Rapport semestriel d'activité - coordonnateur
Programme MDCO - Edition 2007
Projet GEONTO – 1er semestre 2010

Identification

Acronyme du projet	GEONTO
Numéro d'identification de l'acte attributif	ANR-07-MDCO-05
Coordonnateur (société/organisme)	LRI – Université Paris-Sud
Période couverte (date à date)	01/01/2010 – 30/06/2010
Période couverte (t0+n mois à t0+m mois)	T0+24 à T0+30
Rédacteur (nom, téléphone, email)	Chantal Reynaud, 01 72 92 59 87 Chantal.reynaud@lri.fr
Date	12 juillet 2010

URL de la page web du projet et date de dernière mise à jour

http://geonto.lri.fr
La dernière mise à jour a été effectuée le 28/07/2010.

Activités de coordination des activités du projet
(lister les réunions, visites, …)

Réunion plénière :
21/06/2010 : réunion de bilan semestre 5 – LRI, COGIT, IRIT, LIUPPA - Orsay

Réunions de travail par lot :
13/01/2010 : réunion de travail – Lot 1 – IRIT, LIUPPA - Toulouse
16-17/02/2010 : réunion de travail – Lot 1 – IRIT, LIUPPA, COGIT - Toulouse
10/05/2010 : réunion de travail – Lot 1 – IRIT, LIUPPA, COGIT - Pau
18/01/2010 : réunion de travail – sous-lot 2.1 et 2.3 – LRI, COGIT – Saint Mandé
25/02/2010 : réunion de travail – lot 2 – LRI, COGIT - Orsay

Synthèse

Numéro du Partenaire	Conformité des résultats obtenus aux prévisions	Conformité de la consommation des ressources par rapport aux prévisions	Difficultés particulières
1	conforme aux prévisions mises à jour dans la demande de prolongation du projet	conforme	Des retards dans la production de l'ontologie via des techniques automatiques de TAL (lot 1) se répercutent sur l'avancement des travaux portant sur l'alignement et l'enrichissement (lot 2). Retard dans la livraison du logiciel de réconciliation d'instances pour la réconciliation d'ontologies du fait de

			l'embauche d'un ingénieur uniquement à partir du 1/4/10.
2	conforme	conforme	aucune
3	Conforme au nouveau calendrier	Conforme aux nouvelles prévisions	Pas de difficulté particulière
4	conforme	Conforme (compte tenu du décalage précédemment annoncé en terme de recrutement du doctorant)	
Synthèse	Conforme aux prévisions mises à jour dans la demande de prolongation du projet	Conforme	Des retards dans la production de l'ontologie via des techniques automatiques de TAL (lot 1) se répercutent sur l'avancement des travaux portant sur l'alignement et l'enrichissement (lot 2). Retard dans la livraison du logiciel de réconciliation d'instances pour la réconciliation d'ontologies du fait de l'embauche d'un ingénieur uniquement à partir du 1/4/10.

Faits marquants

Indiquer les résultats et/ou réalisations marquants. Préciser s'ils peuvent ou non faire l'objet de communications externes par l'ANR et la Délégation ANR-CI.

Les travaux du lot 1 ont porté sur les aspects suivants :

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

Une 3ème chaîne de traitement de texte a été développée en java de façon à utiliser des primitives de représentation OWL, non accessibles sous *GATE*. Elle permet, à partir des documents de spécification, de produire une ontologie comportant 2 parties : un modèle de la carte et un modèle du « monde » décrit par la carte. Elle est adaptée à la structure du document de spécification BD-Topo. Elle s'appuie sur la définition d'un modèle de représentation des spécifications des bases de données géographiques en OWL2 qui a été mis au point en collaboration avec le Cogit.
Par ailleurs, une partie de la chaîne de traitement requise pour analyser le langage naturel présent dans les documents de spécification a été réalisée en utilisant la plateforme *LinguaStream*. Le résultat visé est une représentation en OWL de propositions de concepts et de relations. Ce travail sera achevé fin Août, date de fin du contrat de post-doctorant de M. Laignelet de l'IRIT.
Enfin, un module de correction des étiquettes des concepts de l'ontologie construite automatiquement a été réalisé. Les corrections sont soit automatiques soit proposées à l'expert. Ce travail a fait l'objet d'une publication à IC2010.

- Sous-lot 1.2 et 1.3 : Enrichissement et restructuration d'une ontologie existante

L'IRIT et le LIUPPA ont travaillé à la mise en place des principes qui guident l'enrichissement de l'ontologie à partir des termes trouvés dans le corpus grand public analysé par le LIUPPA, rapprochés des entrées du thesaurus Rameau. Le LIUPPA a notamment produit une première version de la chaîne de traitement *AugmOnto* pour alimenter ce processus d'enrichissement. Ce travail a fait l'objet d'une publication à TALN 2010.
En parallèle, un travail a été réalisé en collaboration avec le LRI pour spécifier des traitements d'enrichissement d'ontologies au sein de la plate-forme *TaxoMap Framework*. Il s'agit d'un travail préliminaire qui sera poursuivi durant le prochain semestre en travaillant sur la version de

l'ontologie Topo-IRIT livrée en juillet 2010 et à partir des propositions d'enrichissement fournies par le LIUPPA.

Les travaux du lot 2 réalisés sont les suivants :

- Sous-lot 2.1 : Alignement d'ontologies

Le module d'affinement de mappings de *TaxoMap Framework* a été complètement conçu et implémenté. Les patrons qui ont été définis permettent d'affiner les mappings issus de l'alignement de Topo-Cogit et Carto-Cogit. Le Cogit a participé au processus de validation des mappings et de définition des patrons utiles pour les affiner. Ces travaux ont donné lieu à 3 publications, RFIA 2010, Int. Symposium on Matching and Meaning, EKAW 2010 (papier long).

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

Une approche permettant d'attaquer de front le problème de l'alignement d'ontologie et celui de la réconciliation d'instances a été développée par le LRI. Des expérimentations basées sur l'alignement de TopoCarto et des ontologies DBPedia, Yagoo et KIM ont été réalisées. L'implémentation est en cours (non terminée du fait de l'embauche d'un ingénieur assez tardive).

- Sous-lot 2.3 : Analyse des différences entre ontologies

Deux types de travaux sont réalisés en parallèle. Côté LRI, le travail réalisé repose sur l'étude de l'utilisation d'une mesure de similarité entre ontologies développée par Maedche & Staabe combinée avec l'outil de partitionnement *TaxoPart*. Des expérimentations préliminaires ont été réalisées. Côté Cogit, une méthode de comparaison globale d'ontologies a été proposée. Des tests sur des ontologies réelles ont été réalisés. Ce travail a été publié à IC 2010 sous forme d'un poster.

Les travaux du lot 3 réalisés sont les suivants :

- Sous-lot 3.1 : Indexation automatique du contenu des documents

Finalisation de la conception et de la spécification du module logiciel « Indexation automatique du contenu des documents » par le LIUPPA, avec validation des toponymes à partir de ressources locales ou distantes et construction d'une sortie XML.
Conception d'une structure XML pour la production des index après récupération automatique locale ou distante des géométries des toponymes valides.

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

Un modèle de représentations en OWL2 des spécifications des bases de données géographiques a été défini en collaboration avec l'IRIT de façon à dissocier le modèle de la carte, du modèle du « monde ». Des tests ont été effectués sur des spécifications réelles. Ce travail a donné lieu à une publication à la conférence EKAW 2010 acceptée sous la forme d'un poster.

L'article du consortium à la conférence Sageo de novembre 2009 a été sélectionné pour être publié dans la Revue Internationale de Géomatique.

Publications liées au projet :

<u>Conférences et ateliers internationaux (mono-partenaires)</u>

N. Abadie, A. Mechouche, S. Mustière, **OWL based formalisation of geographic databases specifications**, EKAW 2010, 17th International Conference on Knowledge Engineering and Knowledge Management, Poster, 11th October-15th October 2010, Lisbon, Portugal.

F. Hamdi, C. Reynaud, B. Safar, **A framework for mapping refinement specification**, in Proceedings of the International Symposium on Matching and Meaning, Michael Chan and Fiona McNeill (Eds.), at the AISB 2010 convention, 29 March - 1 April 2010, De Montfort University, Leicester, UK.

F. Hamdi, C. Reynaud, B. Safar, **Pattern-based Mapping Refinement**, EKAW 2010, 17th International Conference on Knowledge Engineering and Knowledge Management, 11th October-15th October 2010, Lisbon, Portugal.

Conférences nationales (mono-partenaires)

F. Hamdi, C. Reynaud, B. Safar, **L'approche TaxoMap Framework et son application au raffinement de mappings**, Congrès Francophone Reconnaissance des Formes et Intelligence Artificielle, RFIA 2010, Caen, 19-22 janvier 2010.

M. Kamel, N. Aussenac-Gilles, M. Laignelet, **Correction d'ontologies construites à partir de la structure de documents**, 21èmes Journées Francophones d'Ingénierie des Connaissances, Journées Francophones d'Ingénierie des Connaissances, Nîmes (France), Sylvie Despres (Eds.), Ecole des Mines d'Alès, p. 29-40, 8-11 Juin 2010.

A. Mechouche, N. Abadie, S. Mustière, **Mesure de la distance sémantique entre parties partiellement communes à deux taxonomies**, 21èmes Journées Francophones d'Ingénierie des Connaissances, IC 2010, Poster, Nîmes, 8-11 Juin 2010.

M.-N. Bessagnet, M. Gaio, E. Kergosien, C. Sallaberry, **Extraction automatique d'un lexique à connotation géographique à des fins ontologiques dans un corpus de récits de voyages**, TALN 2010, 19-23 juillet, Montréal.

Revue internationale (multi-partenaires)
L'article du consortium à la conférence Sageo de novembre 2009 a été sélectionné pour être publié dans la Revue Internationale de Géomatique. L'article suivant a été soumis :

S. Mustière, N. Abadie, N. Aussenac-Gilles, M.-N. Bessagnet, M. Kamel, E. Kergosien, C. Reynaud, B. Safar, C. Sallaberry, **Analyses linguistiques et techniques d'alignement pour créer et enrichir une ontologie topographique**.

Difficultés rencontrées

L'avancement des travaux a été ralenti par le fait qu'aujourd'hui nous ne disposons pas encore d'une version de Topo-IRIT prête à être enrichie. Ceci est dû en particulier au recrutement du post-doctorant de l'IRIT tardif (Marion Laignelet a été recrutée seulement en octobre 2009). La demande de prolongation déposée en début d'année intégrait toutefois les retards générés par ce problème.

Le Cogit accuse un retard de quelques mois suite au recrutement différé d'un post-doc (prévu début 2009, début effectif juin 2009) et à une charge de travail du laboratoire imprévue en 2009.

Le LIUPPA accuse également un retard qui a été pris en compte dans la demande de prolongation du projet.

Du point de vue des difficultés rencontrées, il n'a pas été possible de définir une chaîne de traitement paramétrable et applicable à d'autres textes que ceux conformes à la DTD, ce qui remet

en question la production d'un programme unique pour toutes les spécifications des bases de données du Cogit. Par ailleurs, tous les traitements nécessaires à la production de l'ontologie n'ont pu être définis au sein de la même plate-forme informatique. Enfin, la construction d'ontologies de façon totalement automatique s'avère irréaliste. Des traitements manuels sont nécessaires avant d'appliquer des techniques d'alignement.

Suivi des délivrables du projet (d'après le planning accepté lors de la demande de prolongation)

(exemple, le tableau initial est celui contenu en annexe 1)

	Libellé	Nat.	Partenaires	Date	08 S1	08 S2	09 S1	09 S2	10 S1	10 S2
T0	**Coordination – Communication**									
T0a	Mise en place d'une page web pour le projet		Tous	Début 2008	A					
T0b	Mise à jour page Web		Tous	Régulièrement	A	A	A			
T0c	Réunion de lancement		Tous	18/01/08	A					
T0d	Réunion de bilan semestre 1		Tous	13/06/08	A					
T0e	Réunion de bilan semestre 2		Tous	23/01/09		A				
T0f	Réunion de bilan semestre 3		Tous	30/06/09			A			
T0g	Réunion de bilan semestre 4		Tous	04/12/09				A		
T0h	Réunion de bilan semestre 5		Tous	21/06/10					A	
T1	**Lot 1 Construction et enrichissement d'ontologies**									
T1a	Mise au point d'outils d'extraction de concepts et de relations : rapport intermédiaire	R	IRIT, LIUPPA, COGIT	Fin S2		X	A			
T1b	Mise au point d'outils d'extraction de concepts et de relations	Logiciel	IRIT	Fin S3			A			
T1c	Enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction et à partir des ressources lexicales	Logiciel	LIUPPA	Fin S3			A	A	A*	R1
T1d	Mise au point d'outils d'extraction de concepts et de relation	R Module logiciel	IRIT	Fin S4				A	A	
T2	**Lot 2 Appariement d'ontologies hétérogènes**									
T2a	Alignement d'ontologies : rapport intermédiaire	R	LRI, COGIT	Fin S2		A				
T2b	Réconciliation d'instances pour l'alignement d'ontologies	R Logiciel	LRI	Fin S5					A (R)	R2 (Log)
T3	**Lot 3 Exploitation des ontologies créées**									
T3a	Intégration et accès aux schémas des bases de données	R	Cogit	Fin S3			A			
T3b	Indexation automatique de contenu de documents	R	LIUPPA	Fin S3			A			

Nat. : CR = Compte-rendu, R = rapport, …

X : prévision initiale
A : atteint – A* : version livrée non finale
R1, R2, … : reprévision

Commentaires
Préciser en particulier la raison de chaque reprévision de délivrables (Ri)

Concernant le livrable n° 4 (T1c) portant sur l'enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction et à partir de ressources lexicales, on peut distinguer la partie extraction à partir de textes d'une liste de termes candidats (fait par le LIUPPA) de la partie enrichissement de l'ontologie. La chaîne de traitement réalisant l'extraction intégrant *AugmOnto* qui produit des propositions d'enrichissement est livrée ce semestre par le LIUPPA. Une étude de

l'enrichissement réalisé à l'aide de *TaxoMap Framework* à partir des propositions générées par le LIUPPA vient de débuter au LRI. Les résultats seront livrés en T0+36.

Concernant le livrable n°7 (T1d) portant sur la mise au point d'outils d'extraction de concepts et de relations, une première version du logiciel a été livrée en T0+24. Nous livrons fin juillet la version courante. Le rapport ainsi qu'une version plus élaborée seront livrés en septembre 2010.

Enfin, concernant le livrable n°10 (T2b), le rapport est livré en T0+30 mais le logiciel ne sera livré qu'en T0+36 (retard dans l'embauche d'un ingénieur pour réaliser l'implémentation).

Liste des CDD recrutés par des établissements publics dans le cadre du projet
Lister ici tous les CDD recrutés depuis le début du projet.

Numéro du Partenaire	Nom	Prénom	Qualifications	Date de recrutement	Durée du contrat (en mois)
1	HAMDI	Fayçal	Stagiaire M2 recherche	10/03/2008	6 mois
1	HAMDI	Fayçal	Doctorant	05/11/2008	24 mois (avec l'objectif de prolonger de 12 mois)
4	NGUYEN	Van Tien	Doctorant	17/11/2008	36 mois (renouvelable par année)
2	MECHOUCHE	AMMAR	Post-doctorant	18/05/2009	18 mois
3	LAIGNELET	Marion	Post-doctorant	01/10/2009	12 mois à 4/5 de temps
3	CAPELLE	Jérôme	Stage L3	01/07/2009	1 mois
1	NIRAULA	Nobal	Ingénieur	01/04/10	3 mois et 2 semaines

Equipements achetés par les partenaires dans le cadre du projet
Lister ici tous les équipements achetés depuis le début du projet

Numéro du Partenaire	Désignation	Date d'achat	Prix d'achat (en Euros)	Part financée par l'aide ANR (en Euros)
1	Mac Pro (sans écran)	Décembre 2008	2 036,94	2 036,94
4	Disque Dur 250	Mars 2008	119,79	119,79
4	2 Mémoires DDR 333Mhz 1go	Mai 2008	157,87	157,87
4	2 Mémoires SODIMM DDR 333Mhz 1go	Juin 2008	124,38	124,38
4	Portable pour doctorant	Décembre 2008	1205,38	1205,38
4	Ecran de bureau pour doctorant	Décembre 2008	249,00	249,00
3	Ordinateur individuel	Mars 2009	1300,00	1300,00
1	PC portable	Décembre 2009	1227 ,09	1227,09
2	Deux PC	Septembre 2009	2 x 822,95	2 x 822,95
3	4 PC, écrans, licences logiciels	Octobre 2009	5204,00	5204,00
4	Disque dur	Mars 2009	119,79	119,79
4	2 mémoires DDR 333Mhz 1 go	Mai 2009	157,87	157,87
4	2 mémoires SODIMM DDR 333 Mgz 1 go	Juin 2009	124,38	124,38
4	Portable pour doctorant + écran de bureau	Décembre 2009	1454,38	1454,38
4	Portable pour chercheur + station + écran de bureau	Mai 2010	2578,79	2578,79