Compte rendu

Réunion semestrielle n°3 - Projet ANR MDCO – GEONTO

Toulouse le 30 juin 2009 ~ 9h30 – 17h30

Présents : N. Abadie (COGIT), N. Aussenac (IRIT), M.-N. Bessagnet (LIUPPA), M . Gaio (LIUPPA), O. Haemmerlé (IRIT), F. Hamdi (LRI), M. Kamel (IRIT), E. Kergosien (LIUPPA), A. Mechouche (COGIT), S. Mustière (COGIT), N.T. Nguyen (LIUPPA), N. Pernelle (LRI), C. Reynaud (LRI), B. Safar (LRI), F. Saïs (LRI), C. Sallaberry (LIUPPA)

Ordre du jour :
9h30 - 9h35 : Mot d'accueil et rappel du programme de la journée

9h35 - 10h35 : Construction et enrichissement d'ontologies - Lot 1
Sous - lot 1.1 : Exploitation de textes de spécification pour construire une ontologie : dernière évolutions et analyse des parties en langage naturel (IRIT)
Sous - lot 1.2 : Chaîne d'extraction automatique de termes dans des corpus Grand Public (LIUPPA)

10h35 - 11h35 : Appariement d'ontologies - Lot 2
Sous-lot 2.1 : Alignement et fusion d'ontologies (LRI)
Sous-lot 2.2 : Réconciliation d'instances pour l'alignement d'ontologies (LRI)
Sous-lot 2.3 : Premières réflexions sur la comparaison d'ontologies (COGIT)

11h35 - 12h35 : Exploitation des ontologies créées - Lot 3
Sous-lot 3.1 Indexation automatique de contenus de documents patrimoniaux (LIUPPA)
Sous-lot 3.2 : Intégration et accès aux schémas de bases de données : avancement (COGIT)

12h35 - 14h : Déjeuner

14h - 16h : Discussion
16h - 17h30 : Préparation revue septembre : Exposés (plan de la présentation, orateurs), Démos.

Résumé des présentations

1. **Construction automatique d'ontologies à partir de spécifications de bases de données - Sous-lot 1.1 : Mounal Kamel (IRIT)**

1) Rappel de l'approche adoptée : s'appuyer sur la structure du document pour construire une première version d'ontologie (étape 1) et enrichir cette première version par des méthodes d'analyse de textes (étape 2).
Etape 1 : S'appuyer sur l'expert pour donner une sémantique aux balises et aux relations qui les lient. Spécification de règles d'extraction
Etape 2 : analyse des champs définition, écriture de patrons lexico-syntaxiques, annotation du texte, spécification de règles d'extraction.
Pbs : un même terme peut être utilisé à différents endroits du document : même concept ?
Solution : label du concept courant concaténé à tous les labels des concepts pères = nom absolu.
2) Evolutions

Traitement de schémas de spécification normalisés (Norme INSPIRE)
Traitement des énumérations, traitement des définitions (séquence d'annotations – repérage de termes associés à des concepts, d'hyperonymes, de relations partie-de), concepts documentés dans un objectif de traçabilité.
Deux labels associés aux concepts : nom absolu (cf. ci-dessus) + nom relatif
3) Travaux en cours :
Comment rendre l'approche plus générique (moins s'appuyer sur l'expert) ?
Mise en ligne des ontologies construites.
4) Travaux futurs :
Développement d'une interface, étendre l'analyse des définitions, analyser d'autres champs que définitions, enrichir éventuellement l'ontologie construite avec d'autres ressources.

2. **Chaîne d'extraction automatique de termes dans des corpus Grand Public – Sous-lot 1.2 : Eric Kergosien (LIUPPA)**

Un ensemble de termes associés à des entités nommées a été extrait d'un échantillon élargi de textes issus du corpus « récits de voyage ». Cette liste de termes a ensuite été complétée par le thésaurus RAMEAU (généraliste). La confrontation de ces termes avec l'ontologie géographique de l'IGN (TopoCarto-Cogit ou Topo-IRIT) permet d'identifier des termes présents dans le thesaurus mais pas dans l'ontologie à enrichir. Ces termes peuvent venir enrichir l'ontologie s'il existe un lien dans le thesaurus avec des termes déjà présents dans l'ontologie.
La chaîne de traitement actuelle fournit donc un ensemble de termes candidats à l'enrichissement. Il reste à étudier comment automatiser précisément le processus d'enrichissement et si on peut le voir comme un traitement faisant suite à de l'alignement. L'enrichissement peut consister à ajouter des éléments feuilles à l'ontologie enrichie ou des sous-arbres ou branches d'une autre ressource. La solution retenue dans le projet n'est pas encore arrêtée. L'ajout de parties de ressources peut poser des problèmes du fait de la divergence éventuelle de points de vue entre la ressource servant à l'enrichissement et l'ontologie enrichie.

La définition d'un toponyme a été discutée : nom du lieu seul ou expression du type « pic d'Allanz » ou « montagne d'Allanz » ? Dans la présentation, un toponyme correspond au nom du lieu seul.

Perspective : Création d'un Web service, accès à diverses ressources gazetters/SIG locales ou distantes.

3. **Alignement et fusion d'ontologies à travers divers niveaux de richesse – sous-lot 2.1 : F. Hamdi, C. Reynaud, B. Safar (LRI)**

Rappel du planning

Présentation des améliorations apportées sur Taxomap concernant :
- la mesure de similarité
- les techniques

et des tests réalisés à partir de la version améliorée du logiciel consistant à aligner Topo-Cogit et Carto-Cogit.
Une version de TaxoMap est disponible sur le serveur INRIA (svn). Contact : Fayçal Hamdi (Faycal.Hamdi@lri.fr)

Présentation d'un framework d'appariement d'ontologies en cours de conception :

- motivations,
- architecture,
- primitives de raffinement identifiées et exemple de GUI d'aide à la spécification de traitements d'affinement de mappings.

Retour sur l'alignement de Topo-IRIT et de Topo-Cogit : des mappings exploitables pour enrichir Topo-Cogit à l'aide d'éléments liés par des relations du type isA.

4. **Réconciliation de références de données géographiques – Sous-lot 2.2 : N. Pernelle (LRI)**

Travaux ayant débuté en T0+12. Rappel des objectifs à atteindre en T0+18.

Présentation de résultats obtenus par application de la méthode de réconciliation de références conçue au LRI (LN2R) sur des instances de points remarquables du relief issues de BD-Carto et d'instances d'oronymes issues de BD-Topo.
Les conclusions de l'étude sont les suivantes : pas de mécanisme de propagation mis en œuvre du fait de l'absence de relations dans le schéma testé, donc sous-utilisation de l'outil. L2R, méthode logique est intéressante pour pointer les non-réconciliations. N2R, méthode numérique, est applicable mais nécessite des aménagements préalables.

Il avait été convenu au début de l'étude que des tests croisés seraient effectués : données géographiques testées sur l'outil de réconciliation du LRI et données bibliographiques (ayant servi d'expérimentations à l'élaboration de la méthode LN2R du LRI) testées sur les outils du Cogit développées dans la thèse d'Ana-Maria Olteanou. Ces tests n'auront en fait pas lieu sous cette forme. Le Cogit effectuera des tests sur le même échantillon de données que celui sur lequel a travaillé le LRI (instances de points remarquables du relief et d'oronymes).

5. **Comparaison d'ontologies – sous-lot 2.3 : S. Mustière (Cogit)**

Travail qui devait débuter en T0+18, effectué par anticipation. Présentation d'une étude effectuée sur le sujet ayant fait l'objet d'un papier soumis à COSIT'09 (Conference on Spatial Information Theory) non accepté et ayant été publié sous forme de rapport interne de l'Université Paris-Sud.

Rappel du but, des indices recherchés : taux de recouvrement thématique, indices de niveaux de détail, indices de comparaison d'organisation.
Etat de l'art : travaux portant sur l'évaluation d'une ontologie, le partitionnement d'ontologies guidée par l'alignement, la visualisation d'ontologies, l'analogie géographique, l'isomorphisme de graphes.

Ce travail sera poursuivi, côté COGIT, dans le cadre du travail post-doctoral de Ammar Mechouche (contrat de 18 mois).

6. **Indexation automatique de contenus de documents patrimoniaux – sous lot 3.1 : M. Gaio (LIUPPA)**

Utilisation d'entités nommées pour créer les index. Approche retenue : système linguistique à base de règles.
Notion d'entités nommées géographiques (ENG) dérivées (ex : sud de Pau), appelées entités cibles, localisées à partir de sites connus (ex : Pau), les entités site. Les ENG site sont identifiées par recherche de leur géométrie dans des ressources (Gazetters, SIG). Les ENG cibles sont obtenus par raisonnement spatial qualitatif à partir de la géométrie des ENG site et des relations qualitatives.
Dans une telle application, l'identification de la meilleure géométrie ou la géométrie la plus proche des ENG site se fait grâce à une ontologie.
Résultats concernant le typage automatique des termes associés aux ENG : 5% des informations spatiales annotées avec Topo-Cogit, 10% avec Topo-IRIT, 50 % après enrichissement via RAMEAU.
Perspective : Appels directs avec l'ontologie après représentation en OWL.

7. Intégration des bases de données à partir de la formalisation de leurs spécifications – sous-lot 3.2 : N. Abadie (Cogit)

1) Rappel objectif : Besoin d'intégrer des bases de données géographiques multi-thèmes, multi-niveaux à partir de la formalisation de leurs spécifications.
2) Avancement :
Les développements effectués concernent le cas où l'on ne dispose pas de spécifications associées aux schémas des bases de données. La stratégie d'appariement de schémas mise en place consiste, à partir de chacun des schémas à apparier, à générer une ontologie du domaine. Un traducteur automatique, prenant en entrée un schéma de base de données géographique encodé selon la norme ISO 19109, et produisant une ontologie au format OWL en sortie, a été implémenté. Il ne s'agit pas ici d'une simple transformation de format : la sémantique de certaines classes des schémas est enrichie à l'aide de valeurs d'attributs spécifiques faisant directement référence à des labels de concepts géographiques. Les ontologies ainsi produites sont ensuite alignées à l'aide de Taxomap, via la taxonomie TopoCarto-Cogit. Ils permettent le calcul de liens d'appariement entre les schémas des bases de données.
3) Modèle et développements à venir :
La suite du travail va consister, dans un premier temps, à exploiter les résultats d'appariement de schémas déjà obtenus afin de paramétrer l'appariement des données. Il s'agira ensuite de mettre à profit l'appariement des données pour améliorer les résultats d'appariement de schémas dans le cadre d'une boucle de rétroaction. Enfin, nous implémenterons le scénario qui consiste à introduire les connaissances issues des spécifications formelles dans le processus global afin d'affiner le processus d'appariement des schémas.

Remarque : les travaux réalisés par l'IRIT montrent toute la richesse que peut avoir une ontologie. Le Cogit s'interroge sur le fait de représenter à la fois l'ontologie et les spécifications formelles, décrivant les contraintes portant sur les instances contenues dans les bases, au sein d'un unique fichier, dans le langage OWL.

8. Discussion

L'ontologie de référence
Quelle est–elle ?

Il a été réaffirmé qu'un objectif important du projet était la construction d'une ontologie du domaine géographique pouvant servir d'ontologie de référence dans des applications variées,

par exemple pour servir de source externe dans un processus d'alignement d'ontologies géographiques hétérogènes.

L'ontologie de référence est à construire. L'ontologie initiale à considérer est TopoCarto-Cogit. TopoCarto-Cogit sera enrichie à partir de Topo-IRIT, Carto-IRIT, l'ensemble des termes associés à des entités nommées et restructurée.

L'ajout de concepts dans l'ontologie de référence doit être accompagné de la mention de son origine correspondant à la traçabilité documents-ontologie : référence aux spécifications textuelles des bases de données (chemin dans le document XML menant au concept), référence aux textes grand public.

Outils de fusion
Sont-ils utiles ? Si oui, comment le processus de fusion est-il mis en oeuvre ? Si des développements sont nécessaires, qui en a la charge ?

Eléments de réponse :
- Des outils de fusion d'ontologies peuvent être utiles pour rassembler des modèles du domaine réalisés selon un même point de vue. Leur mise en œuvre fait partie du projet.
- Mise en œuvre possible : (1) Protégé fournit un outil d'aide à la fusion d'ontologies : I-Prompt. (2) Le processus de fusion peut résulter de traitements faisant suite à l'alignement. Dans ce cas, nous pouvons étudier comment le mettre en œuvre dans le cadre du framework d'alignements qui sera développé par le LRI.

Post-traitements à effectuer sur l'ontologie Topo-IRIT construite automatiquement à partir des spécifications textuelles de la base BD-Topo

L'ontologie Topo-IRIT nécessite d'être re-travaillée pour pallier aux problèmes suivants :
- présence de concepts dont le label est « Autres »
- présence de deux relations différentes entre deux mêmes concepts
- présence d'un même concept à différents niveaux de la hiérarchie

Ces traitements peuvent-ils être automatisés ?
Une solution simple pourrait être que ce soit l'utilisateur du Cogit qui effectue ce travail en utilisant, par exemple, Protégé, et en étant éventuellement guidé.
La conception d'un outil de vérification qui serait intégré à la plate-forme en cours de conception par le LRI (initialement conçue pour affiner des résultats d'alignement mais dont les fonctionnalités peuvent a priori être élargies) a aussi été discutée. Un tel outil serait utile pour fournir des indications sur les retouches à réaliser sur l'ontologie.
La méthodologie pour réaliser ces post-traitements reste à préciser.

Exploitation des instances pour l'alignement
Lesquelles ?

Pour l'instant, les données ayant fait l'objet de l'étude réalisée dans le cadre du sous-lot 2.2 sont importantes en volume mais ne concernent que deux entités des schémas de BD avec 3 propriétés communes. Le Cogit étudie la possibilité de communiquer un schéma plus complet avec les données associées.

9. Préparation de la présentation du 04/09/09 : revue à T0+18

Se faire préciser la durée exacte de la présentation par l'ANR. ➔ Réponse : 40 minutes.
Tous les partenaires devront être représentés et participeront à l'exposé.
Plan sommaire proposé avec durée ajustée suite aux précisions de l'ANR concernant la durée de la présentation :

I. Contexte et avancement du projet : (7 ') *Coordonnateur du projet*
II. Lot 1 (14 ') – *IRIT (7minutes) + LIUPPA (7 minutes)*
III. Lot 2 (7 ') - *LRI*
IV. Lot 3 (7') - *COGIT*

CL : Conclusion et planning (7') – *Coordonnateur du projet*

Démos à préparer : à enregistrer au préalable. Mouna K. envisage mettre une démo enregistrée sur le site web dès maintenant. A présenter ou y faire référence au cours de l'exposé (si le temps ne le permet pas). Le LIUPPA réalisera une démo enregistrée pour la revue de mi-parcours de septembre.

10. Livrables et rapports d'activité à T0+18

Rappel concernant les rapports d'activité à transmettre au plus vite.
Point sur les livrables à transmettre :

- Discussion autour des modalités de livraison des logiciels. Des précisions sont à demander à l'ANR.

Livrable 3 : Mise au point d'outils d'extraction de concepts et de relations. Resp. IRIT. L'IRIT va rendre le logiciel de construction d'ontologies réalisé accessible sur le site de l'IRIT. Dans ce cas, un lien sur le site web du projet GEONTO (rubriques LIVRABLES) permettra d'y accéder.
Le LIUPPA va indiquer la chaîne de traitement qu'il propose dans la rubrique LIVRABLE du site Web du projet et indiquera comment l'exécuter sachant que l'exécution exige d'utiliser LINGUASTREAM. Une possibilité de téléchargement de LINGUASTREAM sera possible en demandant au laboratoire.
Livrable 4 : Enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction et à partir de ressources lexicales. Resp. IRIT
Le logiciel produit aujourd'hui et développé par le LIUPPA ne permet pas encore d'aller jusqu'à l'enrichissement d'une ontologie. L'implémentation réalisée est partielle. Le LIUPPA propose d'effectuer une livraison des développements tels qu'ils existent aujourd'hui et de les compléter dans les mois qui viennent par la version finale.
Un mail sera envoyé à l'ANR pour demander si ces modalités de livraison de logiciel conviennent ➔ Réponse positive obtenue le 6/7 de l'ANR (Diane Penel).

- Point sur les rapports intermédiaires :

Livrable 5 : Intégration et accès aux schémas de bases de données. Resp. COGIT. Sera envoyé prochainement.
Livrable 6 : Indexation automatique de contenus de documents. Resp. LIUPPA. Sera envoyé prochainement.

11. Publications

Un article avec en co-auteur l'IRIT et le LIUPPA est en cours d'écriture et sera soumis à la conférence JFO (3èmes Journées Francophones sur les Ontologies) 3-4 décembre 2009, Poitiers. Il porte sur la construction d'ontologies à partir de spécifications textuelles (Topo-

IRIT), l'analyse de textes grand public et l'extraction de termes associés à des toponymes, l'utilisation de ces termes pour enrichir TOPO-IRIT.

L'envoi d'un papier à la conférence nationale SAGEO 2009 (Spatial Analysis and GEOmatics) qui se tiendra à Paris en novembre 2009 est discuté. L'objectif est de présenter le projet GEONTO en ciblant sur les ontologies géographiques, plutôt que sur les schémas de bases de données géographiques. Sébastien Mustière fait une proposition de plan pour la fin de semaine. Chaque partenaire étudie la façon dont il peut contribuer. Liste des co-auteurs = liste des personnes participant à l'écriture du papier. Deadline : 15 juillet.

Rappel : chaque partenaire doit penser à transmettre les références des articles publiés au coordonnateur, accompagnées du papier (si public), afin qu'elles figurent sur le site Web du projet.