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