

**Rapport semestriel d'activité - coordonnateur**  
**Programme MDCO - Edition 2007**  
**Projet GEONTO – 1er semestre 2011**

**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 – 30/06/2011
Période couverte (t0+n mois à t0+m mois)	T0+36 à T0+42
Rédacteur (nom, téléphone, email)	Chantal Reynaud, 01 72 92 59 87 chantal.reynaud@lri.fr
Date	26 juillet 2011

**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 22 juillet 2011.

**Activités de coordination des activités du projet**

*(lister les réunions, visites, ...)*

Réunion plénière :

24/06/2011 : réunion de bilan semestre 7 – LRI, COGIT, IRIT, LIUPPA - Pau

Réunions de travail par lot :

Nous n'avons pas eu de réunions par lot. Le travail de chacun a consisté à poursuivre les développements précédemment spécifiés. Les échanges entre partenaires se sont faits essentiellement par courrier électronique ou téléphone.

**Synthèse**

Numéro du Partenaire	Conformité des résultats obtenus aux prévisions	Conformité de la consommation des ressources par rapport aux prévisions	Difficultés particulières
1	Conforme	Conforme	Aucune
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	Aucune

		de recrutement du doctorant)	
Synthèse	Conforme	Conforme	Aucune difficulté particulière

### **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

La chaîne de traitement définie au S6 par l'IRIT en utilisant la plateforme LinguaStream pour analyser le langage présent dans les documents de spécification a subi quelques évolutions suite aux travaux sur l'enrichissement décrits ci-dessous. Les règles d'extraction ont été notamment affinées de façon à mieux gérer l'identification des labels des concepts et la nature des relations. Une nouvelle version de la chaîne de traitement sera livrée en septembre 2011.

Le travail réalisé par l'IRIT pour construire une ontologie à partir de textes via des patrons prenant en compte la structure des textes a été valorisée par une publication à TALN 2011.

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

L'IRIT a développé un module d'enrichissement d'ontologie au format OWL, s'appuyant sur le logiciel d'alignement d'ontologies TaxoMap. Il prend en entrée des fragments d'ontologie et propose systématiquement de les associer à l'ontologie existante. Une interface graphique permet à un expert de valider les propositions générées. Les premières tentatives d'enrichissement ont permis d'affiner les règles d'extraction des fragments extraits de textes définies antérieurement. L'évaluation du logiciel d'enrichissement doit être poursuivie, de même que son application pour enrichir TOPO-IRIT. Le code du module d'enrichissement sera livré en septembre 2011.

TopoCarto\_Cogit a été enrichie par le LRI en appliquant TaxoMap FrameWork sur les graphes construits à partir des concepts venant des définitions des spécifications générées par l'IRIT et situés dans RAMEAU. Les différents patterns d'enrichissement décrits dans le livrable n°9 ont été implémentés et introduits dans TaxoMap Framework. Ce travail a été valorisé par une publication à EGC 2011.

Les travaux réalisés antérieurement par le LIUPPA ont été valorisés par des publications à GeoS'2011 et à la 30<sup>ème</sup> conférence internationale « Lexis and Grammar » (papier accepté pour être présentée en octobre 2011).

Les travaux sur la restructuration n'ont pas donné lieu à des développements particuliers. Le logiciel utilisable est TaxoMap Framework, livré antérieurement. Les patrons spécifiques pour cette fonctionnalité n'ont pu être créés du fait de la non disponibilité de l'ontologie sur laquelle la restructuration devait être réalisée.

Les travaux du lot 2 réalisés sont les suivants :

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

Le développement du logiciel associé aux travaux décrits dans le livrable n°10 remis en juillet 2010 a été terminé et peut être livré.

- Sous-lot 2.3 : Analyse des différences entre ontologies

La valorisation des travaux sur la comparaison d'ontologies s'est poursuivie par une publication à EGC 2011.

Les travaux du lot 3 réalisés sont les suivants :

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

Une collaboration entre le COGIT et le LIUPPA a permis de construire une ontologie utilisée dans les travaux d'indexation du LIUPPA. Cette ontologie s'appuie sur l'ontologie initiale du COGIT au sein de laquelle ont été intégrés manuellement les termes extraits de récits de voyage complétés par des extraits du thesaurus RAMEAU.

Cette ontologie a été utilisée pour le typage des entités spatiales (ES). Un service RechercheConcept de parcours d'ontologie a ainsi pu être développé. Pour un qualificatif donné, contenu dans le syntagme nominal de l'ES, ce service consulte l'ontologie et retourne une liste de couples (*concept, pointeurRessource*). L'analyse du syntagme nominal correspondant aux ES permet de les catégoriser (oronyme, hydronyme, ...). L'élément d'information *pointeurRessource* associé au concept désigne la ressource (BD et table) dans laquelle l'ES est décrite. Les annotations apposées aux ES sont utilisées dans la phase d'interprétation des ES pour les rechercher dans les ressources spécifiques et compléter leur description (géométrie). La phase d'interprétation, qui a pu être optimisée, utilise une seconde ontologie (LSO) issue des spécifications des bases de données. Toponymes.owl, mise au point par le COGIT suite à des discussions avec l'IRIT, intègre une partie des spécifications de BD-Topo. Cette ontologie permet de faire le lien entre l'ontologie du domaine et la base de données.

Le LIUPPA a travaillé sur un ensemble de 11 livres composés de récits de voyage. 15572 ES ont été obtenues.

Des expérimentations ont été réalisées pour tester l'apport d'une ontologie en indexation par rapport à l'utilisation d'un lexique ou d'aucune ressource externe. Les résultats obtenus confirment l'idée que l'utilisation d'une ontologie géographique améliore grandement la chaîne d'indexation.

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

Poursuite de la thèse sur l'appariement de schémas à partir de la formalisation de leurs spécifications.

- Sous-lot 3.3 : Mise à disposition des ontologies réalisées

Suite aux différents outils et techniques mises au point au cours du projet, et aux différentes expérimentations réalisées, les membres du projet se sont mis d'accord, lors de la réunion plénière du 24/06/2011, sur la chaîne de traitement à appliquer de façon à obtenir fin octobre une ontologie topographique riche, en accord avec les besoins applicatifs.

L'idée est de repartir de l'ontologie construite à la main par el COGIT et le LIUPPA (Fusion2), et de l'enrichir à partir des mappings générés par le LRI obtenus après application de TaxoMap sur Fusion2 et sur les graphes (mini-ontologies) produits par le LIUPPA provenant des récits de voyage.

- Les concepts provenant des graphes du LIUPPA, proposés automatiquement pour l'enrichissement à l'aide de TaxoMap, seront introduits avec Protégé dans Fusion2, s'ils n'y figurent pas déjà. La version de l'ontologie obtenue sera alors Fusion3.

- Fusion3 sera enrichie avec les fragments extraits des définitions de BD-TOPO, issus du travail de M. Laignelet, et conduira à l'obtention de Fusion4.

- Les mappings de type « grotte marine » is-a « grotte » seront communiqués par le LRI au COGIT. Les termes correspondant aux concepts considérés comme plus spécifiques seront listés. Le COGIT introduira des propriétés au bon niveau de l'ontologie pour prendre en compte ces concepts plus spécifiques.

- Un alignement de Fusion4 avec Topo-IRIT, issue de l'application des traitements automatiques de langage naturel sur les spécifications de BD-TOPO, sera réalisé.

### **Publications liées au projet**

#### Revue internationale (multi-partenaires)

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, Revue Internationale de Géomatique (RIG), vol. 21 – n°2/2011, p. 155-180, Hermès.

#### Conférences et ateliers internationaux (mono-partenaires)

M. Gaio, V. T. Nguyen, Towards Heterogeneous Resources-based Ambiguity Reduction of sub-typed Geographic Named Entities, in 4<sup>th</sup> Int. Conf. of Geospatial Semantics, LNCS 6631, p. 217-234, Brest, 2011.

N. Pernelle, F. Saïs, LDM: Link Discovery Method for new Resource Integration. In proceedings of RED Fourth International Workshop on REsource Discovery (ESWC 2011), 29-30 mai, 2011.

Notons également une soumission acceptée en papier long à 30th international conference on lexis and grammar, « Utilisation de la relation «verbe–préposition–toponyme » pour un inventaire lexical automatique », Cyprus Octobre 2011.

#### Conférences et ateliers d'audience nationale (mono-partenaires)

F. Hamdi, B. Safar, C. Reynaud, 2011. Utiliser des résultats d'alignement pour enrichir une ontologie, EGC'2011, Brest, 25-28 janvier 2011.

Mechouche A., Abadie N., Prouteau E., Mustière S., 2011, Utilisation d'une ontologie du domaine pour la découverte du contenu de bases de données géographiques, 11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'2011), 25-28 janvier 2011, Brest (France), pp 569-574

Mechouche A., Abadie N., Mustière S., 2011, Une mesure de distance dans l'espace des alignements entre parties potentiellement homologues de deux ontologies légères, 11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'2011), 25-28 janvier 2011, Brest (France), pp 329-340

M. Laignelet, M. Kamel, N. Aussenac-Gilles. Enrichir la notion de patron par la prise en compte de la structure textuelle - Application à la construction d'ontologie (short paper). Dans Traitement Automatique des Langues Naturelles (TALN 2011), Montpellier (F), 27/06/2011-01/07/2011, Vol. 2, Mathieu Lafourcade, Violaine Prince (Eds.), LIRMM, p. 267-272, juin / *june* 2011.

URL : [http://www.lirmm.fr/~lopez/TALN2011/PDF\\_court/Laignelet\\_taln11\\_submission\\_154.pdf](http://www.lirmm.fr/~lopez/TALN2011/PDF_court/Laignelet_taln11_submission_154.pdf)

## Retombées intéressantes

Notons que les collaborations initiées dans le cadre du projet GéOnto ont permis d'initier un réseau de recherche entre les participants appelé à perdurer après le projet : engagement prévu par le COGIT d'engager un doctorant du LRI après sa thèse, prise de contact par un industriel avec le COGIT et l'IRIT lors de la réponse à un appel d'offre d'étude pour la défense, mise en place d'une thèse en co-direction COGIT/LIUPPA.

## Difficultés rencontrées

Un des points difficiles est l'intégration des modules logiciels réalisant les différents traitements : hétérogénéité des logiciels et des langages.

L'IRIT confirme par ailleurs la difficulté de définir une chaîne de traitement paramétrable et applicable à d'autres textes que ceux conformes à la DTD. Certaines parties du programme d'analyse des fichiers XML sont même complètement spécifiques au document BD-Topo, ce qui remet en question la possibilité de produire un programme unique pour toutes spécifications des bases de données du COGIT.

## Suivi des livrables 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	11 S1
<b>T0</b>	<b>Coordination – Communication</b>										
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		
T0i	Réunion de bilan semestre 6		Tous	18/11/10						A	
T0j	Réunion de bilan semestre 6		Tous	24/06/11							A
<b>T1</b>	<b>Lot 1 Construction et enrichissement d'ontologies</b>										
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		
T1d	Mise au point d'outils d'extraction de concepts et de relation	R Module logiciel	IRIT	Fin S4				X		A	
T1e	Enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction	R Logiciel	LRI	Fin S6						R2	A
T1f	Restructuration d'une ontologie construite automatiquement	Module logiciel	LRI	Fin S7							A
<b>T2</b>	<b>Lot 2 Appariement d'ontologies hétérogènes</b>										
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	R1 (Log)	A
T2c	Alignement d'ontologies	R Logiciel	LRI	Fin S6						A	

T2d	Analyse des différences entre ontologies pour faire ressortir les différences de points de vue sous-jacentes	R	COGIT	Fin S6						A	
<b>T3</b>	<b>Lot 3 Exploitation des ontologies créées</b>										
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 livrables (Ri)*

Des livraisons complémentaires seront effectuées courant septembre 2011 :

- Module d'enrichissement développé à l'IRIT
- Nouvelle version de la chaîne de traitement conçue à l'IRIT

### **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 recherche M2	10/03/2008	6 mois
1	HAMDI	Fayçal	Doctorant	05/11/2008	36 mois
4	NGUYEN	Van Tien	Doctorant	17/11/2008	36 mois (renouvelable par année)
2	MECHOUCHE	AMMAR	Post-doctorant	18/05/2009	16,5 mois (contrat initial de 18 mois achevé 1,5 mois avant la fin, le candidat ayant été nommé sur un poste d'ATER)
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	01/04/2010	3 mois et 2 semaines
2	PROUTEAU	Emeric	Stagiaire (master)	08/03/2010	5,4 mois
2	SOUALAH-ALILA	Fayrouz	Stagiaire Master (3A)	18/04/2011	4,5 mois
3	DIALLO	Abdoul	M2R – Stage	01/03/2011	5 mois
3	CARASSUS	Vincent	L3 – stage	20/04/2011	4 mois

### **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
4	Portable pour chercheur	Décembre 2010	1147,83	1147,83