Lot 3 : Exploitation des ontologies créées

# **Lot**	3	**Nom du lot**	Exploitation des ontologies créées
Responsable	COGIT	**Participants**	COGIT, LIUPPA

Résumé : Les ontologies permettent la spécification de connaissances agréées par une communauté de personnes grâce à un langage formel non ambigu. Elles ne peuvent servir directement sous cette forme à l'utilisateur final, en revanche elles acquièrent toute leur importance une fois intégrée, au sein d'une méthodologie complète. Dans ce lot, nous exploiterons et expérimenterons cette intégration dans deux cas d'utilisation : l'indexation de contenu (sous lot 3.1) et l'intégration de bases de donnée géographique (sous lot 3.2). Dans le cas de l'indexation, les représentations conceptuelles (déduites grâce à l'ontologie) et géométriques (obtenues grâce à l'association entre termes de l'ontologie et champ d'un base de donnée géographique) vont permettre de construire des index se basant sur la typologie des objets identifiés et leurs relations topographiques. Dans le cadre de l'intégration de bases de données, les ontologies alignées vont être exploitées pour apparier entre eux des schémas, qui découlent chacun de points de vue et donc d'ontologies différentes. Il s'agit d'étudier dans quelle mesure l'alignement de ces ontologies peut être utilisé pour aller jusqu'à un appariement fin des schémas des bases.

Sous-lot 3.1 : Indexation automatique du contenu de documents

Limites actuelles de l'état de l'art

Au delà du besoin croissant de partage d'informations sur le Web qui passe par la structuration des ressources mises à disposition, le problème adressé ici correspond également aux nouveaux besoins de valorisation des fonds documentaires patrimoniaux suscités par l'importante politique de numérisation mise en oeuvre par les différentes instances de conservation (Archives Régionales, Musées, Médiathèques...) des collections documentaires territorialisées. Une part non négligeable de l'information contenue dans ces documents numériques fait référence de manière plus ou moins explicite à des entités géographiques. Or la plupart des systèmes permettant la gestion et la consultation de documents en ligne propose une indexation reposant sur l'exploitation de méta données produites manuellement combinées à des méthodes de fouille plein texte basées essentiellement sur des méthodes statistiques. Seules quelques propositions préindustrielles proposent actuellement la prise en compte de certaines relations sémantiques (comme, par exemple, les relations spatiales ou temporelles dans le projet SPIRIT (op.cit.)). L'indexation géographique des contenus, quant à elle, se limite encore à l'association de l'Entité Nommée Géographique à une géoréférence. Les prises en compte de concepts spatiaux tels que « au Nord de », « entre », « à proximité de » ou « autour de »... voire de composition ou de dépendance plus thématiques (correspondant à tel ou tel phénomène, par exemple topographique « les pics les plus élevés » ou « les rivières à fort débit ») restent un enjeu scientifique, comme en attestent de nombreux workshops[1].

Innovation apportée

L'innovation portera sur les techniques de construction d'index géographiques complexes capables de répondre à des requêtes d'utilisateurs (composées en texte libre ou en multi-modalité) à connotation géographique forte (spatiale et topographique).

Description

Dans le cadre de ce projet il s'agira de construire des index pouvant avoir une finesse plus ou moins importante selon les ressources qui seront mises à contribution au moment de leur construction. Nous utiliserons, bien évidemment, des ressources de type base de données géographiques, geowebservice, gazeteers, etc. et étudierons les diverses possibilités d'exploitation de ressources de type ontologie telles qu'elles auront été produites par les lots précédemment décrits et notamment le lot 1.3.

Dans le cadre de travaux récemment réalisés, l'équipe-projet Desi du LIUPPA propose, en effet, une plateforme expérimentale (basée sur des services web) permettant de construire des index géographiques à partir de méthodes originales d'extraction d'information dans les textes et d'enrichissement par des ressources de type BD géographiques. La première phase est une phase qui extrait des syntagmes nominaux étendus « candidats » à constituer les Entités Nommées

1 Par exemple les ateliers GIR http://www.geo.unizh.ch/~rsp/gir06/ qui se tiennent conjointement lors des conférences ACM- SIGIR.

Géographiques (ENG). La deuxième phase consiste dans un premier temps à valider grâce à divers ressources les ENG « candidates » puis à leur associer une représentation sémantique et si possible la représentation géométrique la plus adéquate. Nous proposons d'étendre cette plateforme et d'y intégrer de nouvelles procédures de validation/représentation de ENG basées sur des ressources ontologiques.

Planning d'activité prévisionnel

N°	Description	Date de fin de tâche
		T0 + 6
3.1	Etude des apports sémantiques de l'ontologie topographique dans le cadre d'une indexation géographique de contenus textuels	T0 + 12
3.2	Mise au point de la méthodologie pour l'exploitation de l'ontologie à des fins d'indexation	T0 + 18
3.3	Conception et implémentation de nouveaux Services Web intégrables dans la plateforme du LIUPPA afin d'exploiter des ressources de type ontologie	T0 + 24
3.4	Tests d'évaluation des extensions développées. Construction de nouveaux index sur un corpus d'étude à des fins d'expérimentation	T0 + 30
3.5	Finalisation des extensions après expérimentation et publication des nouveaux services web.	T0 + 36

Sous-lot 3.2 : Intégration, accès aux schémas des bases de données et évaluation

Limites actuelles de l'état de l'art

L'intégration de schéma de bases de données géographiques repose sur une compréhension fine de la sémantique des éléments des schémas. Les spécifications des bases sont des sources de connaissances riches décrivant de cette sémantique. Cependant, les rares travaux existants en appariement de schémas dans le domaine géographique n'exploitent pas cette source de connaissances. Il en est de même pour les travaux sur la recherche d'information dans ces bases, qui s'appuient sur une description des schémas et des métadonnées spatiales, mais pas sur les spécifications.

Innovations apportées

L'approche globale d'appariement et d'analyse des bases de données à partir d'une formalisation des spécifications a été initiée au COGIT dans la thèse de Nils Gesbert soutenue en 2005 [Gesbert 2005]. Lors de cette thèse, un modèle de formalisation des spécifications a été proposé, qui suppose l'existence d'ontologies géographiques. Cependant, faute de telles ontologies et de méthode d'instanciation automatique du modèle proposé, celui-ci n'a été instancié que sur des cas particuliers et l'approche n'a pas pu être conduite jusqu'à une réelle intégration des bases. Le but de ces recherches est donc de dépasser ces limites et d'intégrer, au sein d'une méthodologie complète, l'exploitation d'ontologies pour formaliser finement les spécifications et enfin une réelle description et intégration des bases de données qu'elles décrivent.

Description

Ces travaux se situent dans la suite de travaux du COGIT sur l'intégration de bases de données géographiques. Ils seront réalisés en grande partie à travers une thèse financée par l'IGN sur " l'intégration des bases de données à partir de la formalisation de leur spécifications ". La thèse supposera l'existante d'ontologies du domaine riches et alignées, comme celles issues des lots 1 et 2. Dans un premier temps, elle s'attachera à étudier comme exploiter ces ontologies pour instancier, grâce à des méthodes de traitement automatique du langage, un modèle formel de description des spécifications. Si le lot 2 vise à déduire une ontologie à partir de textes tels que les spécifications (par exemple, extraire les concepts de " route ", " rivière "…), la thèse vise à formaliser le lien entre la base de données et l'ontologie (par exemple, représenter le fait que toutes les routes de plus de 100 mètres de long sont représentées dans la base). Dans un deuxième temps, il s'agira d'étudier comment ces descriptions formelles de deux bases de données peuvent être exploitées pour les intégrer.

La thèse sera complétée par des travaux visant à exploiter les ontologies pour mettre au point un portail

d'accès aux schémas des bases ayant pour point d'entrée les ontologies. Cette partie s'appuie sur l'idée que l'alignement d'ontologies permet de relier les termes utilisés par les utilisateurs avec leur propre point de vue lors de requêtes, à ceux plus techniques utilisés par les concepteurs pour décrire leurs bases de données.

Planning d'activité prévisionnel :

N°	Description	Date de fin de tâche
3.6	Démarrage de la thèse, et état de l'art sur le domaine de l'intégration de schémas.	T0 + 6
3.7	Mise au point de techniques de traitement du langage pour analyser et formaliser les spécifications.	T0 + 12
3.8	Test des techniques sur des données réelles et identification des limites.	T0 + 18
3.9	Définition et mise en œuvre d'un portail d'accès aux schémas ayant pour point d'entrée les ontologies.	T0 + 24
3.10	Mise au point d'une méthodologie et d'un module logiciel permettant l'exploitation des spécifications formelles pour apparier des schémas de données.	T0 + 30
3.11	Finalisation de la thèse.	T0 + 36

Sous-lot 3.3 : Mise à disposition des ontologies réalisées

Limites actuelles de l'état de l'art
Les lots 1 et 2 doivent permettre de constituer une ou plusieurs ontologies riches, selon que l'on juge pertinent, à l'issue du sous lot 2.3, de les fusionner ou non. Peu d'ontologies se concentrant sur la description de l'espace topographique sont disponibles actuellement, du moins si on fait abstraction de celles focalisant sur un unique aspect. Les seules qui existent à notre connaissance sont anglophones et moins riches que celles attendues en résultat des lots 1 et 2. Par ailleurs, les résultats du lot 2.3 nous permettront de statuer sur la réelle originalité des ontologies issues du projet par rapport à des ontologies externes.

Innovation apportée
Une ontologie topographique riche sera mise à disposition de la communauté scientifique. Elle sera en bilingue français/anglais. Les groupes de travail associé à la directive européenne INSPIRE ont exprimé le besoin de disposer de telles ontologies.

Description
Ce sous lot ne constitue pas un sujet de recherche en soi mais une valorisation des résultats obtenus dans le projet. Les ontologies résultant des lots 1 et 2 seront traduites, contrôlées, mises en ligne au format OWL. Un moteur adapté sera réalisé pour permettre de les interroger et les parcourir. Selon les résultats du sous-lot 2.3 qui mettront en lumière les points communs et différences entre les ontologies du projet et des ontologies externes, des partenariats pourront être initiés avec des organisations responsables de ces dernières pour aller vers une mise en commun des ontologies. Les ontologies réalisées pourront être également soumises à l'OGC (Open Geospatial Consortium), organisme normalisateur en matière d'information géographique.

Planning d'activité prévisionnel

N°	Description	Date de fin de tâche
		T0 + 6
		T0 + 12
		T0 + 18
		T0 + 24
3.12	Analyse des résultats intermédiaires des lots 1 et 2 pour statuer sur la	T0 + 30

	pertinence de diffuser plusieurs ontologies ou une unique ontologie fusionnée. Fusion éventuelle des ontologies réalisées, traduction des termes associés aux concepts retenus, et finalisation interactive des ontologies. Initiation de collaborations avec d'autres organismes à l'origine d'ontologies géographiques.	
3.13	Conception du site web pérenne réceptacle des ontologies à diffuser, et conception d'un moteur en ligne pour leur parcours et interrogation.	T0 + 36