

**Rapport semestriel d'activité - coordonnateur**  
**Programme MDCO - Edition 2007**  
**Projet GEONTO – 1<sup>er</sup> semestre 2009**

**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)	01/01/2009 – 30/06/2009
Période couverte (t0+n mois à t0+m mois)	T0+12 à T0+18
Rédacteur (nom, téléphone, email)	Chantal Reynaud, 01 72 92 59 87 Chantal.reynaud@lri.fr
Date	17 juillet 2009

**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 20/07/2009

**Activités de coordination des activités du projet**

*(lister les réunions, visites, ...)*

**Réunions plénières :**

23 /01/2009 : réunion de bilan semestre 2 – Orsay – Diane Penel (Délégation ANR – calcul intensif) a présenté le fonctionnement des projets ANR.

30/06/2009 : réunion de bilan semestre 3 – Toulouse.

**Réunions de travail par lot :**

20/01/2009 : réunion de travail – Lot 2 – LRI, COGIT – Saint Mandé

09/03/2009 : réunion de travail téléphonique – Lot 2 – LRI, COGIT

18/03/2009 : réunion de travail – Lot 1 – IRIT, LIUPPA, Toulouse

20/05/2009 : réunion de travail – Lots 1 et 3 – COGIT, IRIT, LIUPPA, Pau

08/06/2009 : réunion de travail – Lot 1 – IRIT, LIUPPA

09/07/2009 : réunion de travail – Lot 0 – LRI, COGIT – Saint Mandé

## Synthèse

Numéro du Partenaire	Conformité des résultats obtenus aux prévisions	Conformité de la consommation des ressources par rapport aux prévisions	Difficultés particulières
1	inférieurs aux prévisions	conforme	Retard pris dans les travaux du fait de l'embauche d'un doctorant uniquement à partir d'octobre 2008
2	conformes	conforme	aucune
3	conformes mais leur intégration et réutilisabilité sont moindres	Moindre, en particulier du budget « financement spécifique » de l'IRIT	Recrutement différé d'un étudiant en post-doctorat (recherche d'un nouvel étudiant après abandon du candidat pressenti).
4	conformes	Conforme (compte tenu du décalage annoncé en terme de recrutement du doctorant)	
Synthèse	Résultats légèrement inférieurs aux prévisions	Conforme pour 3 partenaires sur 4, inférieure pour 1 partenaire	Des retards dans l'embauche du personnel pour tous les partenaires

## 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 d'une part sur la conception de la chaîne de traitement permettant de construire automatiquement une ontologie fidèle aux spécifications associées aux bases de données géographiques du Cogit (sous-lot 1.1), d'autre part sur l'enrichissement d'ontologie (sous-lot 1.2).

Les travaux du sous-lot 1.1 ont été complétés par l'analyse du contenu des spécifications (champs définitions). Les traitements exploitant la structure et le contenu des documents ont été adaptés de façon à s'exécuter sur des schemas XML inspirés des normes ISO communiqués par le Cogit de façon à rendre l'approche d'extraction davantage générique. Une comparaison de l'ontologie construite (Topo-IRIT) et de celle élaborée par le Cogit (Topo-Cogit) a été réalisée ainsi que son alignement avec TopoCarto-Cogit à l'aide de l'outil d'alignement TaxoMap du LRI (travaux en lien avec le lot 2). Les résultats d'alignement obtenus montrent que leur exploitation peut permettre d'enrichir TopoCarto-Cogit. L'alignement est un préalable à l'enrichissement. Il produit des résultats riches qui doivent ensuite faire l'objet de traitements spécifiques complémentaires, en interaction avec l'expert. Un environnement d'aide à la spécification de tels traitements est en cours de conception. Une démo de l'exécution de la chaîne de traitement basée sur l'environnement GATE élaborée dans le sous-lot 1.1 a été mise en ligne sur le site du projet.

Les travaux du sous-lot 1.2 ont porté sur l'extraction de termes associés à des entités nommées dans un échantillon élargi de textes issus du corpus « récits de voyage » dans le but d'obtenir un ensemble de termes-concepts candidats à l'enrichissement d'une ontologie existante (TopoCarto-Cogit). La chaîne de traitement utilisée pour l'extraction est ici basée sur l'environnement LINGUASTREAM. Cette liste de termes a été complétée grâce au thesaurus RAMEAU. Par

ailleurs, l'ontologie ITINERAIRES, précédemment construite, a été traduite en OWL, projetée sur les textes des récits de voyage et une chaîne de traitement dédiée au repérage des descriptions d'itinéraires a alors été conçue. Elle a donné lieu au développement d'un prototype intégré à l'environnement LINGUASTREAM.

Les travaux du lot 2 ont porté sur l'alignement d'ontologies (sous-lot 2.1), la réconciliation d'instances pour l'alignement d'ontologies (sous-lot 2.2) et la comparaison d'ontologies (sous-lot 2.3).

Les travaux du sous-lot 2.1 ont tout d'abord consisté à améliorer l'outil d'alignement TaxoMap (mesure de similarité et techniques adaptées) de façon à produire des résultats d'alignement des ontologies Topo-Cogit et Carto-Cogit de meilleure qualité. De nouveaux tests ont été réalisés sur ces ontologies et les résultats sont en cours de validation par le Cogit. Ces travaux se sont poursuivis par la conception d'une plate-forme d'alignement intégrant TaxoMap et permettant d'affiner encore plus les résultats d'alignement produits de façon à tenir compte des particularités des ontologies alignées. Cette plate-forme devra permettre à l'expert de spécifier de façon déclarative les affinements à réaliser. Elle devra également à terme permettre la fusion, l'enrichissement et la restructuration d'ontologies.

Les travaux du sous-lot 2.2 ont consisté à étudier l'applicabilité de l'approche LN2R mise en œuvre au LRI sur un sous-ensemble des données des bases BDTopo et BDCarto. Il en résulte que des décisions de non-réconciliation peuvent être obtenues mais que les mesures de similarité utilisées doivent être ajustées pour tenir compte des particularités des données géographiques.

Une étude bibliographique a été réalisée conjointement par le LRI et le Cogit sur la comparaison d'ontologies dans le cadre du sous-lot 2.3.

Les travaux du lot 3 concernent l'exploitation des ontologies créées. La première application cible concerne l'indexation spatiale de documents (sous-lot 3.1). Une étude des apports sémantiques de l'ontologie topographique et des questions de recherche soulevées par ce cadre applicatif a été réalisée. La seconde application concerne l'intégration de données topographiques (sous-lot 3.2). Plusieurs scénarios d'intégration ont été identifiés selon l'existence ou non de spécifications associées aux bases de données. Les développements effectués jusqu'alors concernent le cas où on ne dispose pas de spécifications.

## **Publications liées au projet :**

### Conférences et ateliers internationaux

N. Abadie, 2009, Formal specifications to automatically identify heterogeneities, 12th AGILE International Conference on Geographic Information Science Pre-Conference Workshop "Challenges in Spatial Data harmonisation", 2 June 2009, Hannover, Germany.

N. Abadie, 2009, Schema Matching Based on Attribute Values and Background Ontology, 12th AGILE International Conference on Geographic Information Science, 2-5 June, Hannover, Germany.

M. Kamel, N. Aussenac-Gilles, 2009, Ontology Learning by Analysing XML Document Structure and Content, International Conference in Knowledge Engineering and Ontology Development (KEOD), Oct. 2009, Madeira, Portugal, 6/11/2009 – 8/11/2009, Jan Dietz (Eds.), INSTICC – Institute for Systems and Technologies of Information, Control and Communication, p. 1-6, novembre 2009 (à paraître).

### Conférences nationales

M. Kamel, N. Aussenac-Gilles, 2009, Construction automatique d'ontologies à partir de spécifications de bases de données, Dans Journées Francophones d'Ingénierie des Connaissances (IC 2009) 2009, Plate-forme AFIA / Hammamet, Tunisie, 25-29 mai 2009, Fabien Gandon (Eds.), Université Hassan II, p. 85-96, Mai 2009.

#### Rapport interne

S. Mustière, C. Reynaud, B. Safar, N. Abadie, Same words ? Same worlds ? Comparing ontologies underlying geographic data, Rapport interne n°1521, LRI - Université Paris-Sud, Juin 2009.

#### **Articles soumis :**

F. Hamdi, B. Safar, C. Reynaud, H. Zargayouna, "Alignment-based Partitioning of Large-scale Ontologies", chapter in Advances in Knowledge Discovery and Management, Studies in Computer science, Springer.

S. Mustière, N. Abadie, N. Aussenac-Gilles, M.-N. Bessagnet, M. Kamel, E. Kergosien, C. Reynaud, B. Safar, „GéOnto : Enrichissement d'une taxonomie de concepts topographiques“, Spatial Analysis and GEomatics (SAGEO), Paris, 25-26 Nov. 2009.

#### Difficultés rencontrées

Les difficultés rencontrées concernent le retard dans le recrutement de personnel :

- doctorant au LRI recruté le 5/10/2008 au lieu du 1/1/2008
- post-doctorant au Cogit recruté fin mai 2008 au lieu de début 2008
- post-doctorant à l'IRIT recrutée uniquement en septembre 2009

Les dates de début des travaux ont été toutefois respectées mais on note un retard dans la progression du travail réalisé.

Ces retards nécessiteront une demande de prolongation du projet d'un an de façon à assurer le financement de la dernière année de thèse du doctorant du partenaire 1 (LRI). Par ailleurs, l'Université, l'UPPA, ne permettant pas de rémunération de stage, les montants prévus initialement pour la rémunération de stages par le partenaire 4 (LIUPPA) ont été transférés en contribution complémentaire au salaire du doctorant afin d'assurer en partie le financement de sa dernière année.

#### Suivi des livrables du projet

(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
<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			
<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	Logiciel	IRIT	Fin S3			A			

	concepts et de relations								
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	R1	
<b>T2</b>	<b>Lot 2 Appariement d'ontologies hétérogènes</b>								
T2a	Alignement d'ontologies : rapport intermédiaire	R	LRI, COGIT	Fin S2			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

R1, R2, ... : reprévision

### **Commentaires**

*Préciser en particulier la raison de chaque reprévision de livrables (Ri)*

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 correspond à une version V1 partielle du logiciel. La version V2 complète sera livrée à la fin du second semestre 2009.

### **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	24 mois (avec l'objectif de prolonger de 12 mois)
4	NGUYEN	Van Tien	Doctorant	17/11/2008	1 an renouvelable
2	MECHOUCHE	AMMAR	Post-doctorant	18/05/2009	18 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

**Liste des livrables joints au présent rapport (uniquement pour les rapports de fin d'année)**

*Les livrables du projet sont fournis par le coordonnateur.*

Numéro du livrable	Désignation	Forme/Support
3	Mise au point d'outils d'extraction de concepts et de relations	Démonstration + Logiciel accessible sur le site web du projet (rubrique livrables)
4	Enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction et à partir de ressources lexicales	Logiciel accessible sur le site web du projet (rubrique livrables)
5	Intégration et accès aux schémas de bases de données	Rapport sous forme électronique joint
6	Indexation automatique de contenu de documents	Rapport sous forme électronique joint

L'envoi du livrable n°4 « Enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction et à partir de ressources lexicales » correspond à une implémentation partielle du processus d'enrichissement (V1). Il sera suivi de l'envoi d'une version plus complète V2 en décembre 2009.