

**Masse de Données et Connaissances
Appel à projets 2007 (ANR-07-MDCO)
Annexe technique**

Acronyme du projet : GEONTO

1	Introduction	2
2	Contexte et état de l'art.....	4
3	Partenaires	6
3.1	Présentation générale du partenariat.....	6
3.2	Partenaires.....	6
4	Organisation et management du projet	9
5	Structure du projet – Description des sous-projets.....	10
5.1	Lot 0 : Gestion de projet.....	10
5.2	Lot 1 : Construction et enrichissement d'ontologies	10
5.3	Lot 2: Appariement d'ontologies hétérogènes	14
5.4	Lot 3 : Exploitation des ontologies créées	18
6	Liste des livrables	22
7	Résultats escomptés – perspectives	23
7.1	Retombées scientifiques et techniques	23
8	Propriété intellectuelle	24
	Annexe 1	24
	Annexe 2	27

1 Introduction

Ce projet vise l'interopérabilité de diverses données géographiques hétérogènes. Deux objectifs applicatifs sont tout particulièrement visés. Le premier est l'intégration de bases de données géographiques hétérogènes, opération qui est nécessaire pour faciliter leur gestion (saisie, maintenance, mise à jour) et pour permettre des analyses combinant ces différentes bases. Cette tâche passe par la mise en correspondance des schémas des bases qui est complexe du fait que chaque base a été conçue indépendamment des autres et que les choix de modélisation effectués sont fonction des besoins de chaque concepteur. Il en résulte une grande hétérogénéité des schémas, tant du point de vue de leur structure, que des noms des entités, des attributs, des domaines de définition associés aux valeurs des attributs. Le deuxième objectif applicatif est l'interrogation d'une collection importante de documents textuels plus variés et destinés à un plus grand public que les bases de données précédemment mentionnées. Ces documents font référence à divers lieux nommés qui, s'ils sont identifiés, sont un point d'entrée essentiel sur lequel peuvent se baser de nombreuses requêtes. Ces documents s'appuient aussi sur un vocabulaire géographique qui, s'il est explicite et formalisé, peut servir à produire de puissants index sur lesquels peuvent s'appuyer des langages d'interrogations expressifs.

Dans ce contexte, une approche de plus en plus privilégiée pour intégrer des données diverses, autant dans le monde des bases de données que dans celui de la recherche d'information et des systèmes d'information géographiques, est d'appuyer l'intégration sur une ontologie du domaine concerné. Les ontologies jouent un rôle clé en intégration de sources d'information multiples et hétérogènes. Une ontologie est un modèle structuré des objets d'un domaine d'application, une vue sur ce domaine, une conceptualisation définissant des concepts, des propriétés, des relations. Son rôle est double. D'une part, elle précise le sens des concepts d'un domaine en étant le reflet d'un certain consensus au sein d'une communauté. D'autre part, elle fournit une sémantique formelle. Les concepts ne sont pas vus uniquement comme des notions sémantiques. Ils vérifient des propriétés qui ont une définition formelle. Le langage de représentation des connaissances utilisé doit permettre des traitements automatiques. Dans le contexte de l'intégration, ces représentations peuvent aider à comprendre et interpréter des descriptions hétérogènes de contenus relatifs à un même domaine pour ensuite pouvoir plus facilement les mettre en relation. C'est la voie qui est adoptée dans le cadre de ce projet.

Une première question se pose alors : comment constituer les ontologies ? Pour cela, nous proposons d'analyser divers documents textuels grâce aux techniques de traitement automatique du langage naturel. Une particularité du domaine des bases de données géographiques consiste en l'existence de documents textuels correspondant aux spécifications des bases existantes. Ce projet repose alors d'une part sur l'hypothèse que l'ontologie décrivant le domaine d'un schéma peut être construite par exploitation de toute la richesse des spécifications textuelles de ce schéma. D'autre part, les textes grand public décrivant des lieux géographiques s'appuient aussi sur un vocabulaire spécifique qui, à notre avis, peut être identifié automatiquement et est révélateur d'un point de vue géographique moins technique.

Une deuxième question se pose ensuite : les diverses ontologies réalisées, qu'elles proviennent de plusieurs spécifications de bases de données ou de documents moins techniques, peuvent-elles être alignées ou fusionnées? Et dans quelle mesure reflètent-elles des points de vue différents? Ces dernières années, de nombreux travaux ont été réalisés en alignement d'ontologies. L'idée consiste à étudier comment tirer parti au maximum de ces travaux et comment les adapter à notre cas d'étude.

La réalisation de ce projet passe donc par la réalisation de 3 sous objectifs :

- construire des ontologies associées à des bases de données par exploitation de leurs spécifications (documents textuels), ou associées à un corpus de documents géographiques moins techniques.
- aligner les ontologies obtenues et étudier leurs différences.
- apparier les schémas des bases de données via les ontologies, et développer un moteur de

recherche d'information dans une base de documents via ces mêmes ontologies.

Les techniques préconisées pour l'appariement de schémas de bases de données comme pour la recherche d'information devront permettre un déploiement à grande échelle, ce qui requière des traitements les plus automatiques possibles et pose des problèmes du fait de certains verrous technologiques :

- la construction d'ontologies comme support à l'appariement de bases de données ou comme entrée à la recherche d'informations, les données pouvant être nombreuses, sémantiquement hétérogènes et réparties dans des bases multiples. La construction manuelle d'une ontologie, même assistée par des outils conviviaux, est un travail de modélisation long et difficile. L'objectif d'une partie du projet est de tirer parti des spécifications associées aux bases de données pour construire les ontologies décrivant le domaine associé. Il s'agira d'étudier et de mettre en œuvre différentes approches permettant d'automatiser la construction d'ontologies à partir de documents textuels.
- l'alignement d'ontologies hétérogènes. L'application de ces techniques à grande échelle n'a, pour l'instant, pas été au centre des recherches réalisées. Dans notre contexte, selon les techniques et les sources de connaissances utilisées, les ontologies construites à partir d'analyse de textes sont plus ou moins riches, autant en termes de contenu que de structure. Elles peuvent prendre la forme d'une hiérarchie de termes avec très peu de niveaux de structuration, d'une taxonomie de termes liés mais sans que la sémantique des relations soit clairement définie ou à l'inverse d'une ontologie fortement structurée et à la sémantique riche. Il ne s'agit donc plus de comparer des ontologies proches du point de vue du niveau de description et de structuration. Le passage à l'échelle nécessite d'élargir cette notion d'hétérogénéité pour aligner des ontologies très différentes structurellement et également par rapport à la précision avec laquelle les connaissances sont décrites. Elles peuvent avoir également divers degrés de pertinence et de qualité, en particulier selon le niveau d'interactivité humaine utilisé pour les constituer ou les corriger. De la qualité des ontologies dépendront la qualité des alignements et donc de celle de l'indexation des bases de données, il est donc important de prendre en compte l'ensemble de ces caractéristiques.

Les résultats attendus dans ce projet sont les suivants :

- De nouveaux outils d'extraction de concepts et de relations dans des textes, basés sur la définition de patrons, sur le repérage d'Entités Nommées Géographiques, sur l'exploitation de la structure des textes et des relations argumentatives,
- Des techniques et une méthodologie de création, d'enrichissement et de restructuration d'ontologie, combinant l'utilisation de patrons, de ressources lexicales externes, de techniques d'alignement d'ontologies,
- Une étude de l'efficacité des techniques d'alignement, l'adaptation de techniques actuelles et le développement de nouvelles techniques adaptées à la fusion efficace de deux ontologies hétérogènes de qualité moyenne (cas des ontologies réelles),
- Des techniques et une méthodologie de comparaison de deux ontologies reflétant des niveaux d'échelle spatiale différents ou issues de pays différents, afin d'étudier si ces ontologies reflètent de véritables points de vue différents ou si elles n'ont que des différences marginales et gagneraient donc à être fusionnées.
- Une ontologie du domaine de l'information géographique, et plus particulièrement de la description topographique du paysage,
- Des techniques et méthodologies exploitant une ontologie pour indexer automatiquement le contenu de documents diversifiés et pour apparier des schémas de bases de données hétérogènes.

2 Contexte et état de l'art

Avec d'un côté l'essor des techniques de l'information et, d'un autre côté, le développement des techniques de localisation spatiale, les données géographiques sont de plus en plus nombreuses et diverses. La gestion de cette diversité est un problème important qui se révèle en particulier à travers deux initiatives récentes et d'ampleur.

Au niveau national, la direction générale de la modernisation de l'Etat (DGME) a lancé un projet de portail de l'information géographique publique qui a pour objectif de "constituer un point d'entrée le plus large possible pour rechercher les principales données géographiques de l'Etat, de ses établissements publics et des collectivités territoriales, en connaître leurs caractéristiques et les moyens d'y accéder et de les visualiser et les co-visualiser". Ce portail se donne pour but d'être "ouvert et interopérable, permettant ainsi la fédération des données"¹.

Au niveau européen, la commission chargée de l'Environnement a initié la directive INSPIRE qui vient d'être adoptée et qui demande à mettre en place une infrastructure distribuée de données spatiales permettant "qu'il soit aisé de rechercher les données géographiques disponibles, d'évaluer leur adéquation au but poursuivi et de connaître les conditions applicables à leur utilisation [et] qu'il soit possible de combiner de manière cohérente des données géographiques tirées de différentes sources dans la Communauté et de les partager entre plusieurs utilisateurs et applications"².

Ces deux initiatives illustrent les besoins relatifs à la description et l'intégration cohérente de données géographiques, ce qui se révèle difficile en raison de la grande diversité de ces données, autant du point de vue de leur but que de leur niveau de détail.

Depuis plus de 10 ans, l'acquisition de connaissances et l'analyse sémantique de contenus à partir de textes connaît un nouvel essor avec l'arrivée de nouveaux outils mais surtout grâce à un renouvellement de la manière de poser le problème [Enjalbert & Gaio 2004] [Bourigault et al. 2001]. Les textes sont vus désormais comme des sources à exploiter par des logiciels de Traitement Automatique des Langues (TAL) au sein d'un processus supervisé par un analyste qui tient compte des objectifs de la modélisation. On ne cherche plus à reconstituer automatiquement les modes de compréhension du texte par un individu, mais à outiller le repérage, au sein des textes, de connaissances utiles pour la modélisation. La réflexion menée en France au sein des groupes TIA³, Action Spécifique « Corpus et Terminologies »⁴ souligne la complémentarité des études en linguistique de corpus, en terminologie, en ingénierie des connaissances, en apprentissage, en traitement automatique des langues et en Sciences de l'Information. Au plan international, de nombreux workshops⁵ traduisent la dynamique des recherches sur cette question, dans le cas particulier de la construction d'ontologies et du Web Sémantique. Ces travaux distinguent l'identification des concepts et de leurs propriétés ou relations, tâche appelée « ontology learning » même si elle n'est pas toujours complètement automatisée, de la recherche d'instances de concepts et d'instances de relations dans les textes, ce qui correspond à l'« ontology population » [Cimiano et al. 2004].

C'est le problème de la construction d'ontologies qui nous intéresse dans ce projet. Une gamme d'approches est aujourd'hui disponible, comme le montrent des ouvrages de synthèse [Buitelaar et al. 2005] [Cimiano 2006]. Certaines proposent des processus supervisés comme le système Text-to-Onto [Maedche & Staab 2000], d'autres cherchent à complètement l'automatiser [Han & Schulz 2002]. Les analyses des textes peuvent s'appuyer sur des traitements combinant pour la plupart des analyses statistiques et linguistiques pour identifier à partir de textes des concepts, leur organisation hiérarchique et des relations sémantiques propres au domaine [Cimiano 2006]. Les approches par patrons lexico-syntaxiques s'avèrent des outils fins et précis pour repérer des relations sémantiques lorsque les textes

¹ Charte du portail de l'information géographique publique, 21 juin 2006. www.geoportail.fr

² Projet commun de directive européenne INSPIRE, approuvé par le comité de conciliation le 17 janvier 2007.

³ Groupe de travail du GDR I3, <http://tia.loria.fr/>

⁴ <http://www.irit.fr/ASSTICCOT/>

⁵ Ontology Learning and Population – OLP ou Ontology Learning and Texts – OLT, par exemple.

présentent des formes rédigées et régulières. Ils sont d'autant plus efficaces qu'ils peuvent être adaptés aux formulations en cours dans le domaine étudié [Ciravegna et al. 2002].

Pour améliorer cette approche, il est indispensable de capitaliser des patrons par domaine et par type de relation, de disposer d'un environnement permettant de facilement adapter des patrons connus et d'en définir de nouveaux, et surtout d'étudier l'expression linguistique de types particuliers de relations sémantiques comme cela a été initié dans le domaine spécifique de la géographie avec le projet GeoSem⁶. Ces approches sont d'autant plus efficaces que des analyses linguistiques préalables ont été réalisées, et que les textes sont enrichis d'annotations syntaxiques ou sémantiques précises. Le fait de traiter au préalable des phénomènes linguistiques comme l'anaphore ou les ellipses serait par exemple une avancée pour repérer des relations exprimées à l'aide de plusieurs phrases. Ce projet vise à faire progresser ces questions en étudiant des types de textes particuliers dans le domaine de la géographie et en se restreignant aux concepts topographiques.

De nombreux travaux ont porté ces dernières années sur l'alignement d'ontologies. Une synthèse des techniques est présentée dans [Kalfoglou et Schorlemmer 2003] et [Shvaiko et Euzenat 2005]. Les techniques sont variées. Les résultats générés sont, en général, accompagnés d'un degré de confiance. Pour la plupart, ils ne sont donc pas sûrs à 100% et nécessitent une validation manuelle. L'application de ces techniques à grande échelle n'a, pour l'instant, pas été au centre des recherches réalisées. L'accent a été surtout mis sur l'hétérogénéité sémantique du point de vue du contenu : présence de concepts étiquetés différemment ou ontologies structurées différemment. En revanche, l'hétérogénéité des types de représentation n'a pas fait l'objet de recherches ciblées. Cela concerne l'hétérogénéité provenant du degré de précision des connaissances représentées qui peut être variable (une ontologie correspondant à une description très détaillée du domaine doit être alignée avec une ontologie décrivant ce même domaine mais de façon très générale) ou du degré de structuration qui peut être différent (une ontologie très structurée doit être alignée avec une ontologie qui n'est absolument pas ou très peu structurée).

Enfin, la majorité des approches actuelles exploitent des ontologies OWL avec toute leur richesse (propriétés, relations variées). Peu d'approches sont conçues pour faire de l'alignement lorsque toute cette richesse n'est pas disponible, ce qui peut être le cas lorsque le processus adopté pour construire l'ontologie est totalement automatique ou lorsque nous travaillons sur des ontologies réelles souvent réduites à de simples taxonomies (cas des ontologies actuellement disponibles au sein du COGIT).

Les techniques d'indexation qu'elles soient utilisées par les moteurs de recherche du Web ou par les systèmes de gestion documentaire des bibliothèques sont basées pour la plupart, encore aujourd'hui, sur des approches utilisant toutes, soit le principe des « mots-clefs », soit la prise en compte du contenu basée sur des analyses statistiques sur les formes (mots ou graphies) qui constituent un texte. Le taux de rappel est assez fort, mais la précision est faible. À ces techniques, de plus en plus de travaux opposent des méthodes traitant véritablement du « contenu » des documents, mais appréhendé de manière très partielle pour des raisons de complexité. Le gain attendu est à la fois en terme de rappel (plusieurs mots peuvent être associés à un même concept objet de la recherche), de précision car cela permet de dépasser la combinaison booléenne d'indicateurs, mais aussi d'appréhension par l'utilisateur des résultats de sa requête. Dans le domaine de l'information géographique, deux approches majeures prennent en compte aujourd'hui la sémantique du contenu du document. La première porte sur des techniques de récupération d'information dans le contenu documentaire par application de patrons syntaxico-sémantiques associés à des bases lexicales (souvent composées de toponymes). Un des problèmes rencontrés vient du fait qu'un certain nombre d'expressions complexes (la plupart du temps des syntagmes nominaux), potentiellement très significatives, ne sont pas retenues in fine, car les relations qui les lient ne peuvent être explicitées. La seconde approche porte sur des techniques basées sur des ontologies de domaine. Avec ce type de technique, la structure de l'index est donnée par la structure de l'ontologie employée. Un des problèmes rencontrés vient de la difficulté de disposer d'ontologies offrant des concepts spécifiquement dédiés à l'information géographique.

⁶ <http://infodoc.unicaen.fr/geosem/>

Même si des travaux ont été réalisés dans le cadre plus particulier des ontologies dans le domaine géographique et qu'ils mettent en avant la nécessité de ces ontologies [Uitermark 2001], très peu d'ontologies ont été réalisées en pratique [Lemmens 2006] ou alors celles-ci sont restreintes et décrivent des domaines très ciblés. Les travaux sur l'alignement de ces ontologies, que ce soit par analyse des instances ou par utilisation de sources externes comme des dictionnaires, focalisent donc en général sur des cas très particuliers avec un nombre restreint de concepts [Kavouras et al. 2005]. De plus la comparaison globale d'ontologies pour identifier les différences de point de vue a reçu peu d'attention, en dehors d'analyses de l'expression des relations spatiales dans différentes langues [Bowerman 2006].

3 Partenaires

3.1 Présentation générale du partenariat

Le projet rapproche des équipes spécialistes, d'une part, de la construction et de l'alignement d'ontologies et d'autre part, des systèmes d'information et de l'analyse de documents géographiques. Il comporte 4 partenaires académiques. Les 4 partenaires regroupent des équipes dont les compétences recouvrent la totalité des domaines de recherche du projet. Ces équipes sont déjà expérimentées, d'une part dans le domaine de la construction, de l'alignement ou de la fusion d'ontologies, d'autre part dans le domaine du traitement automatique du langage et de l'annotation de documents en particulier géographiques, et enfin dans le domaine de l'intégration des bases de données géographiques. Ceci se reflète dans le nombre de publications faites sur le domaine (voir annexe 1).

3.2 Partenaires

Laboratoire COGIT, Institut Géographique National

Partenaire	COGIT	Type d'organisation	Laboratoire public
Description générale	Le laboratoire COGIT effectue des recherches sur les bases de données géographiques, de leur gestion à leur exploitation. Des travaux du laboratoire portent notamment sur la représentation multiple et l'intégration de bases de données. Les compétences du laboratoire couvrent le domaine de la géomatique (cartographie, bases de données, analyse spatiale) et diverses approches d'intelligence artificielle sont exploitées (apprentissage automatique, systèmes multi-agents, représentation des connaissances).		
Rôle dans le projet	Dans ce projet, le laboratoire apporte dans les premières tâches son expertise sur les données géographiques et de leur intégration. En particulier, les travaux antérieurs du laboratoire permettent de fournir des premières ontologies du monde géographique. Plus en aval, le laboratoire effectue des recherches sur l'exploitation des ontologies créées ou enrichies par le projet, dans le contexte de l'intégration de bases de données. Le laboratoire fournit également aux partenaires les spécifications textuelles des bases de données utiles au lot 1, et un extrait de données géographiques utile au sous-lot 2.2.		
Participations antérieures à des projets coopératifs	Le laboratoire a participé à plusieurs projets européens, sur la cartographie automatique et les systèmes multi-agents (projet AGENT, leader), la représentation multiple (projet MURMUR), la recherche d'information à caractère spatial (projet SPIRIT). Il a participé à l'ACI Tadorne sur le tatouage des données géographiques. Il a déposé cette année trois autres projets ANR, SMASH sur les bases de données géographiques inductives, GeOpenSim sur la simulation d'évolutions en urbanisme, et PASPHIL sur l'analyse et la simulation de mouvements de terrain.		

LRI

Partenaire	LRI, équipe IASI/Gemo	Type d'organisation	Laboratoire public
Description générale	<p>Cette équipe est commune au LRI et à l'INRIA Futurs. Elle regroupe des compétences reconnues au niveau national et international en intelligence artificielle et en bases de données, et travaille sur la gestion d'informations sous toutes ses formes (données, documents XML, services) pouvant être sémantiquement hétérogènes et distribuées sur le Web. Les travaux de l'équipe portent plus particulièrement sur la construction de médiateurs et d'entrepôts de données pour l'intégration sémantique de données hétérogènes, la fouille de données dans des documents XML, le Web sémantique, les services Web.</p>		
Rôle dans le projet	<p>Dans ce projet, l'objectif de l'équipe est d'expérimenter et d'adapter, dans le cadre de nouvelles applications réelles, ses outils de recherche en matière de mise en correspondance d'ontologies (techniques d'alignement d'ontologies, techniques de réconciliation de données) pour aligner, fusionner des ontologies ou affiner les relations représentées dans une ontologie. Il s'agira également d'intervenir dans l'élaboration d'une méthodologie de comparaison d'ontologies représentant des points de vue différents.</p>		
Participations antérieures à des projets coopératifs	<p>L'équipe a l'expérience de la participation à des projets académiques et industriels dont les plus récents sont : PICSEL et MediaD (avec France Telecom R&D) sur l'intégration sémantique de données et l'élaboration d'une plate-forme modulaire pour la médiation sémantique, l'ACI-MDD sur l'Accès au Contenu Informationnel pour les Masses de Données de Documents, e.Dot (projet RNTL pré-compétitif) sur la construction automatique d'entrepôts de données à partir d'informations découvertes sur le Web, WebContent (projet RNTL) sur la construction d'une plate-forme générique et flexible de gestion de contenus intégrant les technologies du Web Sémantique.</p>		

IRIT

Partenaire	IRIT, équipe IC3	Type d'organisation	Laboratoire public
Description générale	<p>Cette équipe possède une compétence reconnue en matière de construction et de maintenance d'ontologies et de terminologies à partir de textes. Les méthodes et logiciels développés font référence à une approche linguistique et terminologique des textes, mettant l'accent sur l'étude de l'usage de la langue, complétée par des analyses statistiques. Inversement, elle s'intéresse à l'utilisation de ces modèles pour caractériser les contenus documentaires et leur évolution dans le temps. Un autre domaine de compétence de IC3 est l'extraction de relations sémantiques à l'aide de marqueurs lexico-syntaxiques pour l'extraction d'informations et la modélisation de connaissances. Ces travaux ont été validés lors de projets en construisant plusieurs terminologies et ontologies dans différents domaines et pour des applications variées, allant de la classification de documents à l'indexation de site web.</p>		
Rôle dans le projet	<p>Dans ce projet, IC3 va expérimenter l'approche par patrons lexico-syntaxiques pour construire automatiquement des propositions d'ontologies, les enrichir et les vérifier en les confrontant à des ressources annexes. L'ensemble des patrons déjà disponibles sera complété et les outils adaptés pour traiter les spécifications des données géographiques, qui ont la particularité d'être peu redondantes et très structurées.</p>		
Participations antérieures à des projets coopératifs	<p>L'équipe a participé à plusieurs projets directement avec des partenaires industriels ou via des programmes nationaux (TCAN, RNTL) sur la construction d'ontologies à partir de textes pour l'annotation sémantique (veille technologique avec Saint-Gobain, aide au diagnostic automobile avec ACTIA, aide à la recherche scientifique, projet Arkeotek-TCAN) et sur l'utilisation de taxinomies pour repérer des évolutions terminologiques et conceptuelles dans la documentation de projets (CNES). Elle développe une</p>		

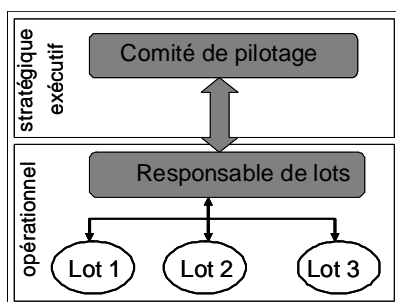
expérience sur l'extraction de relations sémantiques à partir d'articles scientifiques en bioinformatique et contribue sur ce sujet à la plate-forme DAFOE (RNTL).

LIUPPA

Partenaire	LIUPPA, équipe projet DESI	Type d'organisation	Laboratoire public
Description générale	L'équipe projet Document Electronique, Sémantique et Interaction regroupe des compétences reconnues en modélisation et traitements des contenus documentaires à connotation géographique (représentation sémantique de l'information, extraction et recherche d'information) ainsi qu'en interaction et en éducation. Les méthodes et techniques développées dans cette équipe-projet place l'utilisateur au centre de sa thématique de recherche et a fortiori des données multi-modales qu'il manipule via l'outil informatique (ordinateur et réseau). L'interaction avec l'utilisateur suppose donc le développement d'environnements adaptés aux situations d'usage, que ce soit pour l'apprentissage humain, en recherche et extraction dans les données semi-structurées, ou pour la coopération entre utilisateurs de systèmes d'information.		
Rôle dans le projet	Dans ce projet, l'équipe Desi apporte son savoir faire sur l'extraction et l'interprétation d'informations à connotation géographique dans des contenus documentaires, son objectif est de contribuer à améliorer la détection de relations sémantiques autres que hiérarchiques afin d'améliorer l'ontologie, et ce grâce à des méthodes d'extraction d'information (précédemment validées) applicables à des contenus textuels. L'équipe met également à disposition du projet un important corpus documentaire numérisé essentiellement constitué de récits de voyage (riche en concepts géographiques « grand public »).		
Participations antérieures à des projets coopératifs	Les collaborations récentes des membres de l'équipe-projet portent, avec le projet Géosem (CNRS « société de l'information ») et le projet PIV (CAPP/MIDR), sur la mise au point d'une nouvelle démarche pour la construction d'interprétations formelles adaptées aux informations géographiques multimodales (texte et carte) à partir de contenus documentaires (données numériques semi-structurées). Son savoir faire sur l'interaction, les usages et l'apprentissage humain, a permis également de coopérer, grâce au projet Européen « CACTUS », à la réalisation d'un produit d'aide à la lecture critique de l'information (TV, web, etc.), et avec le projet transfrontalier (Conseil Régional d'Aquitaine) « DECUPLE » à la production d'un ensemble de ressources documentées consacrées à la mise en place d'un processus d'e-learning.		

4 Organisation et management du projet

Une structure à 2 niveaux est proposée, conformément au schéma ci-dessous.



Le comité de pilotage (CP) est un organe ayant à la fois des responsabilités au niveau stratégique et exécutif. Le CP est présidé par le coordonnateur du projet, C. Reynaud (LRI / équipe IASI-Gemo). Chaque responsable de lot et chaque partenaire sont représentés. Le CP se réunira une fois par semestre, en alternant les lieux de réunions en fonction de la localisation géographique des partenaires. Des réunions exceptionnelles pourront être convoquées en cas d'urgence. Les décisions ne seront valides que si 3/4 des membres sont présents ou représentés.

Au niveau stratégique, il est l'organe de management dans lequel les décisions stratégiques liées au projet sont prises. Il statue sur les changements décidés dans le programme de travail et résout les litiges au sein du projet. Au niveau exécutif, c'est l'organe dédié au management technique et administratif du projet. Il est chargé de la rédaction/validation des rapports à remettre dans le cadre du projet.

Les responsables de lot sont responsables de l'animation de leur lot et de l'application des différentes actions décidées par le CP. Ils organisent et surveillent le travail au jour le jour, afin de préparer les rapports techniques à fournir. Ils coordonnent également le travail fait dans les différents lots. Afin de coordonner le travail entre les différents lots, les responsables de lot se mettront en contact régulièrement, suivant les besoins. Chaque responsable de lot est libre d'organiser et d'animer son lot en fonction de son expérience et du contexte du lot, tant qu'il suit les directives décidées au niveau du CP. Les tâches de chacun des lots sont, par ailleurs, réparties en sous-lots dont les responsables veillent au bon déroulement des activités.

Les responsables de lots et de sous-lots sont :

Lot 0 «Gestion de projet» : LRI

Lot 1 «Construction et enrichissement d'ontologies» : IRIT

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

Sous-lot 1.2 «Enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction et à partir de ressources lexicales» : LIUPPA

Sous-lot 1.3 : «Restructuration d'ontologie» : LRI

Lot 2 «Appariement d'ontologies hétérogènes» : LRI

Sous-lot 2.1 «Alignement et fusion d'ontologies à divers niveaux de richesse» : LRI

Sous-lot 2.2 «Réconciliation d'instances pour l'alignement d'ontologies» : LRI

Sous-lot 2.3 «Analyse des différences entre ontologies pour faire ressortir des différences de point de vue sous-jacentes» : COGIT

Lot 3 «Exploitation des ontologies créées» : COGIT

Sous-lot 3.1 «Indexation automatique du contenu de documents» : LIUPPA

Sous-lot 3.2 «Intégration et accès aux schémas de bases de données et évaluation» : COGIT

Sous-lot 3.3 «Mise en ligne des ontologies réalisées» : COGIT

5 Structure du projet – Description des sous-projets

Nous proposons de délimiter le domaine d'application du projet en choisissant un sous domaine géographique mettant bien en lumière les problèmes qui se posent. Ce sous domaine correspond à la topographie, c'est-à-dire à la description du paysage. C'est un domaine qui, à lui seul, donne lieu à de nombreuses et volumineuses bases de données, typiquement jusqu'à une centaine de giga-octets en France, auxquelles sont associées des documents textuels correspondant aux spécifications des bases. Ce sous domaine pose à lui seul la question de l'intégration des bases de données pour faciliter leur diffusion. C'est également un domaine qui donne lieu à de nombreux documents textes grand public, dont en particulier les carnets de voyages, les descriptions de panorama ou de paysages. Les travaux seront effectués sur des textes et données réels, issus de l'IGN ou de la médiathèque de Pau.

Le programme de travail se décompose en 3 lots auxquels se rajoute le lot « Gestion de Projet » :

- lot 0 : Gestion de projet
- lot 1 : Construction et enrichissement d'ontologies
- lot 2 : Appariement d'ontologies hétérogènes
- lot 3 : Exploitation des ontologies créées

5.1 Lot 0 : Gestion de projet

# Lot	0	Nom du lot	Gestion de projet
Responsable	LRI	Participants	Membres des organes de management du projet
Objectif : Assurer la gestion administrative, financière et technique du projet.			

La gestion technique est coordonnée au niveau du comité de Pilotage du projet et est assumée au jour le jour par les responsables de lot. Elle consiste à :

- suivre les réalisations des partenaires afin que les délais soient respectés et coordonner le travail entre les différents lots
- collecter les informations nécessaires au suivi de l'avancement du projet,
- préparer et valider les comptes rendus intermédiaires semestriels d'avancement ainsi que le rapport final,
- créer un site Web et veiller à sa mise à jour régulière,
- diffuser les résultats au travers de publications dans des conférences.

La gestion administrative et financière du projet concerne le Comité de Pilotage. Elle consiste à :

- organiser les réunions : rédaction et diffusion de l'ordre du jour, des comptes rendus, réservation de salles, etc.
- mettre à jour le contrat si nécessaire,
- suivre les dépenses.

5.2 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 <i>de référence</i> 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 lexicale 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 (hyponymie, 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 MIDR⁷. 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

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

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.

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

⁷ MIDR: Médiathèque Intercommunale à Dimension Régionale de Pau.

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

5.3 Lot 2: Appariement d'ontologies hétérogènes

# Lot	2	Nom du lot	Appariement d'ontologies hétérogènes
Responsable	LRI	Participants	LRI, COGIT

Résumé:

Une ontologie est un schéma particulier qui décrit, à l'aide d'un vocabulaire structuré et un langage formel non ambigu, les concepts et les propriétés pertinents pour un domaine d'application donné. Nous travaillerons sur des ontologies représentées dans le langage de description d'ontologie OWL recommandé par le W3C.

Dans ce lot, nous exploiterons, dans un premier temps, des ontologies déjà construites et disponibles au sein du COGIT. En effet, les travaux passés réalisés au COGIT permettent de disposer de deux taxonomies géographiques contenant quelques centaines de concepts, réalisées chacune à partir d'analyse de spécifications textuelles de bases de données (d'une centaine de pages chacune) et plus ou moins structurées. Dans un second temps, l'appariement sera réalisé sur les ontologies obtenues en résultat des sous lots 1.1 et 1.2 ainsi que sur d'autres ontologies ou taxonomies externes accessibles.

En premier lieu, il s'agira donc d'apparier les ontologies disponibles pour obtenir une ontologie géographique riche. En deuxième lieu, il s'agira de les comparer pour comprendre leurs différences. L'ensemble des résultats obtenus sera validé d'un point de vue géographique. Ces objectifs motivent le découpage du lot 2 dans les sous lots présentés ci-dessous.

Sous-lot 2.1 : Alignement et fusion d'ontologies à divers niveaux de richesse

Limites actuelles de l'état de l'art

Les travaux actuels d'alignement tirent parti des différents aspects des ontologies (leur structure, les noms des différents éléments, les objets, la sémantique du langage). Ils sont, pour la plupart, basés sur la recherche d'analogies dans les modèles comparés : concepts identiques ou similaires, structures

identiques ou proches, propriétés identiques ou conciliables, etc. [Shvaiko et Euzenat 2005] [Kalfoglou et Schorlemmer 2003]. Ainsi, les techniques mises en œuvre dans ces travaux se trouvent limitées lorsque les ontologies comparées sont réduites à de simples taxonomies qui comprennent uniquement un ensemble de concepts et une hiérarchie de subsomption entre concepts. Elles sont, par ailleurs, inapplicables lorsque certains des éléments sur lesquels se base la comparaison ne sont pas présents dans l'une des deux ontologies mises en correspondance.

Innovations apportées

L'innovation portera sur les techniques d'appariement qui devront être capables de mettre en correspondance de simples taxonomies et de gérer l'hétérogénéité aux niveaux suivants :

- Hétérogénéité structurelle. Il devra être possible d'aligner une ontologie disposant de très nombreux niveaux de structuration avec une ontologie pas ou très peu structurée.
- Hétérogénéité du point de vue du niveau de précision. Il devra être possible d'aligner une ontologie correspondant à une description très détaillée du domaine avec une ontologie décrivant ce même domaine mais de façon grossière.
- Hétérogénéité du point de vue de la qualité de la représentation. Certaines ontologies représentent des concepts liés par des relations hiérarchiques dont la sémantique est ambiguë, d'autres ontologies explicitent la nature des relations en distinguant, par exemple, précisément les relations de subsomption des relations 'partie-de'.
- Hétérogénéité du point de vue de la fiabilité de la représentation. Certaines ontologies, et en particulier les ontologies construites automatiquement, pourront contenir des erreurs (exemple : *bassin d'épuration* classé par erreur sous *bassin de natation*).

Description

Il s'agira de partir de travaux développés par l'équipe IASI du LRI dans des projets précédents et de les étendre pour traiter le plus complètement possible le problème d'appariement de taxonomies hétérogènes décrit ci-dessus. Dans le cadre d'e.Dot, l'équipe IASI a développé une approche générique de mise en correspondance entre taxonomies, correspondant à des ontologies très sommaires avec des définitions de concepts très pauvres. Cette approche propose plusieurs techniques générant des mappings de 2 sortes : des mappings probables et des mappings potentiels qu'un expert doit confirmer. Le processus d'alignement est semi-automatique. Il peut être vu comme une application séquentielle de différentes techniques : terminologiques puis structurelles. Les techniques terminologiques, basées principalement sur des comparaisons de chaînes de caractères, sont appliquées en priorité. Elles exploitent toute la richesse des noms des concepts. Ces techniques sont efficaces. Elles fournissent des mappings de grande qualité que nous qualifions de probables. Même si elles sont efficaces, les techniques terminologiques ne peuvent cependant pas trouver l'ensemble des rapprochements possibles. Les techniques terminologiques sont donc complétées par des techniques basées sur l'exploitation de la structure. Les règles heuristiques communément utilisées dans les travaux d'alignement considèrent que deux entités de deux taxonomies sont similaires si leur voisinage respectif est similaire. Ces règles n'ont pas été appliquées car elles sont inutilisables quand l'une des deux taxonomies est pauvre structurellement. Des techniques différentes adaptées à une dissymétrie structurelle dans les taxonomies comparées sont au contraire proposées. Les mappings supplémentaires générés sont moins sûrs que ceux générés par les techniques terminologiques, ils sont qualifiés de mappings potentiels. Leur validation est indispensable.

Le prototype d'alignement existant sera testé sur les premières taxonomies réelles du COGIT, puis adapté pour prendre en compte complètement les spécificités des ontologies du domaine d'étude et pour aller jusqu'à la fusion d'ontologies, si nécessaire. Le prototype adapté sera ensuite testé sur les ontologies obtenues en résultat du sous lot 1.1. Tous les résultats d'appariement seront évalués d'un point de vue géographique.

Planning d'activité prévisionnel

N°	Description	Date de fin de tâche
2.1	Test des techniques existantes sur les premières taxonomies réelles fournies par le COGIT	T0 + 6

2.2	Identification des limites des techniques d'alignement actuellement mises en œuvre dans le prototype existant et spécification des améliorations à apporter	T0 + 12
2.3	Adaptation du prototype existant	T0 + 18
2.4	Test et validation des résultats obtenus d'un point de vue géographique. Evaluation de la robustesse de l'outil développé par rapport aux ontologies obtenues en résultat du sous lot 1.1.	T0 + 24
2.5	Extension du prototype pour fusionner des ontologies. Tests et validation des résultats obtenus d'un point de vue géographique. Evaluation de la robustesse de l'outil développé.	T0 + 30
		T0 + 36

Sous-lot 2.2 : Réconciliation d'instances pour l'alignement d'ontologies

Limites actuelles de l'état de l'art

Dans le domaine de l'information géographique, il existe des outils de réconciliation automatique d'instances qui considèrent l'appariement de concepts comme une entrée du processus et qui s'appuient principalement sur une comparaison de la localisation spatiale des instances. Mais d'une part ces outils exploitent peu les approches utilisées pour d'autres types de données non localisés, et d'autre part la difficulté de la tâche rend leurs résultats entachés d'erreurs.

Par ailleurs, certains travaux exploitent le fait que deux ontologies soient peuplées (i.e. soient telles qu'il est possible d'accéder aux instances de concepts ou de relations de l'ontologie) pour raisonner sur les correspondances entre concepts de deux ontologies. Cette information supplémentaire peut tout d'abord être utilisée de manière extensionnelle. Les approches peuvent alors utiliser la similarité entre ensembles d'instances appartenant à différents concepts pour raisonner sur la similarité des concepts [Stumme et Maedche 2001], [Euzenat et Valchev 2004]. Ces travaux se basent sur des instances dont la réconciliation ou la non réconciliation est déjà connue. D'autre part, la description des instances peut également être utilisée de manière intensionnelle. En particulier, l'exploitation des ensembles de valeurs participant à la description des instances permet de raisonner sur la similarité des éléments de description des instances et donc des concepts [Dhamankar et al. 2004].

Nous voyons que l'alignement d'ontologies guide la réconciliation des données et que la réconciliation des données peut être une aide pour aligner deux ontologies mais peu d'approches envisagent ces deux aspects du problème.

Innovations apportées

L'innovation apportée consistera à permettre l'alignement d'ontologies par combinaison de techniques de recherche de mises en correspondance entre concepts et entre instances. L'originalité de la méthodologie mise au point sera de gérer simultanément les deux niveaux que sont la réconciliation d'instances et l'alignement de concepts. Il s'agira d'étudier comment un alignement de concepts guide la réconciliation d'instances, et comment à l'inverse on peut induire un alignement de concepts à partir d'une réconciliation d'instances, tout en gérant les incertitudes de cette dernière.

Description

Dans le cadre de ce projet, il s'agira de mettre au point une méthodologie de convergence alternant des phases d'alignement de concepts et de réconciliation d'instances. En effet, cette alternance devra permettre d'exploiter un premier alignement de concepts grossier (ou vide), pour inférer un ensemble de réconciliations d'instances entachées d'incertitudes, à partir duquel il sera possible de raffiner l'alignement de concepts, ce qui permettra ensuite de réconcilier de nouvelles instances, et ainsi de suite.

Les étapes de réconciliation d'instances pourront combiner deux approches différentes de réconciliations proposées par deux des partenaires. Ces approches exploiteront une partie de la masse des données disponibles, des centaines de milliers de données pour certains thèmes géographiques.

Dans le cadre de travaux réalisés au sein du projet PICSEL3, l'équipe IASI du LRI a, en effet, proposé une approche originale pour résoudre le problème de la réconciliation des données en vue de leur fusion dans un entrepôt, en deux étapes. La première étape est une étape logique qui exploite la sémantique du schéma commun des données à réconcilier pour inférer des réconciliations et des non réconciliations qui sont certaines. La seconde étape est une étape numérique qui se concentre sur les couples de

données dont la réconciliation ou la non réconciliation n'ont pu être établies lors de la première étape. Elle exploite toutes les valeurs associées aux propriétés des données, ainsi que les dépendances existant entre données, pour calculer une valeur de similarité globale entre chaque paire de données considérée. Ces deux étapes peuvent être enchaînées mais peuvent être aussi appliquées indépendamment l'une de l'autre.

Les travaux réalisés s'appuieront également sur l'expérience du COGIT en réconciliation de données géographiques. Une partie des travaux réalisés sur ce sujet s'est concentrée sur la comparaison de réseaux géographiques à différentes échelles. Une autre partie, faisant l'objet d'une thèse en cours à soutenir en 2008, s'intéresse à l'application de la théorie de l'évidence pour gérer les incertitudes tout au long du processus d'appariement.

La méthodologie de convergence définie dans le cadre de ce projet devra permettre d'exploiter au mieux les propriétés des méthodes de réconciliation de ces deux partenaires.

Planning d'activité prévisionnel

N°	Description	Date de fin de tâche
		T0 + 6
		T0 + 12
2.6	Etude des différences et complémentarité des approches en réconciliation de concepts développées au COGIT et au LRI. Mise au point d'une méthode de réconciliation d'instances combinant les approches antérieures.	T0 + 18
2.7	Mise au point d'une stratégie de convergence entre alignement de concepts et réconciliation d'instances. Test de la méthodologie sur des données du COGIT à différentes échelles spatiales.	T0 + 24
2.8	Introduction de la méthodologie au sein du processus plus général d'alignement d'ontologie développé au lot 2.1	T0 + 30
		T0 + 36

Sous-lot 2.3 : Analyse des différences entre ontologies pour faire ressortir les différences de points de vue sous-jacentes

Limites actuelles de l'état de l'art

Le problème de l'identification du point de vue utilisé dans un processus de modélisation rejoint le problème d'identification du contexte auquel sont actuellement confrontés les chercheurs en alignement qui ont recours à des connaissances supplémentaires (WordNet, ontologie accessible via Swoogle, autre ontologie du domaine) pour pallier l'insuffisance de connaissances contenues dans les modèles alignés. Aucune solution n'a, pour l'instant, été trouvée.

Innovations apportées

L'innovation portera sur l'élaboration d'une méthodologie générique de comparaison d'ontologies aidant un concepteur à prendre la décision de fusionner des ontologies ou de simplement les mettre en correspondance.

Description

Il s'agira de proposer une méthodologie de comparaison d'ontologies aidant à comprendre les différences entre ontologies, selon différents critères : le pays d'origine des concepteurs, le niveau de détail des ontologies, l'usage qui en est fait, etc. Le résultat de ce travail de comparaison pourra aider à décider si des ontologies doivent être fusionnées car elles modélisent un domaine d'application d'un point de vue similaire, ou au contraire, si l'appariement doit se réduire à établir des mises en correspondance car les différences de modélisation sont plus *profondes*.

Ce travail de comparaison pourra prendre appui sur les résultats fournis par l'application des techniques d'alignement, en exploitant les traces du processus d'alignement et en proposant des outils pour aider à interpréter les résultats. On s'attachera à repérer les indices intéressants dans un processus de comparaison : sous arbres de la hiérarchie de concepts communs, proportion de concepts communs,

position des concepts communs ou similaires dans la hiérarchie des concepts, etc. et on tentera de proposer des critères quantitatifs et qualitatifs à prendre en compte pour mesurer les différences entre deux ontologies.

Des tests seront réalisés à partir des ontologies obtenues dans les sous-lots 1.1 et 1.2, puis étendus à la comparaison avec des ontologies externes. Parmi ces ontologies externes à exploiter, on peut envisager la taxonomie Feature Data Dictionary Register du groupement international « Digital Geospatial Information Working Group », et l'ontologie en cours de construction de l'Ordonance Survey (équivalent de l'IGN du Royaume-Uni, actif en ce domaine).

Planning d'activité prévisionnel

N°	Description	Date de fin de tâche
		T0 + 6
		T0 + 12
		T0 + 18
2.9	Analyse des besoins en prenant appui sur la comparaison d'ontologies du domaine géographique. Confrontation des besoins avec les techniques d'alignement mises en œuvre dans le prototype d'alignement utilisé. Exploitation de la trace du processus d'alignement	T0 + 24
2.10	Spécification de techniques d'aide à l'interprétation des résultats d'alignement Module logiciel de comparaison d'ontologies.	T0 + 30
2.11	Evaluation de l'outil développé. Tests et validation des résultats sur des ontologies variées du domaine géographique.	T0 + 36

5.4 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 données géographique (sous lot 3.2). Dans le cas de l'indexation, les représentations conceptuelles (dédites grâce à l'ontologie) et géométriques (obtenues grâce à l'association entre termes de l'ontologie et champ d'une base de données 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 workshops⁸.

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, geowebsevice, 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 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
3.1	Etude des apports sémantiques de l'ontologie topographique dans le cadre d'une indexation géographique de contenus textuels	T0 + 6 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

⁸ Par exemple les ateliers GIR <http://www.geo.unizh.ch/~rsp/gir06/> qui se tiennent conjointement lors des conférences ACM- SIGIR.

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'existence d'ontologies du domaine riches et alignées, comme celles issues des lots 1 et 2. Dans un premier temps, elle s'attachera à étudier comment 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 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.	T0 + 30
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

Diagramme d'enchaînement des tâches

Le diagramme ci-dessous donne une vue synthétique des dates de début et de fin des tâches par sous-lot, ainsi que les principaux partenaires impliqués dans chaque sous-lot.

N° du sous-lot	Libellé du sous-lot	T0 + 6	T0 + 12	T0 + 18	T0 + 24	T0 + 30	T0 + 36
1.1	Mise au point d'outils d'extraction de concepts et de relations	IRIT et LIUPPA (COGIT/LIUPPA pour expertise sur specs et validation)					
1.2	Enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction et à partir de ressources lexicales		IRIT, LIUPPA et LRI (et COGIT/LIUPPA pour validation)				
1.3	Restructuration d'ontologie				LRI, IRIT (et COGIT pour validation)		
2.1	Alignement et fusion d'ontologies à divers niveaux de richesse	LRI (et COGIT pour validation)					
2.2	Réconciliation d'instances pour l'alignement d'ontologies			LRI et COGIT			
2.3	Analyse des différences entre ontologies pour faire ressortir les différences de point de vue sous-jacentes				LRI et COGIT		
3.1	Indexation automatique de contenu de documents		LIUPPA				
3.2	Intégration et accès aux schémas de bases de données et validation	COGIT					
3.3	Mise en ligne des ontologies réalisées					COGIT	

6 Liste des livrables

	Libellé du livrable	Type	Responsable	Partenaires participants	Date
0	Site web du projet – Mise en place au plus tard 6 mois après le démarrage du projet et mise à jour au moins semestrielle	Web	Coordonnateur	Tous	T0+6
1	Mise au point d'outils d'extraction de concepts et de relations	Rapport intermédiaire	IRIT	IRIT, LIUPPA, COGIT	T0 + 12
2	Alignement d'ontologies	Rapport intermédiaire	LRI	LRI	T0 + 12
3	Mise au point d'outils d'extraction de concepts et de relations	Module logiciel (V1)	IRIT	IRIT, LIUPPA, COGIT	T0 + 18
4	Enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction et à partir de ressources lexicales	Module logiciel	IRIT	IRIT, LIUPPA, LRI, COGIT	T0 + 18
5	Intégration et accès aux schémas de bases de données	Rapport intermédiaire	COGIT	COGIT	T0 + 18
6	Indexation automatique de contenu de documents	Rapport intermédiaire	LIUPPA	LIUPPA	T0 + 18
7	Mise au point d'outils d'extraction de concepts et de relations	Rapport final Module logiciel (V2)	IRIT	IRIT, LIUPPA	T0 + 24
8	Alignement d'ontologies	Rapport final Module logiciel	LRI	LRI	T0 + 30
9	Enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction et à partir de ressources lexicales	Rapport final Module logiciel	LRI	IRIT, LRI	T0 + 30
10	Réconciliation d'instances pour l'alignement d'ontologies	Rapport final Module logiciel	LRI	LRI	T0 + 30
11	Restructuration d'une ontologie construite automatiquement ⁹	Module logiciel	LRI	LRI	T0 + 36
12	Analyse des différences entre ontologies pour faire ressortir les différences de points de vue sous-jacentes ¹⁰	Module logiciel	COGIT	COGIT, LRI	T0 + 36
13	Intégration et accès aux schémas de bases de données ¹⁰	Module logiciel	COGIT	COGIT	T0 + 36
14	Indexation automatique du contenu des documents ¹⁰	Module logiciel	LIUPPA	LIUPPA	T0 + 36
15	Constitution, alignement, comparaison et exploitation d'ontologies géographiques hétérogènes	Logiciels expérimentaux, site web, Rapport final	Coordonnateur	Tous	T0 + 36
16	Diffusion scientifique	Publications	Coordonnateur	Tous	Tout au long du projet

⁹ Le rapport final du sous-lot est intégré au rapport final du projet rendu à la même date.

7.1 Retombées scientifiques et techniques

D'un point de vue méthodologique, le principal objectif du projet est d'éprouver et d'adapter des outils et méthodes existants pour créer, enrichir, aligner et exploiter des ontologies à partir de données et de connaissances qui ont la particularité d'être, d'une part réelles, et donc entachées d'erreurs ou d'imprécisions, et d'autre part hétérogènes en terme de structure et de niveaux de structuration.

Du point de vue du cadre applicatif, à savoir la description et l'intégration de données géographiques, le principal objectif est de créer des ontologies riches en termes de quantité de concepts manipulés et en termes de sémantique afin de permettre l'accès à une masse de données géographiques hétérogènes, qu'il s'agisse de contenus de bases de données ou de textes grand public. Le deuxième objectif est d'appréhender de manière fine les différences de point de vue sous-jacentes à diverses sources de connaissances du domaine.

Les résultats attendus dans ce projet sont les suivants :

- Des outils nouveaux d'extraction de concepts et de relations dans des textes, basés sur la définition de patrons, sur le repérage d'Entités Nommées Géographiques, sur l'exploitation de la structure des textes et des relations argumentatives,
- Des techniques et une méthodologie de création, d'enrichissement et de restructuration d'ontologie, combinant l'utilisation de patrons, de ressources lexicales externes, de techniques d'alignement d'ontologies,
- Une étude de l'efficacité des techniques d'alignement, l'adaptation de techniques actuelles et le développement de nouvelles techniques adaptées à la fusion efficace de deux ontologies hétérogènes de qualité moyenne (cas des ontologies réelles),
- Des techniques et une méthodologie de comparaison de deux ontologies reflétant des niveaux d'échelle spatiale différents ou issues de pays différents, afin d'étudier si ces ontologies reflètent de véritables points de vue différents ou si elles n'ont que des différences marginales et gagneraient donc à être fusionnées.
- Une ontologie du domaine de l'information géographique, et plus particulièrement de la description topographique du paysage,
- Des techniques et méthodologies exploitant une ontologie pour indexer automatiquement le contenu de documents diversifiés et pour apparier des schémas de bases de données hétérogènes.

Le projet étant motivé par la volonté d'intégrer des informations à caractère géographiques, les techniques et les méthodes proposées, bien que génériques, c'est-à-dire applicables a priori à tout domaine d'application, seront validées sur ce domaine. Tous les résultats seront évalués d'un point de vue géographique par les chercheurs du laboratoire COGIT de l'IGN, experts du domaine.

Une ontologie topographique riche sera mise à disposition de la communauté scientifique. Elle sera en bilingue français/anglais et répondra aux attentes des groupes de travail associés à la directive européenne INSPIRE qui ont exprimé le besoin de disposer d'ontologies dans le domaine géographique, les ontologies existant aujourd'hui étant très restreintes et décrivant des domaines très ciblés.

Le contexte applicatif du projet permettra d'assurer la pérennité des résultats. Cela aura lieu d'une part, à travers un site web maintenu par l'IGN mettant à disposition les ontologies définies et fournissant un certain nombre d'outils associés (moteur d'accès aux schémas des bases de données, résultats de l'appariement de schémas, etc.). Par ailleurs, comme déjà indiqué, le LIUPPA dispose d'une plateforme expérimentale basée sur des Services Web, ces services, décrits en WSDL et instanciés dans un service UDDI, sont, aujourd'hui déjà, une preuve de notre capacité à pérenniser nos travaux¹⁰.

¹⁰ A cause de tests nécessaires liés à la sécurité, ces services, bien qu'étant déclarés dans un annuaire, ne sont, pour le moment, disponibles

Les outils mis au point pour l'appariement de schémas bases de données au COGIT seront développés dans le système d'information géographique OpenSource GéOxygène. L'opportunité de leur dépôt en OpenSource ainsi que leur éventuelle intégration au sein de services web sera étudiée.

Au niveau dissémination scientifique des résultats, ceux-ci seront publiés dans des conférences internationales (International Conference on Knowledge Capture (K-CAP), International Conference on Knowledge Engineering and Knowledge Management (EKAW), International Semantic Web Conference (ISWC), Ontology-Mapping Workshop), nationales (RFIA, IC, TIA, EGC, BDA) ou revues réputées sur l'intégration de bases de données ou les ontologies (Knowledge Engineering Review, Journal on Data and Knowledge Engineering, Very Large Database Journal, TSI), ainsi que sur l'information géographique (par exemple Revue Internationale de Géomatique, IJGIS, AGILE). Les résultats seront également publiés sur le site Web du projet, ce qui permettra aux organismes intéressés de suivre l'avancement du projet et de prendre connaissance des résultats obtenus.

8 Propriété intellectuelle

Le projet n'est soumis à aucune restriction de confidentialité. Chaque partenaire gardera la propriété intellectuelle des outils/techniques développées durant le projet. Si les partenaires échangent des données dans le cadre du projet, et notamment dans le cas de données géographiques de l'IGN, une convention d'échange sera établie. Par ailleurs, les données ne peuvent pas être diffusées en dehors du projet sans l'autorisation de l'IGN.

Annexe 1 Bibliographie

Références bibliographiques

- [Aussenac & Jacques 2006] Aussenac-Gilles N., Jacques M.-P., *Variabilité des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques*. Dans : Traitement Automatique des Langues, Numéro spécial Non Thématique, Paris : Hermès Sciences, V. 47, N. 2, 2006 (à paraître).
- [Bourigault et al. 2001] Bourigault, D., Jacquemin, C., L'Homme M.-Cl., editors, [Recent Advances in Computational Terminology](#), John Benjamins, Amsterdam. 2001.
- [Buitelaar et al. 2005] Buitelaar P., Cimiano P. and Magnini B. (eds.), *Ontology Learning from Text: Methods, Applications and Evaluation*, Frontiers in Artificial Intelligence, IOS Press, Amsterdam, 2005
- [Cimiano et al. 2004] Cimiano P., Pivk A., Schmidt-Thieme L., Staab S., *Learning taxonomic relations from heterogeneous sources*. Workshop on Ontology Learning and Population at ECAI 2004, Valencia, Spain, August 2004.
- [Cimiano 2006] Cimiano P. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*, Springer Verlag, 2006.
- [Ciravegna et al. 2002] Ciravegna F., Dingli A., Petrelli D., Wilks Y., *User-system cooperation in document annotation based on information extraction*. In Gómez-Pérez A., Benjamins V.R., (Eds), proc. of the 13th International conference in Knowledge Engineering and Knowledge Management, EKAW 2002. LNAI 2473. Berlin, Springer Verlag.
- [Dhamankhar et al. 2004] Dhamankar R., Lee Y., Doan A., Halevy A., Domingos P., *Imap: discovering Complex Semantic Matches between database schemas*, In SIGMOD 2004, pp. 383-394.
- [Enjalbert & Gaio 2004] Enjalbert P., Gaio M. *Actualité d'une approche sémantique du document numérique*. In P. Enjalbert and M. Gaio, editors, Actes du 7^{ième} Colloque International sur le Document Electronique (CIDE.7), pp. 13–28, LaRochele, France, juin 2004.
- [Euzenat & Valchev 2004] Euzenat J., Valchev P., *Similarity based Ontology Alignment*. In 16th European Conference on Artificial Intelligence, ECAI 2004, pp 333-337, Valencia, Spain, August 2004.

que sur demande explicite.

[Gesbert 2005] Nils Gesbert, Etude de la formalisation des spécifications de bases de données géographiques en vue de leur intégration, Thèse de doctorat de Marne La Vallée, 2 Décembre 2005.

[Han & Schulz 2002] Han U., Schulz S., *Turning lead into gold? Feeding a Formal Knowledge Base with Informal Conceptual Knowledge*. Proc. of EKAW2002, Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. A. Gomez-Pérez and R. Benjamins Eds. LNAI 2473. Springer, pp. 182-196, 2002.

[Kalfoglou & Schorlemmer 2003] Kalfoglou Y., Schorlemmer M., *Ontology mapping: The state of the art*, The Knowledge Engineering Review, 18:1-31, 2003.

[Kavouras et al. 2005] M. Kavouras, M. Kokla & E. Tomai, Computers & Geosciences, 31(2): pp 145-154, Special Issue, Geospatial Research in Europe: AGILE 2003, Michael Gould (Ed.), March 2005.

[Lemmens 2006] : Lemmens, R.L.G. Exploitation of ontology mappings for the discovery of geo - web services In: Geographic information science: proceedings of the 4th international conference GIScience 2006, Münster, Germany, 20-23 September, 2006.

[Maetche & Staab 2000] Maedche A., Staab S. *Semi-automatic Engineering of Ontologies from texts*, Conference on Software Engineering and Knowledge Engineering SEKE2000. Chicago, USA, 2000.

[Shvaiko & Euzenat 2005] Shvaiko P., Euzenat J., *A survey of schema-based Matching Approaches*. Journal of Data Semantics, 4:146-171, 2005.

[Stumme & Maedche 2001] Stumme G., Maedche A., *FCA-Merge : Bottom-up Merging of Ontologies*, IJCAI 2001, pp 225-234, 2001.

[Uitermark 2001] : Uitermark H. Ontology-Based Geographic Data Set Integration. PhD thesis, Universiteit Twente, the Netherlands, 2001.

Publications récentes en lien avec le projet

LRI / Equipe IASI-GEMO

Reynaud C., Safar B. *Techniques structurelles d'alignement pour portails Web*. Revue RNTI, N° spécial Fouille du Web. à paraître en 2007.

Reynaud C., Safar B.. *Utilisation de connaissances supplémentaires pour la découverte de mappings dans le système TaxoMap*. Atelier DECOR, 7^{ièmes} journées francophones « Extraction et Gestion des Connaissances », EGC'2007, Namur, Belgique, janvier 2007.

Sais F., Pernelle N., Rousset M.-Ch. *Approche logique pour la réconciliation de références*. 7^{ièmes} journées francophones « Extraction et Gestion des Connaissances », n° spécial de la revue RNTI, EGC'2007, Namur, Belgique. janvier 2007.

Sais F., Pernelle N., Rousset M.-Ch. *Passage à l'échelle de la réconciliation de concepts et de la réconciliation de références : quelques points de comparaison*. Atelier DECOR, 7^{ièmes} journées francophones « Extraction et gestion des Connaissances », EGC'2007, Namur, Belgique, janvier 2007.

Reynaud C., Safar B., *When Usual Structural Alignment Techniques don't apply*, ISWC'06 Workshop on Ontology Matching, OM'06, Athens, USA, 5-9 Nov. 2006.

Kefi H., Safar B., Reynaud C., *Alignement de taxonomies pour l'interrogation de sources d'information hétérogènes*, 15^{ième} congrès francophone Reconnaissance des Formes et Intelligence Artificielle RFIA2006, Tours, 25-27 janvier 2006.

Delobel C., Reynaud C., Rousset M.-Ch., Sirot J.-P., Vodislav D., *Semantic Integration in Xyleme: a Uniform Tree-based Approach*, Journal on Data and Knowledge Engineering, 44(2), pp 267-298, 2003.

COGIT

Abadie N., Gesbert N., Mustière S. 2006. *Création d'une ontologie à partir des spécifications textuelles pour l'intégration des bases de données géographiques*. Dans les actes de la conférences Ingénierie des Connaissances (IC'2006), Nantes, Juin 2006.

Gesbert Nils, Libourel Thérèse, Mustière Sébastien, 2004. *Formalisation des spécifications de bases de données géographiques pour une meilleure compréhension des données*. Revue Internationale de Géomatique, numéro spécial sur les ontologies, vol.14, pp.239-257.

Mustière S., Abadie N., Laurens F. 2007. *Appariement de schémas de BD géographiques à l'aide d'ontologies déduites des spécifications*. Actes de l'atelier DECOR (passage à l'échelle des techniques de découverte de correspondances), Namur - Belgique, 23 janvier 2007, pp.22-27.

Mustière Sébastien et Smaalen John van. 2007. "Chapter 6 : *Database Requirements for Generalisation and Multiple Representations*. Dans *The Generalisation of Geographic Information : Models and Applications*, sous la direction de William Mackaness, Anne Ruas et Tiina Sarjakoski. Elsevier.

Mustière Sébastien 2005. *Cartographic generalization of roads in a local and adaptive approach : A knowledge acquisition problem*. International Journal of Geographical Information Science, Special issue on cartographic generalisation. vol. 19, n.8-9, sept-oct.2005, pp.937-955.

Olteanu A.M., Mustière S., Ruas A. 2006. *Matching Imperfect Data*. Dans les actes de International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Science (Accuracy'2006), Lisbon, Juillet 2006, pp.694-704.

Sheeren David, Mustière Sébastien, Zucker Jean-Daniel, 2004. *Consistency Assessment Between Multiple Representations of Geographical Databases: a Specification-Based Approach*. International Symposium on Spatial Data Handling (SDH), pp.617-627.

IRIT / Equipe IC3

Aussenac-Gilles N., Séguéla P., (2000) *Les relations sémantiques : du linguistique au formel*. Cahiers de grammaire, Numéro spécial sur la linguistique de corpus. A. Condamines (Ed.) Vol 25.Toulouse : Presse de l'UTM. pp 175-198, Déc. 2000.

Aussenac-Gilles N., Jacques M.-P., *Variabilité des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques*. Dans : *Traitement Automatique des Langues*, Numéro spécial Non Thématique, Paris : Hermès Sciences, V. 47, N. 2, 2006 (à paraître).

Aussenac-Gilles N., Jacques M.-P.. *Designing and Evaluating patterns for Ontology Enrichment from texts*. Dans : International Conference on Knowledge Engineering and Knowledge Management EKAW 2006, Prague, 2-6 Oct.2006, Steffen Staab, Vojtek Svatek (Eds.), Springer-Verlag, Lecture Notes in AI, V.4248, pp. 158-165, Oct 2006. Accès: <ftp://ftp.irit.fr/IRIT/CSC/EKAW2006defin-LNCS4248-0158.pdf>

Aussenac-Gilles N., Biébow B., Szulman S., *D'une méthode à un guide pratique de modélisation de connaissances à partir de textes*. 5e rencontres TIA 2003. Ed. F. Rousselot. Strasbourg (F), ENSSAIS, Avril 2003. pp 41-53.

Kamel M. *Extraction d'information pour l'aide au ciblage des gènes*. Séminaire ERSS-IRIT. Toulouse, Mars 2007.

Kamel M. Benamara F. *Intégrer une ontologie de domaine dans un système question-réponse*. Soumis à la Conférence d'Ingénierie des Connaissances IC 2007. Grenoble, Juillet 2007.

Kamel M., *Outil d'extraction d'information pour le ciblage des gènes impliqués dans les maladies*. Soumis à Jobim 2007.

LIUPPA

Gaio, M., Sallaberry, C., Etcheverry, P., Marquesuzaà, C., and Lesbegueries, J. *A Geographical information access for non-structured data*. Journal Of Visual Languages And Computing - Elsevier. To be published in 2007.

Sallaberry, C., Gaio, M., Lesbegueries, J., and Loustau, P. *A Semantic Approach for Geospatial Information Extraction from Unstructured Documents*. Chapter for an upcoming book on the Geospatial Web, published by Springer in the Advanced Information and Knowledge Processing Series. To be published in 2007.

Enjalbert P. et Gaio M. Géosem. *Traitements sémantiques pour l'information géographique, textes et cartes*, Revue Internationale de Géomatique, 16(2):181-194, Editions Hermès, Paris, 2006.

Lesbegueries, J., Gaio, M., Loustau, P., and Sallaberry, C. *Geographical information access for non-structured data*. 21st ACM Symposium on Applied Computing - Advances in Spatial and Image based Information Systems track, pp. 83-89, Dijon, 2006, ISBN : 1-59593-108-2.

Lesbegueries, J., Sallaberry, C., and Gaio, M. *Associating spatial patterns to text-units for summarizing geographic information*. 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval - GIR (Geographic Information Retrieval) Workshop, pp. 40-43, Seattle, ACM SIGIR 2006.

Nodenot, T., Loustau, P., Gaio, M., Sallaberry, C., and Lopistéguy, P. *From Electronic Documents to Problem-Based Learning Environments : an ongoing Challenge for Educational Modelling Languages*. 7th International

Conference on Information Technology Based Higher Education and Training, pp. 75-86, Sydney, Australia, July 06, ITHET 2006.

Sallaberry, C., Etcheverry, P., and Marquesuzaà, C. *Information Retrieval and Visualization Based on Documents' Geospatial Semantics*. 4th IEEE International Conference on Information Technology: Research and Education, Tel Aviv, Israel, October 06, ITRE 2006.

Sallaberry, C., Marquesuzaà, C., and Etcheverry, P. *Spatial Information Management within Digital Libraries*. 1st IEEE International Conference on Digital Information Management (ICDIM 2006), pp. 465-475, Bangalore, India, December 06, ISBN : 1-4244-0682-X.

Annexe 2

Sujets de thèse associés au projet

Thèse COGIT :

Cette thèse est entièrement financée par l'IGN et concerne une ingénieur IGN permanente.

Co-encadrants : Anne Ruas et Sébastien Mustière

Descriptif du sujet :

Le sujet de la thèse est "l'intégration des bases de données à partir de la formalisation de leur spécifications". Il correspond au sous-lot 3.2. La thèse supposera l'existence d'ontologies du domaine riches et alignées, comme celles issues des lots 1 et 2. Dans un premier temps, elle s'attachera à étudier comment 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.

Thèse LRI : « Enrichir, restructurer et fusionner des ontologies par alignement »

Co-encadrement : Brigitte Safar, Chantal Reynaud

Descriptif du sujet :

Le thème central de la thèse porte sur l'alignement d'ontologies. L'équipe IASI du LRI dispose d'un environnement d'alignement, Taxomap, développé à l'origine dans le cadre du projet e.Dot (RNTL), pour interroger de façon unifiée des sources hétérogènes. La thèse consistera à partir de cet existant pour le confronter aux besoins du projet GEONTO. Il s'agira d'adapter et d'étendre le système Taxomap pour prendre en compte les spécificités des ontologies et des besoins reposant sur l'alignement dans le domaine de la topographie.

Dans un premier temps, il s'agira d'aligner des taxonomies réelles fournies par le COGIT (sous-lot 2.1), et également de contribuer à leur enrichissement (sous-lot 1.2) par des méthodes d'alignement (alignement de l'ontologie à enrichir avec d'autres ontologies ou ressources externes).

Dans un second temps, il s'agira de travailler sur les ontologies construites automatiquement à partir de textes fournies par le sous-lot 1.1 du projet dans l'objectif de les aligner (sous-lot 2.1) mais aussi les restructurer (sous-lot 1.3). Le processus d'alignement mis en œuvre devra permettre la fusion d'ontologies, si nécessaire (sous-lot 2.1). L'accent sera mis sur l'alignement de structures très hétérogènes (cf. description du sous-lot 2.1), nécessitant de ce fait la mise en œuvre de techniques d'alignement innovantes et originales. Pour l'aspect restructuration, l'originalité du travail consistera à étudier et exploiter les atouts d'une combinaison de techniques d'alignement et de techniques de construction automatique d'ontologie à partir de textes.

L'ensemble des résultats sera validé dans le domaine géographique mais les techniques et méthodes mises au point seront génériques.

Thèse LIUPPA :

Co-encadrement : Gaio Mauro, Sallaberry Christian

Descriptif du sujet :

La première hypothèse sous-tendant le sujet de cette thèse considère que dans des ressources d'informations, hétérogènes du point de vue formes d'expressions, mais homogènes du point de vue

référence à un territoire, « la régularité des références faites à des objets de ce territoire doit rendre possible la réalisation de chaînes d'analyses et d'interprétations sémantiques ciblées, raisonnablement générales et donc portables à coût contrôlé ». Cette hypothèse a déjà acquis une première assise grâce aux travaux initiés dans le cadre des projets GeoSem (CNRS) et PIV (CAPP/MIDR) qui ont abouti à la réalisation d'une chaîne automatique « minimale » d'extraction d'information et d'indexation, d'entités géographiques et de certaines expressions spatiales associées, à partir de documents textuels.

Il s'agit désormais d'étendre cette assise pour intégrer aux index produits des concepts plus étendus grâce à l'intégration dans la chaîne d'extraction et d'indexation des expressions géographiques d'une collection de ressources documentaires, de ou des ontologies produites dans ce projet.

L'approche que nous avons envisagée reprend globalement les principes des systèmes d'extraction et d'indexation traditionnels : un traitement « off-line » permet d'analyser la collection de documents et de produire des index. Puis, pour chaque besoin d'information formulé par un utilisateur, un moteur se charge de chercher dans les index les entrées témoignant des plus grands appariements.

Notre approche présentera néanmoins un certain nombre de spécificités quant aux méthodes qui seront utilisées pour l'indexation. Notamment elle ne s'appuiera pas uniquement sur un ensemble de mots clés mais sur les représentations conceptuelles (dédites grâce à l'ontologie) et géométriques (obtenues grâce à l'association entre termes de l'ontologie et champ d'une base de donnée géographique) des objets identifiés et de leurs relations. Il en résultera un pouvoir expressif beaucoup plus riche, pierre angulaire pour permettre une interrogation polyvalente pour tout public du simple utilisateur occasionnel à l'expert.