Lot 1 : Construction et enrichissement d'ontologies

# **Lot**	1	**Nom du lot**	Construction et enrichissement d'ontologies
Responsable	IRIT	**Participants**	IRIT, LIUPPA, LRI, COGIT

Résumé :
L'automatisation de la construction d'ontologies est un verrou scientifique pour beaucoup d'applications. Nous proposons une solution consistant, dans un premier temps, à construire un premier noyau d'ontologie en mettant en œuvre les techniques les plus adaptées au type de corpus disponible (sous lots 1.1 et 1.2). Conscient des limites de toutes les techniques du fait de leur sensibilité au bruit, le résultat de ce processus d'extraction et de structuration de terminologie sera ensuite réorganisé (sous-lot 1.3) par confrontation à une taxonomie *de référence* de bonne qualité. En effet, le COGIT dispose aujourd'hui d'une taxonomie construite au départ automatiquement, qui a ensuite été modifiée par des experts du domaine et qui peut, aujourd'hui, être considérée comme une *référence* pour le domaine dans le cadre de l'étude. Il s'agira aussi d'enrichir des ontologies de l'IGN (sous-lot 1.2). Des experts du domaine valideront toutes les ontologies construites. Cela sera réalisé en interne au projet par le COGIT et le LIUPPA, spécialistes d'information géographique. Des experts externes seront aussi sollicités autant que possible, comme les responsables des produits de l'IGN, le CNIG (Conseil National de l'Information Géographique) ou l'AFIGEO (Association Française pour l'Information Géographique).

Sous-lot 1.1 : Mise au point d'outils d'extraction de concepts et de relations

Limites actuelles de l'état de l'art

Aujourd'hui, la construction d'ontologies à partir de textes s'appuie sur des logiciels de traitement du langage naturel et sur des ressources combinant lexique et concepts. L'extraction de concepts fait appel à des extracteurs de termes.
Deux approches différentes existent pour déceler les relations entre concepts. La première est basée sur la définition de patrons lexico-syntaxiques qui établissent une relation entre concepts du domaine. Ces relations ne sont décelées que lorsque les concepts appartiennent à la même phrase. Deux courants complémentaires se sont développés. Dans une tradition linguistique, des patrons relatifs aux relations hiérarchiques (hyperonymie, définition, méronymie) ou de synonymie, ont été capitalisés avec l'espoir de pouvoir les réutiliser sur tout type de textes. L'expérience montre que ces patrons sont plus ou moins pertinents et doivent toujours être adaptés. Dans la lignée de l'extraction d'information, de nouveaux patrons sont redéfinis pour repérer des relations spécifiques au domaine étudié.
La deuxième approche, dite statistique, décèle des relations entre concepts (co-occurrences de termes, etc.) sans toutefois interpréter ces relations.

Innovations apportées

Les travaux liés à l'extraction de relations ne prennent généralement pas en compte la structure du document ni sa mise en forme. Par ailleurs, les relations syntagmatiques ne sont pas toutes décelables par les techniques de recherche classique telles que l'approche par patron.
Les innovations porteront sur :

- la prise en compte du type de document dans l'extraction de relations et de termes,
- l'exploitation de ressources lexicales pour le repérage de concepts,
- la recherche des paramètres d'une relation en exploitant les relations argumentatives (sujet, objet) autour des verbes,
- la recherche de relations exprimées à l'aide de plusieurs phrases,
- la prise en compte de la disposition matérielle et de la ponctuation dans les patrons,
- l'exploitation de la complémentarité de textes de genres différents (spécifications versus textes grand public).

Ces innovations portent essentiellement sur le repérage de relations. Ce choix justifie d'adopter une démarche automatique pour l'extraction des termes, traitée de manière classique, et supervisée pour le repérage des relations, où les innovations sont nombreuses. Une approche basée sur l'apprentissage des patrons de relation pourra être envisagée en perspective. Elle ne sera mise en œuvre en priorité durant le projet que si elle s'avère simple et qu'elle puisse s'appuyer sur des travaux analogues.

Description
Pour l'extraction des concepts, il s'agira d'abord d'évaluer les ressources lexicales et ontologiques disponibles. On utilisera éventuellement des règles de nommage ou on s'appuiera sur des éléments linguistiques comme la morphologie des termes. De plus, on évaluera la possibilité d'utiliser des outils performants et éprouvés (statistiques ou syntaxiques) pour l'extraction de termes et pour la structuration de l'ontologie (résultats du projet RNTL DAFOE).

Pour l'identification de relations entre concepts à partir de documents géographiques « grand public », les techniques utilisées (développées au LIUPPA) s'appuieront sur des méthodes de marquage des dépendances grammatico-sémantiques locales à partir de noyaux composés d'Entités Nommées Géographiques (ENG) (résultats du projet Geosem). Le travail sera réalisé sur un corpus de récits de voyages mis à disposition par la MIDR[1]. Il vise à obtenir un inventaire des relations existantes dans un tel corpus entre les termes topographiques du lexique de la taxonomie de référence produite par le COGIT et les ENG d'un territoire particulier (les Pyrénées).

Pour l'identification de relations entre concepts à partir de textes de spécifications, l'équipe IC3 a utilisé les outils d'extraction de relations Caméléon (développé par IC3) et Gate. Une base de patrons a été encodée et est opérationnelle pour la recherche de définitions et de relations lexicales. Ce lot permettra de définir des patrons spécifiques aux textes du COGIT et d'intégrer dans la recherche de patrons de nouveaux traitements destinés à résoudre les problèmes cités ci-dessus. En plus de l'approche par patrons, nous développerons des modules s'appuyant sur d'autres éléments textuels :
a. La structuration du texte aidera à l'identification des relations. En exploitant les données contenues dans les champs tels que *Définition* ou *Regroupement*, des relations de type 'est-un' ou 'partie-de' peuvent, par exemple, être identifiées entre concepts.
b. Après une analyse syntaxique de la phrase, l'exploitation des relations argumentatives (sujet, objet) de concepts autour d'un verbe peut aider à identifier de nouvelles relations, en se référant par exemple aux classes WordNet des verbes. Une chaîne de traitement sera développée qui inclura donc l'analyse syntaxique et l'accès à des ressources externes.

Planning d'activité prévisionnel

N°	Description	Date de fin de tâche
1.1	Définition des patrons lexico-syntaxiques à partir des textes, et détection d'autres relations à partir d'un corpus d'apprentissage.	T0 + 6
1.2	Définition d'un ensemble de patrons géographiques. Spécification de techniques de repérage de concepts et de relations, basées sur le marquage grammatico-sémantique.	T0 + 12
1.3	Conception d'une chaîne de traitement pour appliquer les patrons lexico-syntaxiques.	T0 + 18
1.4	Tests et validation des résultats. Rapport final du lot.	T0 + 24
		T0 + 30
		T0 + 36

<u>Sous-lot 1.2 : Enrichissement d'une ontologie existante à partir de textes à l'aide d'outils d'extraction et à partir de ressources lexicales</u>

Limites actuelles de l'état de l'art
Les outils de construction automatique d'ontologies génèrent des hiérarchies de concepts pour lesquelles les relations ne sont pas toujours bien définies. Certaines relations correspondent à une relation de subsomption, d'autres à une relation de composition ou d'agrégation. Il arrive même que des liens soient associés à des relations de synonymie entre termes.

1 MIDR: Médiathèque Intercommunale à Dimension Régionale de Pau.

Innovation apportée
L'innovation portera sur la combinaison des patrons définis au sous-lot 1.1 pour typer les relations, et de l'exploitation de techniques d'alignement à partir de ressources lexicales (ressource Mémodata, équivalent de WordNet pour le français). D'un point de vue géographique, le typage des relations permettra d'obtenir une ontologie de meilleure qualité car plus précise.

Description
Une taxonomie/ontologie est actuellement disponible au COGIT. C'est une ontologie de type thesaurus de bonne qualité mais elle présente un certain nombre de défauts en tant qu'ontologie descriptive. En particulier, les relations hiérarchiques ont une sémantique mal définie. Nous nous proposons de tester l'exploitation de plusieurs types de ressources pour la corriger et la compléter, et définir ainsi une ontologie : des thésaurus, des bases de données lexicales, ou encore les textes qui ont constitué le corpus à partir duquel la taxonomie a été construite.
Différentes techniques seront utilisées dans ce processus de correction et d'enrichissement :

- extraction de concepts à partir des lexiques et des textes,
- extraction de relations entre concepts, par application de patrons et de ressources lexicales du domaine tirées de textes "grand public", pour produire un thésaurus (LIUPPA et IRIT),
- extraction de relations entre concepts et de concepts, par application de patrons sur des textes de spécifications (IRIT et LIUPPA),
- confrontation du modèle obtenu à une base de données lexicale par une technique d'alignement d'ontologies (LRI).

Un thésaurus sera construit à partir des textes grand public, et dans ce cas les patrons issus de l'exploitation des relations argumentatives autour d'une forme prédicative seront appliqués. Il sera intégré à l'ontologie existante du COGIT, fournissant de nouveaux concepts et des termes associés, ainsi que leur organisation hiérarchique.
Ensuite, d'autres types de relations seront recherchés dans les textes de spécification à l'aide des patrons adaptés à ce corpus et définis au 1.1. Ces relations viendront enrichir l'ontologie, qui sera ensuite révisée par des experts du domaine.

Enfin, l'exploitation d'une base de données lexicale prendra appui sur l'expérience du LRI en matière d'alignement d'ontologies. Une des techniques de l'environnement d'alignement existant exploite en effet une base de données lexicale (ressource de Mémodata ou EuroWordNet pour le français) d'une manière originale. La base de données lexicale n'est pas considérée uniquement comme un moyen de fournir des synonymes, hyperonymes, hyponymes. Elle fournit un support structurel exploité pour détecter des relations entre concepts. Il serait intéressant de tester cette technique dans le contexte du domaine géographique et de la compléter pour, non seulement savoir que 2 concepts sont liés, mais également pour identifier la nature de ce lien.

Planning d'activité prévisionnel

N°	Description	Date de fin de tâche
		T0 + 6
1.5	Premiers tests d'extraction de relations en utilisant chaque technique séparément sur les taxonomies du COGIT. Construction d'un premier thésaurus à partir des textes « grand public »	T0 + 12
1.6	Version 1 du module logiciel de construction Version 1 de l'ontologie enrichie Tests combinant différentes techniques sur les taxonomies du COGIT	T0 + 18
1.7	Tests sur les ontologies obtenues en résultat du sous-lot 1.1. Evaluation de la robustesse du prototype développé. Validation des résultats d'un point de vue géographique.	T0 + 24
1.8	Version 2.0 de l'ontologie enrichie par les différentes techniques Rapport final de présentation des techniques de construction et de l'ontologie produite	T0 + 30

		T0 + 36

Sous-lot 1.3 : Restructuration d'ontologie

Limites actuelles de l'état de l'art

Isolées, les techniques de construction automatique d'ontologie à partir de corpus ou les techniques d'alignement sont limitées. Les textes servant à l'élaboration automatique d'une ontologie ne présentent pas nécessairement toutes les définitions et concepts utiles à l'élaboration d'une ontologie bien structurée. Par ailleurs, des sous-arbres indépendants peuvent être générés, qu'il convient de relier et de niveler en définissant le bon niveau d'abstraction pour chaque concept. L'organisation des concepts au sein de l'ontologie reflète alors davantage le mode d'utilisation des connaissances que leur essence ontologique..

Innovation apportée

L'innovation portera sur le fait de combiner une approche de construction d'ontologie à partir de textes, avec des techniques automatiques de comparaison d'ontologies (comparer l'ontologie obtenue à l'issue du processus de construction automatique avec une ontologie *de référence*). Cette composition de processus devrait permettre d'obtenir une ontologie de qualité de façon totalement automatique.

Description

Il s'agira de concevoir des outils automatisés d'aide à la réorganisation d'une ontologie, pour identifier les redondances (sous arbres identiques dans la hiérarchie), les sous-classes à regrouper, de modifier la structuration, identifier les concepts équivalents, les homonymes (sous-classes différentes), etc. D'un point de vue géographique, il s'agira d'améliorer la qualité de l'ontologie et, de ce fait de diminuer les interventions manuelles.

Les restructurations à effectuer seront fortement dépendantes de la méthode de construction de l'ontologie (des techniques utilisées) et de la nature des textes sur lesquelles les techniques d'extraction et de structuration de la terminologie ont été appliquées. Là encore, disposer d'une ontologie *de référence* est un atout pour proposer des réorganisations automatiques via des techniques d'alignement.

Planning d'activité prévisionnel

N°	Description	Date de fin de tâche
		T0 + 6
		T0 + 12
		T0 + 18
1.9	Analyse des besoins sur la base des ontologies obtenues en sortie du lot 1.2	T0 + 24
1.10	Mise au point d'une méthodologie de restructuration	T0 + 30
1.11	Module logiciel de restructuration automatique. Tests. Evaluation et validation d'un point de vue géographique.	T0 + 36