet de fouille de données multi-stratégie pour extraire et qualifier la végétation urbaine à partir de bases de données d'images* ». 3 laboratoires sont à l'oeuvre dans ce projet : LSIT (Laboratoire des Sciences de l'Images, de l'Informatique et de la Télédétection) et LIV (laboratoire Image et Ville) du CNRS et le laboratoire ERIC (Equipe de Recherche en Ingénierie des Connaissances) de l'Université Lumière Lyon. Pour l'interprétation des images les chercheurs ont besoin de donner un sens aux éléments identifiés sur les images, pour cela ils ont décidé de créer une ontologie qui leur permet d'avoir une représentation des connaissances du domaine. L'ontologie a été construite à partir d'entretiens avec les experts du LIV, par apprentissage automatique, et d'un dictionnaire de données recensant les objets géographiques.

Je n'ai pas trouvé l'ontologie sur le Web, en revanche on la trouve rédigée avec la syntaxe OWL dans l'annexe de l'article « Ontologie des objets urbains » de Bertille Fremaux et Thi-Thuy N'Guyen, mais ni Protégé ni Swoop n'est capable de la lire. Dans cette ontologie je compte 112 concepts. La figure 8 présente un extrait de la structuration de l'ontologie en OWL.

¹⁷ www.towntology.net

¹⁸ <http://lsiit.u-strasbg.fr/afd/fodomust>

```

<rdf:RDF
  xmlns:protege="http://protege.stanford.edu/plugins/owl/protege#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns=""="http://www.owl-ontologies.com/unnamed.owl#"
  xml:base="http://www.owl-ontologies.com/unnamed.owl">
  <owl:Ontology rdf:about="">
    <owl:imports
      rdf:resource="http://protege.stanford.edu/plugins/owl/protege" />
    </owl:Ontology>
    <owl:Class rdf:ID="Alignement_d_arbres">
      <rdfs:label>Alignement d arbres</rdfs:label>
      <protege:abstract>true</protege:abstract>
      <rdfs:subClassOf>
        <owl:Class rdf:ID="Surfaces_arborees" />
      </rdfs:subClassOf>
    </owl:Class>
    <owl:Class rdf:ID="Naturels">
      <protege:abstract>true</protege:abstract>
      <rdfs:subClassOf>
        <owl:Class rdf:ID="Plans_d_eau" />
      </rdfs:subClassOf>
    </owl:Class>
    <owl:Class rdf:ID="Batiment">
      <rdfs:subClassOf>
        <owl:Class rdf:ID="Autres_espaces_urbains_specialises" />
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <owl:Class rdf:ID="Espaces_associes_au_reseau_routier" />
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <owl:Class rdf:ID="Decharges_et_depots_de_remblais_-
          _deblais" />
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <owl:Class rdf:ID="Zones_industrielles" />
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <owl:Class rdf:ID="Fortes_densite_de_contruit_-
          _Lotissement" />
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <owl:Class rdf:ID="Emprises_militaires" />
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <owl:Class rdf:ID="Friches_industrielles" />
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <owl:Class rdf:ID="Carrières_actives" />
      </rdfs:subClassOf>
    </owl:Class>
  </rdf:RDF>

```

Figure 8 Extrait de l'ontologie des objets urbains du projet Fodomust

- **GIEA¹⁹** :

Le nombre et la diversité des informations demandées par les entreprises économiques, l'administration, ou autres, aux agriculteurs est en forte augmentation. Pour standardiser les échanges de données, entre l'exploitant agricole et ses interlocuteurs, le projet GIEA (Gestion des informations de l'exploitation agricole) a été lancé. Il a pour objectif de coordonner les travaux de standardisation menés par les différents organismes, de définir les informations demandées de façon commune pour les agriculteurs et de permettre l'interopérabilité des systèmes d'information publics et privés du domaine agricole.

¹⁹ www.projetgiea.fr/

Dans le cadre de ce projet ils ont élaboré une ontologie du domaine agricole. L'ontologie couvre les concepts communs utilisés dans les principales chaînes de production agricole et elle a pour objectif « *de fournir des formats d'échange de données complets pour faciliter l'interopérabilité des systèmes d'information agricoles* »²⁰. Les classes décrites sont par exemple *Alimentation, Bâtiments agricoles, Occupation du sol*, etc. Je n'ai pour l'instant pas pu récupérer l'ontologie, j'ai en revanche pu avoir les présentations des données concernant le sol, l'exploitation et l'élevage, où pour chaque concept il y a une définition, une description des principales relations, des règles de gestion, des attributs, des détails sur les attributs, des remarques et les termes proches. Ces présentations sont très proches des spécifications des bases de données de l'IGN (cf. figure 9).

Voirie et réseau divers (CF Données permanentes)

Définition :
 Tout ou partie de réseaux permettant la circulation de véhicules (voirie), de fluides (canalisation), d'énergie (câbles) ou d'information.

Description des principales relations :
[Infrastructure](#) : une voirie et réseau divers peut être liée à une d'infrastructure
[Occupation du sol](#) : une voirie ou réseau divers peut être un type d'occupation du sol
[Objet géoréférencé](#) : une voirie ou réseau divers peut être associée à un objet géographique

Règles de gestion :

Attributs :

Nom attribut	Statut	Format	Liste
Libellé	O	AN,,100	

Détail sur les attributs :

Remarques :

Termes proches : [Retour liste de concepts](#)

Élément du paysage (pré-validé 15/11/2006)

Définition :
 Élément visible naturel ou historique du territoire, éventuellement aménagé pour les besoins de l'exploitation, mais n'entrant pas dans le cadre d'une production agricole

Description des principales relations :

Figure 9 Projet GIEA : présentation des données sur le sol

²⁰ Catherine Roussey et Myoung-Ah Kang, 2008, « Les Ontologies et leurs Applications en Agriculture ».

2.2.2. Ontologies en anglais

- *Ordnance Survey*²¹ :

Les chercheurs de l'Ordnance Survey (l'équivalent britannique de l'IGN) ont produit des ontologies, qui associées à une série d'opérations automatiques pourront permettre la combinaison de bases de données géographiques différentes. Leur désir était de décrire les classes contenant des caractéristiques hydrologiques étudiées par Ordnance Survey. L'intérêt de ces ontologies était d'améliorer l'utilisation de leurs données par leurs clients et de faciliter le traitement semi-automatique des données. Ils ont créé trois ontologies de domaines, une représentant le domaine de connaissances de l'hydrologie, une deuxième qui décrit les lieux et les bâtiments et la dernière les entités géographiques administratives.

Ces ontologies ont été, il me semble, construites à la main, des experts ont listé les concepts, puis ils ont recherché des informations complémentaires dans des ouvrages du domaine (pour les définitions, les relations, etc.). Ils ont créé à partir de ces informations un glossaire de connaissances, qui décrit précisément tous les concepts et toutes les relations, à partir de ce glossaire ils ont construits l'ontologie.

Je ne vais parler ici que de l'ontologie qui décrit le domaine de l'hydrologie. Elle présente les caractéristiques topographiques impliquées dans les phénomènes de retenue et de transport d'eaux superficielles. Elle fait appel à d'autres ontologies, appelées ontologies modulaires, qui décrivent les relations spatiales, les relations météorologiques, etc. Elle est composée de 186 concepts, avec une profondeur de 4 niveaux, les concepts sont reliés entre eux grâce à 37 relations sémantiques. Cette ontologie est très riche et très détaillée : elle fournit, pour chaque concept, une définition. Elle contient également 101 instances. L'Ordnance Survey propose deux versions de cette ontologie, la dernière date de janvier 2008. Un extrait de la dernière version est présenté dans la figure 10.

Une ontologie de haut niveau regroupant les termes généraux du domaine de la topographie est en cours de construction depuis février 2008, elle contient pour l'instant 17 concepts, 7 individus et 7 relations reliant les concepts entre eux, et les concepts et les individus.

²¹ <http://www.ordnancesurvey.co.uk/oswebsite/ontology/>

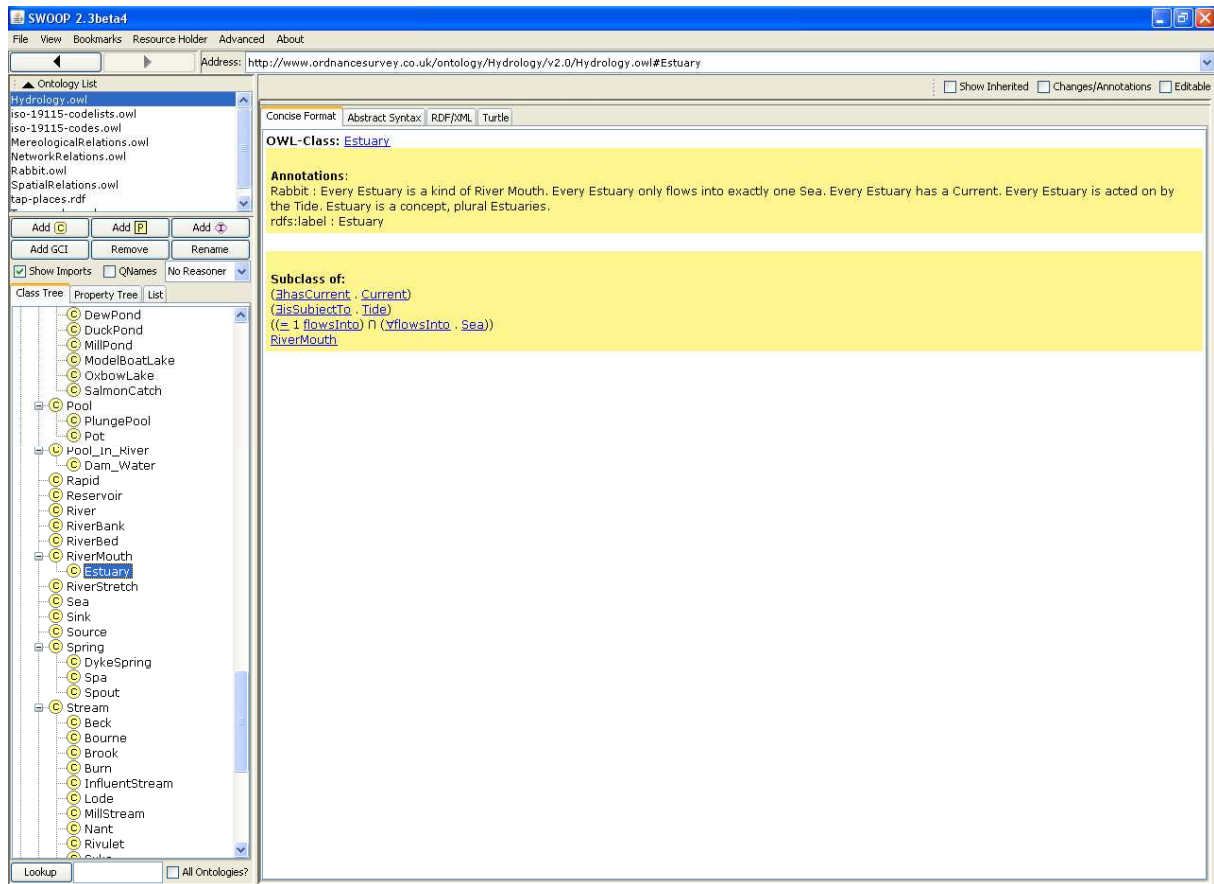


Figure 10 Visualisation de l'ontologie *hydrology* de l'Ordnance Survey avec Swoop

- **SUMO²²** :

L'IEEE Standard Upper Ontology Working Group a travaillé à l'élaboration d'une ontologie de haut-niveau appelée SUMO (Suggested Upper Merged Ontology). Dans le cadre de ce travail des ontologies dans le domaine de la géographie et dans le domaine des transports ont été créées. Celle portant sur la géographie est très détaillée, elle comporte 432 classes, avec une profondeur de 7 niveaux, 651 instances reliées grâce à 99 relations. De plus cette ontologie fournit une définition pour la plupart des concepts (cf. figure 11). Voici quelques classes décrites dans cette ontologie : *Static Water Area*, *Water Area*, *Wood*, *Geographic Area*, *Geopolitical Area*.

²² <http://www.ontologyportal.org/>

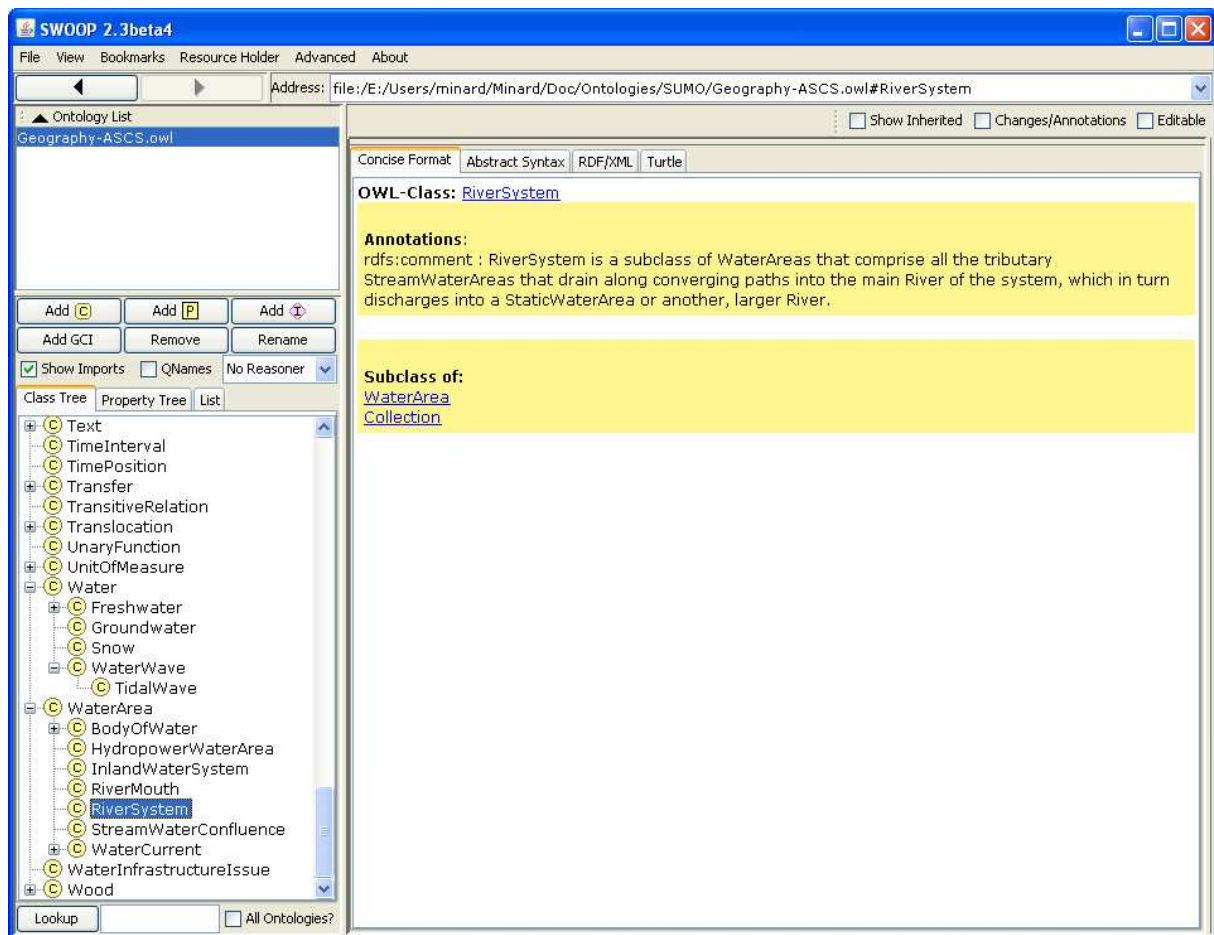


Figure 11 Visualisation de l'ontologie géographique du projet SUMO avec Swoop

2.2.3. Ontologies multilingues

- *WalkOnWeb*²³ :

Le projet WalkOnWeb a pour objectif de définir un nouveau modèle de publication d'information pour les randonneurs et les touristes. Ce modèle doit permettre de créer un itinéraire de randonnée s'étendant sur plusieurs pays, ainsi que la description de l'itinéraire dans plusieurs langues, un itinéraire en sens inverse, ...

Pour ce projet 5 ontologies ont été construites. Une ontologie contenant des concepts généraux (*general purpose ontology*) qui concernent l'état des routes, les types d'indicateur de temps, etc. Une ontologie de navigation (*navigation ontology*) renferme des concepts comme les directions. La *topo-ontology* est formée des concepts et des relations existant dans une base vectorielle de données géographiques. Les deux dernières, enfin, sont des ontologies de tâches qui varient selon la randonnée. La première : *walk-ontology* est utilisée pour formaliser les informations concernant la randonnée et la deuxième sert à capturer tous les types d'applications spécifiques de cette randonnée, par exemple les centres d'intérêts de la marche. En plus de ces 5 ontologies l'ontologie de haut-niveau *WordNet* a été utilisée.

L'ontologie à laquelle j'ai eu accès sur leur site Internet regroupe trois ontologies : *general purpose ontology*, *navigation ontology* et *walk ontology*. Elle est écrite en anglais, comporte 108 concepts, avec une profondeur de 6 niveaux, 4463 instances et 75 types de relations entre les concepts et entre les instances. La plupart des concepts sont accompagnés d'une définition.

La *walk ontology* a été créée pour que les auteurs d'itinéraires puissent décrire une balade en utilisant des modèles d'information prédéfinis. Dans ces modèles il y a des directives de parcours, comme par exemple *changez de direction*, des orientations, comme *droite*, des modes de balisage, des caractéristiques géographiques, et entre autres des objets géographiques, ou des relations géospatiales, par exemple *devant*. Les informations enregistrées par les auteurs peuvent ensuite être modifiées automatiquement (par exemple pour avoir l'itinéraire en sens inverse) et sont traduites en langage naturel suivant la demande du randonneur.

²³ www.walkonweb.org/

Voici deux exemples des instances que l'on a dans *walk ontology* :

- le concept *Topographic feature* contient 383 instances, on y trouve des noms généraux de lieu en français, anglais ou néerlandais (*parking, piste, etc.*) et des toponymes en français ou en néerlandais (*allée des rhododendrons, chemin du pavé, etc.*).
- le concept *Nl Element Navigation Directions* contient 857 instances, ils correspondent à des instructions de la direction à prendre, les instructions sont en français ou en néerlandais (ex : *GRPtourdeChartreuse_33_1FR* a comme texte *environ 50m après*).

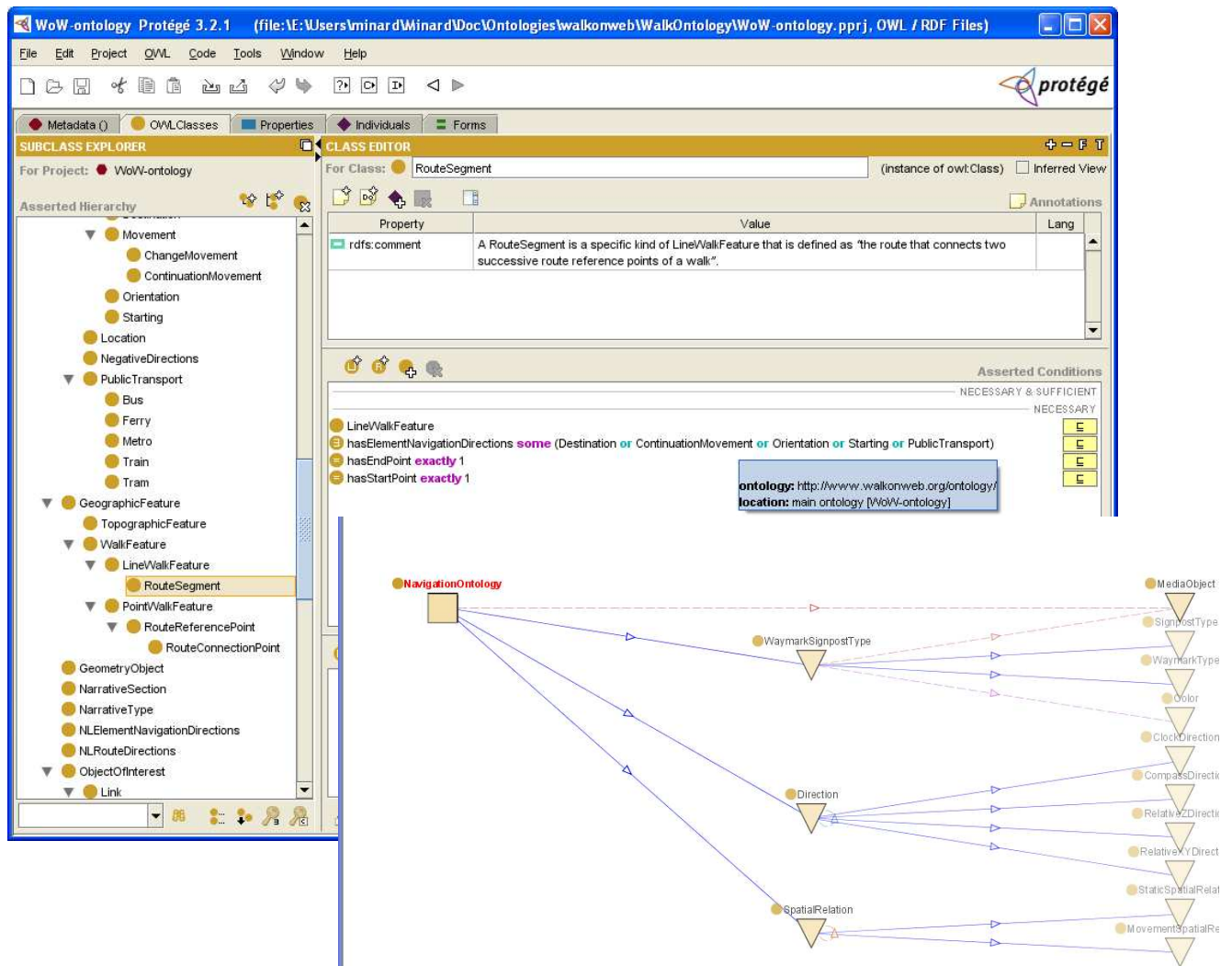


Figure 12 Visualisation de l'ontologie WalkOnWeb sur Protégé

2.2.4. Thesaurus

- *Activités gouvernementales*²⁴ :

L'université de Montréal a développé un thesaurus en français qui décrit le domaine des activités gouvernementales. Les termes sont reliés grâce à 6 relations : EP (Employé Pour), EM (Employer), TA (Terme Associé), CT (Catégorie), TG (Terme Générique) et TS (Terme Spécifique). Voici quelques unes des thématiques du thesaurus : *Environnement, Gouvernement en vie politique, Ressources naturelles, Territoire, Tourisme, Transport*, etc. La figure 13 présente les termes de la catégorie Environnement et la figure 14 la description du terme *cours d'eau*.

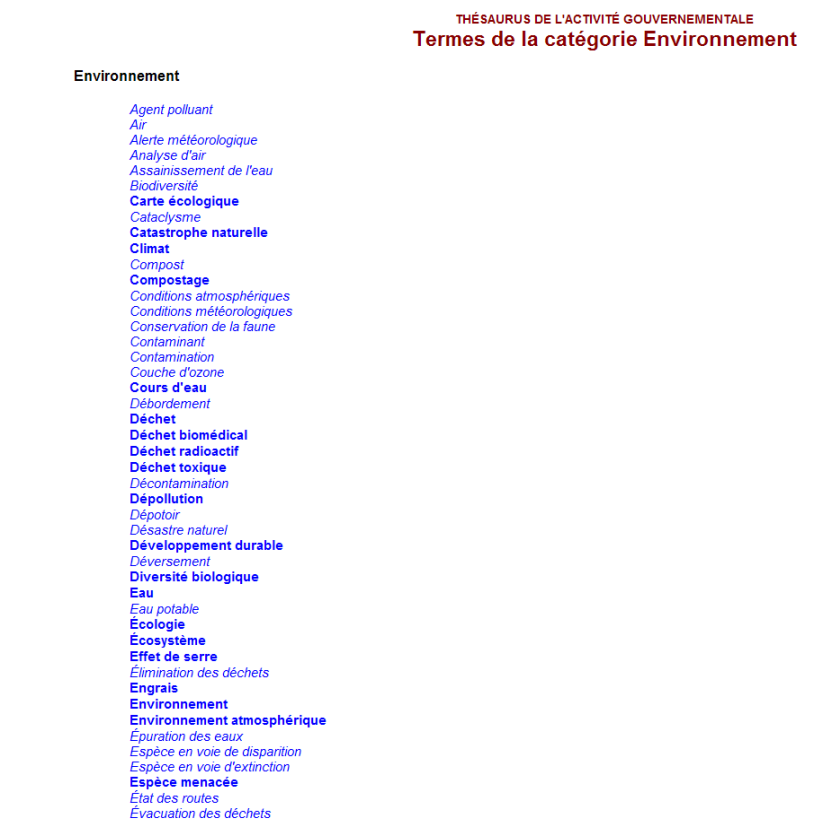


Figure 13 Extrait du thesaurus de l'université de Montréal



EP - Employé pour, EM - Employer, TA - Terme Associé, CT - Catégorie, TG - Terme Générique, TS - Terme Spécifique

Figure 14 Description du terme *Cours d'eau* du thesaurus de l'université de Montréal

²⁴ <http://grds.ebsi.umontreal.ca/>

2.2.5. Bilan

A partir des grilles de renseignements et de l'état de l'art, j'ai trié les ontologies pour faire un bilan de l'état de l'art. Je les ai d'abord classées selon leur type, c'est-à-dire en séparant les ontologies de référence de celles de domaine et de celles de tâches.

Le premier bilan que je peux tirer de la recherche d'ontologies d'objets géographiques est qu'il y a très peu d'ontologies géographiques dont le langage de description est le français. On trouve par contre une large gamme d'ontologie en anglais. On peut les classer en fonction de la présence ou non d'une définition pour chaque concept, d'une traduction, des conditions nécessaires et suffisantes, des relations autres que « is-a », ... J'ai essayé de trier toutes ces ontologies, le premier classement sépare les ontologies de référence, de domaine et de tâche, puis à l'intérieur de ces trois classes je propose de trier ces ontologies selon la profondeur de la hiérarchie. Et le deuxième effectue un tri selon le niveau de complexité (lightweight ou heavyweight), selon le type d'ontologie, puis selon la présence ou non de définition.

J'ai recensé 49 ontologies du domaine de la géographie et des objets géographiques, dont 8 ontologies de références, 31 ontologies de domaine, 4 ontologies de tâches et 7 thesaurus. J'ai cherché à la fois des ontologies en français et en anglais, j'ai trouvé 3 ontologies écrites uniquement en français, 37 en anglais et 10 multilingues, c'est-à-dire que les termes sont traduits dans plusieurs langues, dans ces dernières le langage de description est souvent l'anglais.

Dans l'état de l'art 17 ontologies sont décrites, sa version complète est proposée dans le document annexe joint à ce rapport.

3. D'un thesaurus à une ontologie

Quand j'ai fait l'état de l'art des ontologies d'objets géographiques, j'ai trouvé des thesaurus spécifiques au domaine de la géographie, mais aussi des thesaurus traitant de domaines plus larges. Un thesaurus a particulièrement intéressé les chercheurs du laboratoire, il a été construit au Canada en français (cf. 2.2.4). Les concepts décrits sont du domaine de l'activité gouvernementale et sont regroupés en catégorie, l'une d'elle concerne le domaine de la géographie. Pour que ce thesaurus puisse servir dans l'amélioration de l'ontologie de l'IGN, il m'a été demandé de le transformer en une ontologie au format OWL.

Le deuxième objectif du stage a donc été d'écrire un programme en Java permettant de convertir un thesaurus au format PDF en une ontologie au format OWL. Le langage OWL est un langage XML utilisé pour structurer les ontologies. Une ontologie en OWL peut être facilement utilisée par une application informatique. Contrairement au format PDF qui est un format d'impression préservant la mise en forme, le format OWL des ontologies écrites avec le langage du même nom ne prend pas en compte la mise en forme, mais permet uniquement de structurer les informations.

Dans cette partie je présente le thesaurus, l'ontologie que je souhaite obtenir, le processus d'écriture du programme en JAVA, du thesaurus au format PDF des activités gouvernementales du Canada à l'ontologie au format OWL en passant par le thesaurus structuré en XML. Je termine par un point sur les relations entre les concepts et plus précisément comment faire correspondre les relations présentes dans un thesaurus et celles d'une ontologie.

3.1. *Thesaurus*

Le thesaurus contient des concepts classés par ordre alphabétique, en dessous de chacun d'entre eux sont indiquées les relations taxonomiques qu'ils entretiennent avec d'autres. Il y a 6 types de relations : *employer plutôt*, *employé pour*, *générique*, *spécifique*, *catégorie* et *associé à*.

Exemple :

Sciences pures

Employer plutôt: Sciences naturelles

Catégorie: Science et technologie

Sciences spatiales

Employé pour: Aérospatiale

Employé pour: Agence spatiale

Employé pour: Espace

Employé pour: Recherche spatiale

Générique: Sciences

Spécifique: Aéronautique

Spécifique: Astronautique

Spécifique: Astronomie

Spécifique: Astrophysique

Associé à: Technologie spatiale

Catégorie: Science et technologie

3.2. *Ontologie*

L'ontologie à obtenir doit être écrite avec le langage OWL. Elle hiérarchise les concepts grâce à des relations « is-a », et instaure des relations entre les concepts appelées *Object Property*. En théorie ces relations relient les individus de classes différentes entre eux, mais pas directement les classes. Mais reliés les individus en fonction de la classe auxquelles ils appartiennent, revient à dire que les classes sont reliées entre elles par ces mêmes relations.

L'ontologie sera composée de 20 classes correspondant aux 20 catégories du thesaurus, chacune d'elles contiendra des concepts, eux-mêmes hiérarchisés grâce à une relation de spécialisation.

3.3. *Prétraitement du thesaurus*

J'ai implémenté cette transformation avec le langage dirigé objet JAVA. J'ai trouvé le thesaurus au format PDF (cf. figure 15), je l'ai ensuite enregistré au format TXT (cf. figure 16). Le thesaurus au format texte avait récupéré que partiellement la mise en forme du thesaurus initial. En effet les espaces et les tabulations, présentes dans le format de départ, apparaissent de façon irrégulière dans le fichier TXT.

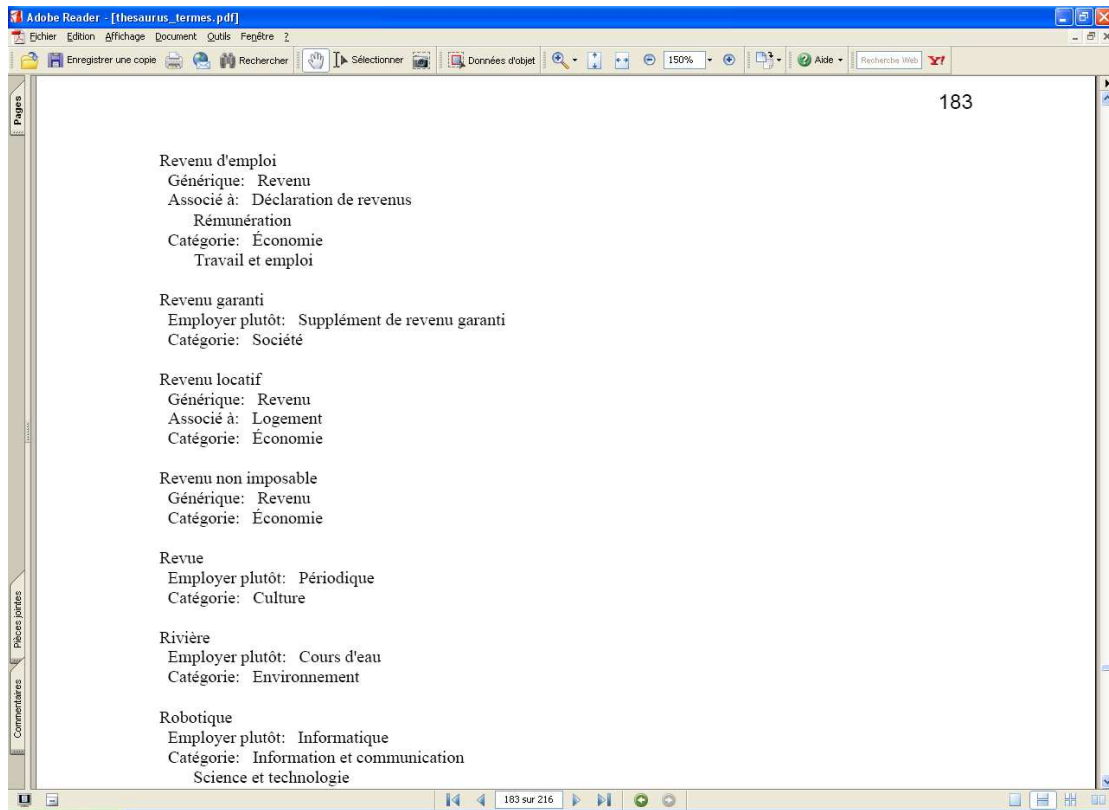


Figure 15 Extrait du thesaurus au format PDF.

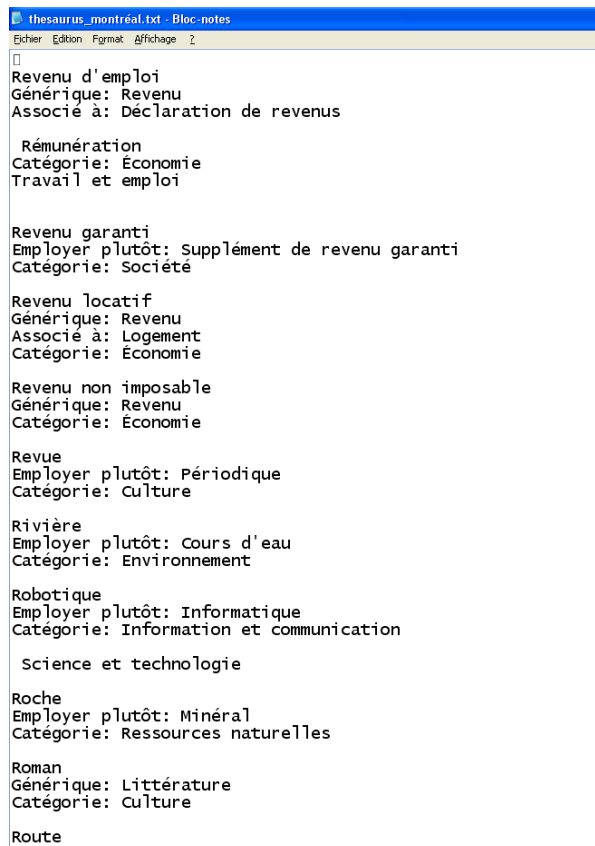


Figure 16 Extrait du thesaurus au format TXT.

La mise en forme permet de distinguer le concept décrit du type de relations qu'il entretient avec d'autres concepts. Sans mise en forme il est difficile de savoir où commence et où s'arrête la description d'un concept. Pour récupérer la mise en forme et pouvoir distinguer les concepts des propriétés, il a fallu que je prétraite le document TXT.

Le programme JAVA, que j'ai écrit pour prétraiter le thesaurus, prend en entrée le fichier au format TXT. Ensuite il découpe en ligne le thesaurus et traite une ligne à la fois (une ligne s'arrête à l'endroit d'un saut de ligne). A la fin du prétraitement le programme produit un nouveau fichier au format TXT (cf. figure 17).

```
Sans titre - Bloc-notes
Fichier  Edition  Format  Affichage  ?
Revenu d'emploi
  Générique: Revenu
  Associé à: Déclaration de revenus
  Associé à: Rémunération
  Catégorie: Économie
  Catégorie: Travail et emploi

Revenu garanti
  Employer plutôt: Supplément de revenu garanti
  Catégorie: Société

Revenu locatif
  Générique: Revenu
  Associé à: Logement
  Catégorie: Économie

Revenu non imposable
  Générique: Revenu
  Catégorie: Économie

Revue
  Employer plutôt: Périodique
  Catégorie: Culture

Rivière
  Employer plutôt: Cours d'eau
  Catégorie: Environnement

Robotique
  Employer plutôt: Informatique
  Catégorie: Information et communication
  Catégorie: Science et technologie

Roche
  Employer plutôt: Minéral
  Catégorie: Ressources naturelles
```

Figure 17 Extrait du thesaurus obtenu après traitement.

Les modifications effectuées sur le fichier de départ sont présentées ci-dessous grâce aux figures 18 et 19. La première représente un extrait du fichier TXT de départ, la deuxième le résultat du prétraitement de l'extrait.

183

Revenu d'emploi#

Générique: Revenu#

Associé à: Déclaration de revenus

#Rémunération#

Catégorie: Économie#

Travail et emploi

Revenu garanti#

Employer plutôt: Supplément de revenu garanti#

Catégorie: Société#

Robotique#

Employer plutôt: Informatique#

Catégorie: Information et communication

#Science et technologie

Figure 18 Extrait du fichier TXT de départ.

183 : remplacement des lignes contenant les numéros de pages par des lignes vides.

: suppression des espaces placées après les mots.

: remplacement des espaces précédant une chaîne de caractères par une tabulation.

Travail et emploi : ajout d'une tabulation et du nom de la propriété devant les chaînes de caractères contenant la valeur de la propriété mais pas son nom.

Science et technologie : ajout d'une tabulation et du type de propriétés, devant les chaînes de caractères précédées d'un espace.

Générique: Revenu : ajout d'une tabulation devant les types de propriétés.

Revenu d'emploi

Générique: Revenu

Associé à: Déclaration de revenus

Associé à: Rémunération

Catégorie: Économie

<p>Catégorie: Travail et emploi</p> <p>Revenu garanti</p> <p>Employer plutôt: Supplément de revenu garanti</p> <p>Catégorie: Société</p> <p>Robotique</p> <p>Employer plutôt: Informatique</p> <p>Catégorie: Information et communication</p> <p>Catégorie: Science et technologie</p>
--

Figure 19 Extrait de la figure 18 après traitement.

3.4. Structuration en XML

J'ai ensuite transformé le fichier TXT obtenu en 3.3 en un fichier XML. Le programme repère dans le fichier texte les propriétés et les balisent <propriete>. Chaque balise <propriete> a un attribut xsi:type="" qui a comme valeur le type de propriété. La valeur de la propriété est balisée <valeur>. Et pour finir les concepts sont entourés de la balise <concept> et <label>.

Ce balisage est ensuite modifié pour que la balise <concept> regroupe le concept et ses propriétés. J'ai remarqué que certains concepts ne possédaient pas de propriétés, dans ce cas je leur crée une propriété du type *catégorie* et qui a pour valeur *Autre*.

La figure 20 présente un extrait du balisage XML du thesaurus.

```

<concept>
  <label>Revenu d'emploi</label>
  <propriete xsi:type="generique">
    <valeur>Revenu</valeur>
  </propriete>
  <propriete xsi:type="associe">
    <valeur>Déclaration de revenus</valeur>
  </propriete>
  <propriete xsi:type="associe">
    <valeur>Rémunération</valeur>
  </propriete>
  <propriete xsi:type="categorie">
    <valeur>Économie</valeur>
  </propriete>
  <propriete xsi:type="categorie">
    <valeur>Travail et emploi</valeur>
  </propriete>
</concept>
<concept>
  <label>Revenu garanti</label>
  <propriete xsi:type="employer_plutot">
    <valeur>Supplément de revenu garanti</valeur>
  </propriete>
  <propriete xsi:type="categorie">
    <valeur>Société</valeur>
  </propriete>
</concept>
<concept>
  <label>Robotique</label>
  <propriete xsi:type="employer_plutot">
    <valeur>Informatique</valeur>
  </propriete>
  <propriete xsi:type="categorie">
    <valeur>Information et communication</valeur>
  </propriete>
  <propriete xsi:type="categorie">
    <valeur>Science et technologie</valeur>
  </propriete>
</concept>

```

Figure 20 Balisage XML de l'extrait de la figure 19.

3.5. Création de l'ontologie en OWL

Grâce à un troisième programme je crée l'ontologie au format OWL à partir du fichier XML. Cette ontologie est éditée grâce à l'API du logiciel Protege.

Convertir un thesaurus en une ontologie implique de transformer les relations entre les termes. En effet dans une ontologie les concepts sont hiérarchisés grâce à une relation de spécialisation, les autres relations correspondent à des propriétés nommées *ObjectProperty* dans Protege.

Dans ce fichier les propriétés *Catégorie* sont des superclasses des concepts, tout comme les propriétés *Générique* et celles du type *Spécifique* sont des sous-classes des concepts décrits.

Les trois autres types de propriétés ont été plus difficiles à transposer dans l'ontologie, car ils n'ont pas de correspondance directe. Pour représenter ces propriétés il faut utiliser les

ObjectProperty, ce sont des propriétés qui permettent de relier les individus de classes différentes. Nous n'avons pas d'individus mais que des classes, nous allons quand même nous servir de ces *ObjectProperty* pour rendre compte des relations de type *Associé à*, *Employer plutôt* et *Employé pour* du thesaurus.

Employer plutôt est traduit par *use*, c'est une relation de type norme linguistique, c'est-à-dire qu'il est conseillé d'utiliser le concept qui porte cette propriété plutôt que le concept décrit. Son contraire est *Employé pour*, que l'on traduit par la relation *isusedfor*, le concept décrit doit être utilisé pour parler du concept qui suit cette propriété.

Associé à semble relié des concepts du même domaine, mais que signifie domaine sachant que chaque concept est déjà classé par catégorie ? Le type de cette relation est dur à discerner, nous considérerons que les concepts reliés par elle sont utilisés dans des domaines connexes. Avec le langage OWL ceci correspond à définir des classes disjointes.

3.6. Résultat

Après l'exécution des trois programmes JAVA vus précédemment on obtient l'ontologie. La figure 21 en présente un extrait. Elle ne correspond pas entièrement à l'ontologie décrite en 3.2. En effet mes connaissances en JAVA et en OWL et le manque de temps ne m'ont pas permis de réaliser l'ontologie correctement. Seules les relations de spécialisation et de généralisation du thesaurus sont présentes. Les autres relations ont été transposées en *ObjectProperty*.

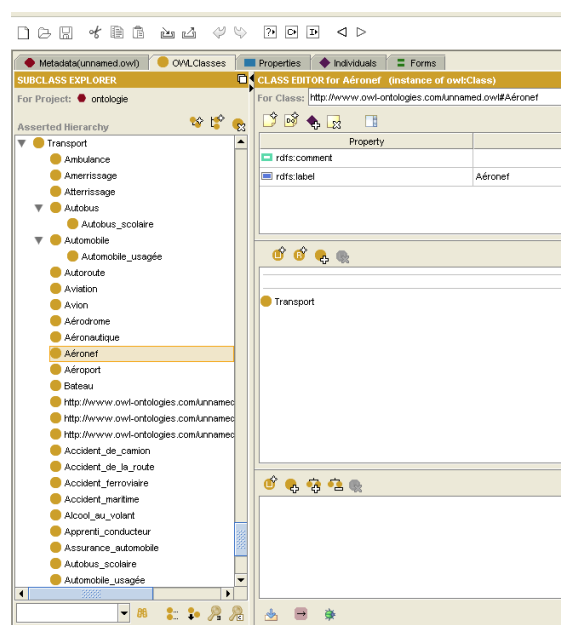


Figure 21 Extrait de l'ontologie obtenue

Bilan personnel

La mission qui m'avait été confiée a été réalisée entièrement. En effet les chercheurs du COGIT attendaient de moi que je dresse un état de l'art des ontologies d'objets géographiques permettant d'avoir une meilleure connaissance de celles-ci, ainsi que de situer le projet GEONTO par rapport à l'existant. J'ai dû commencer mon travail sans savoir quelles étaient les informations importantes à sélectionner, cela m'a ralenti dans la progression de mon travail. De ce fait je n'ai pas pu analyser toutes les ontologies répertoriées. Par manque de temps je n'ai pas pu contacter directement les organismes pouvant me renseigner, j'ai uniquement traité les données issues du Web.

La durée du stage ne m'a pas permis d'aborder les aspects techniques de ce domaine. Mon travail consistait d'avantage à de la recherche d'informations plus qu'à une application technique de mes connaissances.

L'apprentissage du langage de programmation JAVA m'a été très bénéfique. Comme ce langage était nouveau pour moi mon travail n'a pas pu aboutir à une version complète du programme. Cette familiarisation avec le langage m'a donné l'envie d'en savoir plus pour pouvoir mener à terme la réalisation de ce programme.

Il s'agissait de mon premier stage et ce fut très enrichissant pour moi de pouvoir le réaliser dans un laboratoire de recherche. J'ai ainsi pu voir concrètement ce qu'était un laboratoire de recherche, et changer les aprioris que j'avais concernant le monde de la recherche. En effet j'ai pu échanger avec des chercheurs et des thésards à propos de différents sujets d'études. J'ai également assisté à des réunions de laboratoire, et ainsi suivi la préparation d'articles et d'une soutenance de thèse dans le domaine des sciences de l'information géographique (SIG).

Effectuer mon stage dans le laboratoire COGIT de l'IGN m'a permis de découvrir les problématiques d'un domaine de recherche que je ne connaissais pas, dont certaines sont dans le domaine de la terminologie et d'autre du traitement automatique des langues.

Conclusion

Après deux mois de stage l'état de l'art était terminé, et le programme fonctionnait en partie. J'ai trouvé 49 ontologies provenant de projets différents. Dans tous les projets pour lesquels j'ai pu avoir des renseignements, aucun n'est semblable au projet GEONTO. Le contenu de certaines ontologies se rapproche de celui de l'ontologie de l'IGN, mais l'objectif pour lequel elles ont été construites est toujours différent de celui de GEONTO. Le premier objectif de mon stage a été réalisé puisque l'état de l'art conçu permet de mieux situer le projet GEONTO par rapport aux ontologies existantes. Pour réaliser entièrement cet objectif j'ai fourni à l'IGN un document rassemblant l'état de l'art, le tableau de tri des ontologies et les fiches de renseignements (cf. document annexe).

Le deuxième objectif du stage qui était l'écriture d'un programme convertissant un thesaurus en une ontologie, n'a pas été complètement réalisé. J'obtiens une ontologie au format OWL qui contient une partie des concepts du thesaurus, ils sont reliés entre eux par des relations de spécialisation.

Même si le deuxième objectif n'a pas été entièrement réalisé, sa réalisation m'a permis de découvrir le langage de programmation JAVA, je complète ainsi ma formation en programmation. J'ai aussi découvert un peu plus en détail le langage OWL avec lequel était structurée l'ontologie obtenue.

Le stage m'aura permis d'utiliser mes connaissances théoriques en terminologie et en science de l'information. Il m'a surtout servi à acquérir des connaissances sur les ontologies. Ce type de structuration est de plus en plus utilisé dans le domaine de la recherche d'information et en particulier pour le Web sémantique, domaine dans lequel je souhaite travailler à la fin de ma formation.

Pour continuer l'état de l'art il faudrait maintenant rechercher sur le Web des projets ayant conduit à la création d'ontologies, même si l'ontologie n'est pas visible. Et pour améliorer le programme il faudrait travailler sur l'écriture des relations en OWL et ensuite pouvoir l'adapter à d'autres thesaurus.

Bibliographie

Brade J., 2005, « Mutualisation de données et de connaissances pour la Gestion Intégrée des Zones Côtières. Application au projet SYSCOLAG. », 287p,

Charlet, J.; Bachimont, B. & Troncy, R., Jean Charlet, P. L. & C. R. (ed.), Web sémantique - Rapport Final, Ontologies pour le Web sémantique, Action spécifique 32 CNRS/STIC, 2003, 43 - 63

Agarwal P., Ontological considerations in GIScience. International Journal of Geographic Information Science, vol 19, n° 5, pages 501-536, Taylor & Francis, 2005.

Cullot N., Parent C., Spaccapietra S., et Vangenot C., Des SIG aux ontologies géographiques. Revue internationale de Géomatique, vol 13, n° 3, pages 285-306, Lavoisier, 2003.

Laurens F., « Construction d'une ontologie à partir de textes en langage naturel », rapport de stage, 2006.

Abadie N., Mustière S., « Constitution d'une taxonomie géographique à partir des spécifications de bases de données », article Sageo, 2008.

Roussey C., Kang M., « Les Ontologies et leurs Applications en Agriculture », congrès Inforsid, 2008.

Roussey C., Laurini R., Beaulieu C., Tardy Y., Zimmermann M., « Le projet Towntology. Un retour pour la construction d'une ontologie urbaine », in *Revue internationale de Géomatique*, vol. 14, 2004.