

Thèse COGIT :

Cette thèse est entièrement financée par l'IGN et concerne une ingénieur IGN permanente.

Co-encadrants : Anne Ruas et Sébastien Mustière

Descriptif du sujet :

Le sujet de la thèse est "l'intégration des bases de données à partir de la formalisation de leur spécifications". Il correspond au sous-lot 3.2. La thèse supposera l'existence d'ontologies du domaine riches et alignées, comme celles issues des lots 1 et 2. Dans un premier temps, elle s'attachera à étudier comment exploiter ces ontologies pour instancier grâce à des méthodes de traitement automatique du langage un modèle formel de description des spécifications. Si le lot 2 vise à déduire une ontologie à partir de textes tels que les spécifications (par exemple, extraire les concepts de "route", "rivière"...), la thèse vise à formaliser le lien entre la base de données et l'ontologie (par exemple, représenter le fait que toutes les routes de plus de 100 mètres de long sont représentées dans la base). Dans un deuxième temps, il s'agira d'étudier comment ces descriptions formelles de deux bases de données peuvent être exploitées pour les intégrer.

Thèse LRI : « Enrichir, restructurer et fusionner des ontologies par alignement »

Co-encadrement : Brigitte Safar, Chantal Reynaud

Descriptif du sujet :

Le thème central de la thèse porte sur l'alignement d'ontologies. L'équipe IASI du LRI dispose d'un environnement d'alignement, Taxomap, développé à l'origine dans le cadre du projet e.Dot (RNTL), pour interroger de façon unifiée des sources hétérogènes. La thèse consistera à partir de cet existant pour le confronter aux besoins du projet GEONTO. Il s'agira d'adapter et d'étendre le système Taxomap pour prendre en compte les spécificités des ontologies et des besoins reposant sur l'alignement dans le domaine de la topographie.

Dans un premier temps, il s'agira d'aligner des taxonomies réelles fournies par le COGIT (sous-lot 2.1), et également de contribuer à leur enrichissement (sous-lot 1.2) par des méthodes d'alignement (alignement de l'ontologie à enrichir avec d'autres ontologies ou ressources externes).

Dans un second temps, il s'agira de travailler sur les ontologies construites automatiquement à partir de textes fournies par le sous-lot 1.1 du projet dans l'objectif de les aligner (sous-lot 2.1) mais aussi les restructurer (sous-lot 1.3). Le processus d'alignement mis en œuvre devra permettre la fusion d'ontologies, si nécessaire (sous-lot 2.1). L'accent sera mis sur l'alignement de structures très hétérogènes (cf. description du sous-lot 2.1), nécessitant de ce fait la mise en œuvre de techniques d'alignement innovantes et originales. Pour l'aspect restructuration, l'originalité du travail consistera à étudier et exploiter les atouts d'une combinaison de techniques d'alignement et de techniques de construction automatique d'ontologie à partir de textes.

L'ensemble des résultats sera validé dans le domaine géographique mais les techniques et méthodes mises au point seront génériques.

Thèse LIUPPA :

Co-encadrement : Gaio Mauro, Sallaberry Christian

Descriptif du sujet :

La première hypothèse sous-tendant le sujet de cette thèse considère que dans des ressources d'informations, hétérogènes du point de vue formes d'expressions, mais homogènes du point de vue référence à un territoire, « la régularité des références faites à des objets de ce territoire doit rendre possible la réalisation de chaînes d'analyses et d'interprétations sémantiques ciblées, raisonnablement générales et donc portables à coût contrôlé ». Cette hypothèse a déjà acquis une première assise grâce aux travaux initiés dans le cadre des projets GeoSem (CNRS) et PIV (CAPP/MIDR) qui ont abouti à la réalisation d'une chaîne automatique « minimale » d'extraction d'information et d'indexation, d'entités géographiques et de certaines expressions spatiales associées, à partir de documents textuels.

Il s'agit désormais d'étendre cette assise pour intégrer aux index produits des concepts plus étendus grâce à l'intégration dans la chaîne d'extraction et d'indexation des expressions géographiques d'une collection de ressources documentaires, de ou des ontologies produites dans ce projet.

L'approche que nous avons envisagée reprend globalement les principes des systèmes d'extraction et

d'indexation traditionnels : un traitement « off-line » permet d'analyser la collection de documents et de produire des index. Puis, pour chaque besoin d'information formulé par un utilisateur, un moteur se charge de chercher dans les index les entrées témoignant des plus grands appariements.

Notre approche présentera néanmoins un certain nombre de spécificités quant aux méthodes qui seront utilisées pour l'indexation. Notamment elle ne s'appuiera pas uniquement sur un ensemble de mots clés mais sur les représentations conceptuelles (déduites grâce à l'ontologie) et géométriques (obtenues grâce à l'association entre termes de l'ontologie et champ d'une base de donnée géographique) des objets identifiés et de leurs relations. Il en résultera un pouvoir expressif beaucoup plus riche, pierre angulaire pour permettre une interrogation polyvalente pour tout public du simple utilisateur occasionnel à l'expert.