Compte rendu rédigé par C. Reynaud

Réunion Semestre 2 du projet ANR MDCO GEONTO

Orsay le 23 janvier 2009 ~ 10h - 16h30

Présents : N. Abadie (COGIT), N. Aussenac (IRIT), N. Bénnacer (Supelec), M. Gaio (LIUPPA), O. Haemmerlé (IRIT), F. Hamdi (LRI), M. Kamel (IRIT), P. Loustau (LIUPPA), Y. Mrabet (LRI), S. Mustière (COGIT), D. Penel (CEA ANR-CI), N. Pernelle (LRI), C. Reynaud (LRI), B. Safar (LRI), F. Saïs (LRI), C. Sallaberry (LIUPPA)

Sommaire des présentations :

1. Introduction à la journée – C. Reynaud (LRI)
2. Le fonctionnement des projets ANR - Diane Penel (ANR)
3. Construction et enrichissement d'ontologies – Lot 1 (Resp. N. Aussenac)
4. Construction automatique d'ontologies à partir de spécifications de bases de données (Sous- Lot 1.1) : Mouna Kamel (IRIT)
5. Vers une ontologie géographique des récits de voyage (Sous-Lot 1.2) : Mauro Gaio / Pierre Loustau (LIUPPA)
6. Appariement d'ontologies hétérogènes (Sous-Lot 2.1) : Brigitte Safar (LRI)
7. Lot 3 : Bilan COGIT : S. Mustière (COGIT)
8. Sous- Lot 3.2 Intégration des bases de données à partir de la formalisation de leurs spécifications : Nathalie Abadie (COGIT)
9. Discussion

1. Introduction - C. Reynaud (LRI)

Rappel des objectifs généraux du projet et de l'avancement des tâches à T0+12. Présentation des objectifs par lot tels qu'ils ont été redéfinis en T0+6.
Eléments de gestion de projet : rappel des tâches à réaliser (rapports d'activités semestriels, livrables) pour le 31 janvier 2009 (fin Semestre 2), liste des publications liées au projet, point sur les CDD.

2. Le déroulement des projets ANR - D. Penel (CEA ANR-CI)

1. Contexte : ANR et Unité support. Site de la délégation : http://www-anr-ci.cea.fr. Interlocuteurs ANR (François Jacquenet, interlocuteur ANR Masse de Données) et Unité Support (interlocuteur privilégié).
2. Rôle de l'unité support
3. Droits et devoirs des projets. Rôle important des actes attributifs qui spécifient les droits et devoirs de chaque partie. Rappel des éléments à fournir périodiquement. Mentions à porter sur les publications ou présentations (texte avec les références du projet (**ANR-07-MDCO-005**), logo sur le site http://www-anr-ci.cea.fr. Communication prévue sur les projets ANR STIC à l'automne 2009 (poster ou exposé) et au colloque de fin d'édition de l'appel. Réunions de mi-parcours (en septembre 2009). S'attacher à montrer l'aspect collaboratif et les résultats de la collaboration. Bien mettre à jour le site web du projet. Réunion de fin de parcours.
4. Conseils pour la bonne marche des projets. *En cas de changements dans l'utilisation des fonds, il est nécessaire de contacter l'US* (mail à D. Penel). Si plus de 30% du

financement est concerné, l'ANR sera contactée. En cas de retard, une prorogation peut être accordée. *Faire la demande 6 ou 12 mois avant la fin initialement prévue*. Envoi d'une **lettre signée de l'ensemble des partenaires** car c'est l'ensemble du projet qui est prolongé.

5. Fonctionnement des versements. *Si dépenses inférieures au cumul des versements reçus, engagement à rembourser le trop-perçu*.
6. Précisions sur l'accord de consortium, la labellisation « pôle de compétitivité » (des fonds qui peuvent aussi servir à l'organisation d'un workshop).
7. Clôture du projet. Bilan scientifique + Réunion de clôture + dossier de clôture (Modèles sur le site).

3. Construction et enrichissement d'ontologies – Lot 1 - N. Aussenac (IRIT)

1. Rappel du contenu défini en T0+6 et des enjeux
   - 1.1 Mise au point d'outils d'extraction de concepts et de relations à partir de textes : 2 corpus utilisés : les récits de voyage et les spécifications de BD-Topo. Automatisation maximale. Construction de premières ontologies (Topo-IRIT et Itinéraire)
   - 1.2 Enrichissement d'une ontologie existante à partir de textes et de ressources lexicales : effectué sur Itinéraire.
   - 1.3 Restructuration d'ontologie : Etape finale - utilisation de techniques d'alignement sur Topo-IRIT et Itinéraire (début : T0+24).
2. Présentations de l'activité du lot 1 sur l'année :
   Développement de chaînes automatiques de traitement de textes (collaborations fortes IRIT et LIUPPA), construction automatique d'ontologie (IRIT), exploitation de ressources lexicales (LIUPPA). Une réunion de travail.
3. Perspectives et moyens :
   Perspectives scientifiques : application aux autres documents de spécification, systématiser l'exploitation des documents grands publics, définir le processus d'enrichissement et le lien avec l'alignement.
   Perspectives techniques : intégration des outils LinguaStream et Gate, choix d'une plate-forme
   Moyens : recrutement conjoint d'un post-doc LIUPPA-IRIT.

4. Construction automatique d'ontologies à partir de spécifications de bases de données - Sous- Lot 1.1 - M. Kamel (IRIT)

1. Analyse des documents de spécification associés à BD-Topo :
   Différences avec les textes habituels : peu de textes rédigés.
   Analyse de la structure : repérage de concepts présents dans certains champs – repérage de relations hiérarchiques ou du domaine ou de propriétés. Analyse du texte lui-même : texte contenu dans les champs définitions et modélisation géométrique, par exemple.
2. Méthode de construction automatique d'ontologies à partir de spécifications : Combinaison d'une approche exploitant la structure d'un document XML (tags) et d'une approche exploitant le texte. Définition de règles d'extraction de concepts et de relations basées sur la structure du document, définition de patrons lexico-syntaxiques pour déceler des concepts et des relations dans les parties textes.
   Illustration sur des exemples tirés des spécifications associées à BD-Topo.
3. Mise en œuvre de la méthode à l'aide de GATE (General Architecture for Text Engineering).
   Règles java pour l'exploitation de la structure, règles JAPE correspondant aux patrons de méronymie et de définitions de propriétés. Chaîne de traitement présentée.

4. Evaluation :
   Comparaison avec l'approche de F. Laurens (Cogit) du point de vue méthodologique et selon différents critères : nombre de concepts obtenus, profondeur, nature des éléments présents dans l'ontologie.
   Apport de l'approche présentée : extraction de propriétés, relations de méronymie, autres relations conceptuelles, méthode non supervisée. A permis de détecter des incohérences dans les spécifications.
5. Evolution et perspectives :
   - Vers un traitement automatique avec trace entre ontologie et document d'origine. La validation de l'expert ne se fera qu'après l'alignement des différentes ontologies.
   - Etre capable d'analyser tous les documents de spécifications du projet Geonto et étudier la portabilité de l'analyse de la structure. Expérimentations à faire à partir de documents munis d'un schéma XML tels qu'il en existe dans le domaine de la géographie.
   - Tirer parti de la forme matérielle du texte.

5. Vers une ontologie géographique des récits de voyage – Sous-lot 1.2 - M. Gaio et P. Loustau (LIUPPA)

Redéfinition de l'objectif du sous-lot 1.2 : (1) création d'une ontologie d'un domaine particulier à partir de textes, d'outils d'extraction et selon un objectif applicatif donné, (2) faire collaborer cette ontologie avec l'ontologie TopoCarto-Cogit. Les textes choisis sont les textes des récits de voyage. L'usage est de type pédagogique.

La nécessité de construction d'une telle ontologie provient du fait que les documents exploités ne sont pas structurés, qu'il n'existe pas de schéma XML. L'application de techniques analogues à celles définies par l'IRIT n'est pas possible.

L'ontologie créée pour des besoins pédagogiques visés comporte les concepts d'itinéraire, de trajectoire, d'activité, d'observation, de jugement, … et des relations de comparaison, d'opposition, de séquentialité, …Un modèle conceptuel MADS (adapté aux applications comportant des données spatio-temporelles) a tout d'abord été construit manuellement. Il a ensuite été traduit en OWL.

L'ontologie ainsi construite a été projetée sur les textes (récits de voyage) de façon à identifier les endroits où ces éléments apparaissent. Ceci a permis d'améliorer les patrons syntaxico-sémantiques utilisés dans la chaîne de traitement LinguaStream, mais également de peupler l'ontologie.

L'intérêt de confronter cette ontologie Itinéraire à celle de l'IGN (TopoCarto-Cogit) sera de l'enrichir avec un point de vue « itinéraire » (et inversement). L'utilisation de Wikipedia pourrait permettre de faciliter le lien entre ces deux ontologies.

6. Appariement d'ontologies hétérogènes (Sous-Lot 2.1) : Brigitte Safar (LRI)

1. Présentation des tâches réalisées sur 2008.
2. Rappel bref des techniques mises en oeuvre dans TaxoMap.
3. Evaluation :
   Méthodologie et résultats.
4. Améliorations à apporter :

Traitement d'alignement à compléter par une analyse de la structure différente selon les cas - Accepter les mappings 1:n - Utiliser le partitionnement d'ontologies pour définir le contexte d'interprétation de concepts - Affiner la mesure de similarité utilisée dans l'outil - S'appuyer sur des alignements « sûrs » pour en découvrir d'autres - Utilisation de connaissances autres (ressources externes).

5. Problème à l'étude :
   Faire évoluer le contrôle des tâches dans TaxoMap de façon à rendre l'outil plus flexible et permettre sa réutilisation pour d'autres tâches (fusion, enrichissement, restructuration).

Questions/Remarques

- Les ontologies du lot 1 comportent des propriétés. Comment faire pour les traiter ? R : s'appuyer sur les domaines et les co-domaines.
- Les fausses relations *is-A* de TopoCarto-Cogit pourraient être trouvées par comparaison avec les relations du domaine trouvées dans les ontologies produites dans le lot 1.

7. Avancement GEONTO – Lot 3 - S. Mustière (COGIT)

1. Point sur les activités de 2008 concernant le lot 1, le lot 2 et le lot 3.
2. Introduction de la thèse « Intégration de bases de données à partir de la formalisation de leurs spécifications ».
   Objectif premier = propagation des mises à jour. Ensuite : détection d'incohérences, analyses multi-niveaux exploitant des données de plusieurs bases, partage de données, requêtes et navigation plus simples (ne nécessitera pas d'avoir une très bonne connaissance des bases).
   Trois cas seront étudiés dans la thèse : (1) des spécifications existent, (2) une seule des bases a des spécifications, (3) aucune des bases n'a de spécifications.

Rq : Un des intérêts de travailler à partir d'ontologie est que la richesse des spécifications peut être exploitée dans une ontologie. Elle ne pourrait pas forcément l'être dans un schéma.

8. Intégration des bases de données à partir de la formalisation de leurs spécifications – Lot 3.2 - N. Abadie (COGIT)

1. Contexte
   Besoin d'intégrer des bases de données géographiques multi thèmes, multi-niveaux et transfrontalières, avec ou sans spécifications.
2. Sujet : Intégration des bases de données à partir de la formalisation de leurs spécifications
3. Cas d'application : (1) Intégration de BDG avec spécifications, (2) Intégration de BDG avec et sans spécifications, (3) Intégration de BDG sans spécifications.
4. Approches envisagées : Besoin de construire et aligner les ontologies, besoin de décrire les liens schéma-ontologie, besoin d'apparier les schémas et les données.
5. Approche testée pour le cas d'application n°3 : appariement de schémas sans spécifications. Construction d'ontologies – Alignement (techniques similaires à celles de TaxoMap + utilisation de la taxonomie de background produite à partir du texte des spécifications) – Exploitation liens entre ontologies et liens schémas-ontologies pour apparier les schémas.
   Résultats : Intérêt d'utiliser les concepts cachés – Intérêt de la taxonomie de background.
6. Conclusion : développer les autres scénarii, définir un modèle de correspondance plus expressif, intégrer les différentes approches d'appariement des données, définir une stratégie d'intégration globale exploitant l'appariement de schémas et de données.

9. Discussion

Rappel des besoins de l'IGN : techniques d'extraction, d'alignement + une ontologie de référence riche.

Sous-lot 1.1 :
1) Il serait bon de tester la généricité de l'approche. Cette généricité proviendra de l'adoption du Schema XML en cours de standardisation dans le domaine. Cela nécessite de prendre en compte les travaux faits au niveau européen dans le domaine. *A communiquer par le Cogit.*
2) L'IRIT poursuit son travail sur l'extraction en définissant de nouvelles règles (exploitation de la typographie). La chaîne de traitement mise au point sera appliquée à au moins 2 schémas. Les résultats obtenus devront ensuite être alignés.
3) L'idée de fusionner TopoCarto-Cogit avec d'une part Topo-IRIT puis avec Carto-IRIT en ajoutant la traçabilité documents-ontologie est intéressante, de même que l'ajout de relations du domaine et de propriétés (travail d'enrichissement à inclure dans le sous-lot 1.2)

Sous-lot 1.2 :
1) L'enrichissement de TopoCarto-Cogit avec des données sur les itinéraires n'est pas très pertinent. Il serait préférable de faire coopérer différents points de vue en établissant des liens de mise en correspondance.
2) Il faut étudier l'ajout d'autres informations et se poser donc la question des sources les plus pertinentes à utiliser pour cela.
3) La construction d'un thesaurus contenant des termes tels que pic, lac, …à côté de toponymes qui sont des noms de lieux doit être réalisée. Ce thésaurus devra être aligné avec TopoCarto-Cogit de façon à déceler les termes absents dans TopoCarto-Cogit, et l'enrichir.
4) Le travail sur la chaîne de traitement LinguaStream doit être poursuivi.
5) L'ontologie Itinéraire doit être finalisée.

Lot 2 :
1) L'extension d'AlignViz, outil de visualisation d'ontologies et de mappings, pour effectuer des modifications dans l'ontologie cible sur la base des alignements produits, serait souhaitable. *Voir si cela n'est pas déjà fait dans Prompt à Stanford.*
2) Récapitulatif des fusions à réaliser :
   - Topo-IRIT avec TopoCarto-Cogit pour fusion/enrichissement de TopoCarto-Cogit → Fusion1
   - Carto-IRIT avec TopoCarto-Cogit pour fusion/enrichissement de TopoCarto-Cogit → Fusion 2.
   - Fusion 1 avec Fusion2
   - Itineraire avec TopoCarto-Cogit pour un alignement sans fusion en vue d'étudier les différences.
   - In fine, si cela est possible, TopoCarto-Cogit enrichi avec les ontologies construites par l'IRIT avec des ontologies externes (Towntology, GEIA, …) afin d'étudier les différences.

Communication

1) Faire un papier avec tous les partenaires sur l'ensemble du projet dans les 6 mois
2) Organiser un workshop en 2010. Peut se faire avec le versement reçu du pôle de compétitivité Cap-Digital.

Conférences à venir

COSIT 2009 : Conference on Spatial Information Theory. 21-25 sept 2009, Aber-Wrac'h (soumission : 2 mars)

K-CAP 2009 : The fifth International Conference on Knowledge Capture, 1-4 september 2009, California, USA (submission: 15/04)

CIAO 2009, 5th international Workshop on Cooperation & Interoperability – Architecture & Ontology, 8-9 juin 2009, The Netherlands (in conjunction with CAiSE) (submission: 22/02)

DEXA 2009, 20th Conference on Database and Expert Systems Applications, August 31 – Sept. 4, Austria (submission: 08/03)

IC 2009 : Journées Francophones d'Ingénierie des Connaissances, 25-29 mai 2009, Hammamet (Tunisie). Date de soumission dépassée.

GDR MAGIS : Pau ou Paris en 2009, en octobre ou novembre. En 2010, il auralieu à Toulouse.

TIA 2009 : Toulouse. Novembre. Organisatrice : N. Aussenac (IRIT).

Retards/changement d'affectation des fonds :

Les retards dans l'embauche de CDD justifient le décalage dans les travaux réalisés.

L'ANR a rappelé que le changement d'affectation des fonds doit être signalé. **A chaque partenaire de le faire individuellement en envoyant un mail à Diane Penel.**

Accord de consortium :

A mettre à l'ordre du jour de la prochaine réunion.

Ne semble pas vraiment nécessaire actuellement.

Site web :

1) mettre à jour les publications et permettre d'accéder aux papiers.
2) Ajouter les listes de ressources géographiques
3) Ajouter logo ANR + Cap Digital

Divers :

Demande de la part du LIUPPA de bibliographie sur les ontologies :

R : L' HDR de Jean Charlet (http://www-test.biomath.jussieu.fr/~jc/) soutenue en 2002 et intitulée « L'ingénierie des connaissances – développements, résultats et perspectives pour la gestion des connaissances médicales » contient beaucoup d'informations sur le sujet. Les différents membres du projet sont par ailleurs invités à envoyer au LIUPPA les références qu'ils jugent intéressantes sur le domaine.

Prochaine réunion plénière :

1) Date : en juin 2009
2) Ordre du jour : réunion de bilan + préparation revue de septembre.
3) Lieu : Toulouse ?