

Compte rendu

Réunion semestrielle n°5 – Projet ANR MDCO – GEONTO

Orsay le 21 Juin 2010 – 9h30 – 17h30

Présents : Nathalie Abadie (COGIT), Nathalie Aussenac (IRIT), Nacéra Bennacer (Supelec), Marie-Noëlle Bessagnet (LIUPPA), Mauro Gaio (LIUPPA), Fayçal Hamdi (LRI), Marion Laignelet (IRIT), Sébastien Mustière (COGIT), Yassine Mrabet (LRI), Nathalie Pernelle (LRI), Emeric Prouteau (COGIT), Chantal Reynaud (LRI), Fatiha Saïs (LRI), Brigitte Safar (LRI), Christian Sallaberry (LIUPPA)

Ordre du jour :

9h15 : Accueil

9h30 : Lot 1 – Construction et enrichissement d'ontologies (Nathalie Aussenac)
Présentations de l'IRIT et du LIUPPA

10h30 : Lot 2 - Appariement d'ontologies (Chantal Reynaud - LRI)

Sous-lot 2.1 : TaxoMap Framework utilisé pour l'enrichissement – (Brigitte Safar - Fayçal Hamdi – LRI)

Sous-lot 2.2 : Réconciliation combinée de schémas et de données (Nathalie Pernelle, Fatiha Saïs, Yassine Mrabet – LRI)

Sous-lot 2.3 : Comparaison d'ontologies (Ammar Mechouche – COGIT)

11h30 : Lot 3 - Exploitation des ontologies créées (Sébastien Mustière – COGIT)

Sous-lot 3.1 : LIUPPA

Sous-lot 3.2 : Une interface de consultation des spécifications (Emeric Prouteau – COGIT)

12h30 – 14h : Déjeuner

14h - 17h :

Discussion

+ Questions diverses :

Organisation d'un atelier à Sageo 2010

Soumission à la revue de Géomatique

Suite du projet

Résumé des présentations

1. Avancement du lot 1 – Construction et enrichissement d'ontologies : N. Aussenac (IRIT)

1) Construction d'ontologies à partir de spécifications de bases de données (IRIT : N. Aussenac-Gilles, Mouna Kamel, Marion Laignelet)

Présentation de la nouvelle approche adoptée consistant à séparer les concepts du monde réel de ceux du monde de la base de données dont les vocabulaires ne se recouvrent pas totalement. Conséquences : Utilisation de DataRange nécessaire. OWL Lite n'est plus suffisant. Programmation faite en java.

Résultats de l'extraction :

- A partir des spécifications de BDTopo : 1262 concepts, 8 relations et 102 propriétés
- A partir des spécifications de BDCarto : 884 concepts, 0 relation, 159 propriétés.

Problèmes :

- Présence de concepts dont le label est une valeur numérique
- Pré-traitement sur le Fichier XML nécessaire : des éléments d'énumérations traitées différemment en XML (dans BD Carto)
- La valeur dans le DataRange peut être un paragraphe (dans BDCarto)
- Attributs non typés non traités (nombreux dans BD Carto)

Comparaison par rapport à l'approche précédente :

L'approche précédente offrait un certain degré de généralité, les définitions étaient associées aux concepts, les ontologies produites étaient exploitables MAIS les concepts issus des attributs étaient rattachés sous Top.

Nouvelle approche : Problème des rattachements résolu avec DataRange, approche utilisable pour BD Topo mais pas BD Carto. Perte en généralité.

En parallèle, Marion Laignelet a travaillé sur la recherche de relations et de concepts dans les parties définitions des spécifications. Recherche de relations entre des concepts identifiés par Mouna + nouveaux concepts. Utilise LinguaStream.

Projection de lexiques créés à partir des documents de spécification (concepts bruts + concepts du monde carto et du monde réel + autres) – Repérage des SN simples, coordonnés, suivis d'un SP. Reste à traiter les adjectifs + les coordinations de SN + les textes entre parenthèses. – 3 types de relations repérées : hyperonymie, méronymie, artefact (pourrait permettre de passer du monde de la carte au monde réel). Beaucoup de relations d'hyperonymie trouvées, peu d'autres relations. A faire : affiner + évaluer les traitements implémentés.

CL : ce qui est produit : au niveau de la langue. Pas intégrable aujourd'hui dans l'ontologie produite par Mouna → des pistes pour construire une ontologie. Approche qui permet de traiter tout type de texte.

Fin du post-doc possible au 1^{er} septembre.

2) Recherche des qualifiants dans le corpus de récits de voyage – Amélioration de la chaîne de traitement (LIUPPA : M.-N. Bessagnet, M. Gaio, E. Kergosien, T. Nguyen, C. Sallaberry)

Problématique : définir les limites d'une entité nommée (EN) de type lieu (ex : Pau) et catégoriser les qualifiants (ex : ville) qui y sont associés.

Définition de ce qu'est une EN. Présentation des structures complexes possibles et de la gestion des indications (ex : au sud de) avec des noms toponymiques complexes.

3 cas de figure de qualifiant acquérant le statut de concept :

- La chaîne de traitement regarde dans une ontologie l'existence du concept. On récupère le graphe dans lequel il se trouve.
- On utilise un thesaurus mais on ne sait si ce qu'on récupère est géographique ou non.
- On ne trouve pas le concept ni dans une ontologie, ni dans un thesaurus.

Pb : comment savoir si un qualificatif est réellement un candidat à l'enrichissement ?
Hypothèse : (V,P ?,E) (Verbe, Préposition facultative, Entité spatiale) = connotation plus géographique car les documents décrivent des itinéraires et les EN les plus impliquées dans l'espace géographique = celles associées à des verbes de déplacement. → filtrage plus important et plus pertinent des termes de Rameau. Limites : certaines constructions posent encore problème.

2. Avancement du lot 2 – Alignement d'ontologies (C. Reynaud)

1) Avancement du sous-lot 2.1 et présentation du module d'enrichissement de Taxomap Framework: (LRI : B. Safar, F. Hamdi)

- Avancement du sous-lot 2.1 :

Implémentation de TaxoMap Framework pour le raffinement de mappings – Intégration de divers outils à cet environnement – Implémentation des 1ers patterns d'enrichissement.

- Enrichissement

Présentation du processus d'enrichissement avec un enrichissement venant d'ontologies construites par l'IRIT ou d'ontologies externes.

Problèmes à traiter : Evaluer l'intérêt d'ontologies - extraire les parties intéressantes – Utiliser les parties intéressantes pour l'enrichissement.

Rappel de la mesure « Semantic Cotopy » qui, associée à la technique de partitionnement permet de comparer 2 ontologies.

Présentation d'une méthodologie pour l'enrichissement partant du partitionnement, passant par l'alignement et se terminant par l'enrichissement.

- Outils intégrés dans TaxoMap Framework : Mesure de comparaison de Maedche-Staab, AlignViz, Analyseur d'ontologies, Partitionneur.

- Présentation de l'utilisation des mappings pour l'enrichissement et des patrons implémentés. Besoin d'outils pour traiter les mappings restants.

2) Avancement du sous-lot 2.2 : Réconciliation combinée d'ontologies et de données (LRI : N. Pernelle et F. Saï's)

Rappel des objectifs : améliorer la réconciliation d'ontologies en utilisant les instances + réconcilier des instances décrites dans des ontologies distinctes.

Possibilités de réconciliation : ref/ref, élément/élément, ref/élément, élément/ref

Hypothèse : on suppose qu'on dispose de mappings corrects sur les relations et les propriétés.

Pour simplifier : 2 systèmes d'équations résolus alternativement : calcul des similarités d'éléments en tenant compte des similarités au niveau données mais sans propagation des scores sur les références + calcul des similarités entre références en tenant compte des similarités au niveau ontologie mais sans propagation des scores sur les éléments de l'ontologie.

Intérêts : calculs simplifiés, interventions possibles à n'importe quel moment de l'expert. L'espace de recherche est ainsi plus restreint. Définition de différentes techniques de similarité selon les éléments sur lesquels ils s'appliquent – Moyenne pondérée pour les concepts. Pour les tuples : cf. LN2R + similarité conceptuelle.

Expérimentation sur els données suivantes : KIM, DBpedia, YAGO, IGN (TopoCarto)
Alignement des ontologies avec TaxoMap. Synthèse des instances communes pour les concepts alignés qui vont permettre d'amorcer l'approche (application du double système d'équations).

3) Avancement du sous-lot 2.3 : Comparaison d'ontologies « distance sémantique entre ontologies (COGIT : Ammar Mechouche)

3 indices de comparaison recherchés : recouvrement thématique, indice de niveau de détail et de comparaison des structures.

Méthodes proposées :

Pour le recouvrement thématique : calcul des sommets importants et calcul d'une distance entre ontologies analogue au calcul de distance entre documents (modèle vectoriel adapté).

Pour la comparaison des structures : adaptation de la distance d'édition entre arbres ordonnés

Pour le niveau de détail : distinction niveau horizontal et vertical.

Pb : valider els résultats obtenus. On peut imaginer beaucoup de mesures. Les résultats ont-ils un sens ?

3. Avancement du lot 3 : S. Mustière (COGIT)

4) Avancement du sous-lot 3.1 – Chaîne de traitement PIV (Pyrénées – Itinéraires – Virtuels)
(LIUPPA: C. Sallaberry)

Chaîne de traitement en 2 parties : marquage et représentation symbolique des EN spatiales (typage et détection des relations) + représentation numérique (nécessité de calculer des géométries pour l'indexation) et indexation des EN spatiales.

Etape 1 : On vérifie que les qualificants sont dans l'ontologie de référence et on récupère le sous arbre correspondant au qualificant. (père + frère + fils éventuels)

Etape 2 : on parcourt l'ontologie de la BD en donnant le graphe. Objectif : récupérer la valeur de l'attribut nature correspondant. On consulte ensuite la BD correspondante et on récupère la géométrie.

Aujourd'hui, il manque des informations sur l'ontologie construite par l'IRIT pour pouvoir l'exploiter.

Expérimentations à réaliser pour voir si on peut étiqueter davantage avec l'ontologie Geonto, évaluer la qualité des géométries, calcul des précisions et rappels des requêtes traitées par PIV par comparaison à PIV + ontologie Geonto.

5) Avancement du sous-lot 3.2 – découverte de contenus des bases de données IGN (COGIT: E. Prouteau). Exposé suivi d'une démo.

Objectif : fournir une interface Web pour découvrir et visualiser des données des BD de l'IGN.

Interrogation avec le vocabulaire de l'ontologie du monde réel de l'ontologie de la BD. Les expérimentations ont été faites sur des ontologies construites par l'IGN, réduites à la partie hydrographie.

7. Discussion

Place des définitions : dans l'ontologie du monde réel ou de la base ?

Après plusieurs échanges, il a été décidé de mettre les définitions dans les 2 ontologies.

Etude de Topo-IRIT et de Carto-IRIT :

- Un même objet avec la même URI peut avoir la même description. Ex : zone d'habitat. La classe est re-crée à chaque fois qu'elle est complétée. *Vérifier la syntaxe OWL pour ajouter une propriété à une classe.*
- Les labels ne doivent pas comporter de blancs → *seront éliminés.*
- Concept Autres, sans objets, Inconnu. Côté monde réel : *éliminer le concept Autres et raccrocher ses fils au père de Autres. Idem pour Sans objets et Inconnu.*
- Labels numériques : à *remplacer* par la valeur de définition pour le monde réel. Côté base, laisser en l'état.
- Des concepts apparaissent à différents niveaux. Ex : A fils de A. Si c'est pour enrichir, *on enlève le concept le plus spécifique.*
- Science fils de Type d'établissement. *Devrait être renommé.*
- Plate-forme multi-sport : seul sport est gardé comme concept → *bug à modifier*

Remarques

BD Carto est de moins bonne qualité que BD Topo.

Certaines modifications jugées nécessaires peuvent être faites à la main.

Des choix de modélisation sont faits par l'IRIT. Dans l'idéal, le Cogit devrait pouvoir revenir dessus. Peu de chances que ce soit fait par manque de temps.

Calendrier pour l'obtention des ontologies produites par l'IRIT :

Rapidement, l'IRIT fournir une V1 à Pau et au LRI = ontologie de Mouna nettoyée et limitée à BDTopo.

En septembre, V2 = V1 enrichie du travail de Marion.

On laisse de côté BD Carto pour l'instant. On y reviendra après si le temps le permet.

Processus d'enrichissement

Aujourd'hui le LIUPPA délivre un ensemble de termes candidats à l'enrichissement en SKOS mais ne va pas jusqu'à enrichir l'ontologie de l'IRIT.

Solution à étudier : Les sous-graphes sont considérés comme des mini ontologies qu'on aligne avant d'enrichir. Le LIUPPA fournit ces mini-ontologies au LRI et le LRI enrichit à l'aide de patrons.

8. Questions diverses

- Organisation d'un atelier à SAGEO 2010.

Mercredi 19 novembre à Toulouse. Demi-journée (après-midi) pas en parallèle avec la conférence.

- Membres à associer au CP : Thérèse Libourel, Christelle vangenot, Jérôme gensel, Alain Bouju, Raphaël Troncy ;
- 1 conférencier invité : demander à Oscar Corcho de Madrid qui travaille avec l'équivalent de l'IGN en Espagne.
- Laisser le temps (3/4 h ou 1h) pour des démos/posters à la fin de l'après-midi. Posters seront intégrés dans les actes.

- Point sur le financement permis par CapDigital d'intervenants évoqués mais pas tranché.
- Dates importantes discutées : versions finales pour le 29/10, notification des auteurs au 15/10 et date limite de soumission 1^{er} Octobre.

- **Suite de GeOnto :**

Associer Thérèse Libourel.

Partenaire industriel souvent nécessaire : plutôt une grosse structure. Thalès ?? EADS ??? La Sagem avait contacté Nathalie Abadie sur le problème de la construction d'ontologies. Voir du côté INRIA. Calendrier : Se donner jusque mi-septembre. On fera le point à ce moment là Premier draft de projet pour la mi-octobre.