Fadjimba KOUYATE **11312697**
Amine ADIB **11005061**

Construction d'une base de données d'apprentissage pour le spatial

Encadré par **Fabien DUCHATEAU**

Université Claude Bernard Lyon1 Master1-Informatique 2015/2016

Introduction:

Dans le cadre de l'unité d'enseignement MIF20 - Projet de recherche, nous participerons au développement d'un sujet qui traite les données du spatial. L'intitulé de ce sujet est "*Construction d'une base de données d'apprentissage pour le spatial*".

Ce sujet nous a été proposé par Monsieur Fabien DUCHATEAU Enseignant à l'université Claude Bernard Lyon1 et chercheur au laboratoire LIRIS.
Monsieur DUCHATEAU sera notre encadrant tout au long de la réalisation de ce projet de recherche.

Contexte:

De nos jours, les applications de système d'informations Géographiques aux problèmes d'aménagement, de localisation et de gestion des territoires sont en pleine expansion.

Ainsi, le nombre de bases de données géographiques et le volume de données associées ne cessent d'augmenter, notamment grâce aux techniques nouvelles de géocodage.

Ces données fournissent plus de détails sur les lieux mais peu d'informations sur les types de relations non topographiques [1] qui existent entre ces lieux, par exemple la phrase "**L'architecture du musée de Valence est similaire à celle du musée de Confluence**" décrit une relation non topographique entre deux lieux.

Il y a effectivement des logiciels permettant d'extraire des relations, mais qui nécessitent une base de données d'apprentissage.
La problématique majeure est que cette base de données n'existe pas encore, elle devra donc être mise en place afin de collecter des relations non topographiques entre ces entités spatiales.

Objectifs:

L'existence de relations entre lieux dans des documents textuels nécessite la mise en place d'une base de données d'apprentissage. Cette dernière servira de stockage de données préalablement extraites automatiquement.
Notre objectif au cours de ce projet est de construire cette base de données d'apprentissage contenant des relations non topographiques entre entités spatial.
Ainsi nous devrons mettre en place un algorithme permettant d'analyser des documents textuels et de faciliter la découverte de relations entre lieux en exploitant l'extraction de ces documents textuels.

Description des besoins:

Le programme développé proposera plusieurs fonctionnalités à l'utilisateur. Tout d'abord l'utilisateur devra pouvoir importer des documents textuels à analyser (en saisissant directement le texte ou en générant un texte aléatoire depuis *wikipédia*). Celui-ci découpera les champs textuels en sous parties (des paragraphes par exemple) afin de faciliter leurs exploitations.
Nous exploiterons les résultats d'un outil d'extraction d'informations (comme DBpedia Spotlight [2] , Alchemy API [3] …) pour repérer les différentes entités spatiales (les points géographiques qui représentent des lieux), pour ensuite dérouler notre algorithme qui permettra d'analyser et enfin reconnaître les relations spatiales (non topographiques) qui les lie.
Une fois l'analyse des documents finie le programme sollicitera l'intervention de l'utilisateur afin qu'il expertise les relations suggérées par l'outil. Il décidera donc de les valider (ou pas).
A la fin du processus le programme permettra à l'utilisateur de peupler sa base de données avec les relations spatiales extraites.

Les contraintes:

La principale contrainte à laquelle nous serons confrontés est le fait d'établir une règle générale plus ou moins logique qui sera la base de notre algorithme d'extraction de relations spatiales entre les différents lieux.

Livrables:

- Cahier des charges.
- Liste des types de relations spatiales (surtout les non topographiques). Établir une règle plus ou moins générale pour les identifier.
- Comparaison et évaluation des différents outils de détection d'entités spatiales. En choisir un selon notre besoin.
- Proposition d'algorithme pouvant analyser des documents textuels afin d'extraire les différentes informations (relations topographiques et non topographiques) portant sur les lieux repérés grâce à l'outil de détection choisi.
- Une interface graphique intuitive facilitant l'intervention des utilisateurs pour la validation des relations découvertes par l'algorithme.
- Rapport final.

Organisation du projet: *(Figure 1)*

- Rédiger le cahier des charges. [3 jours]
- Lire et analyser un article sur l'extraction d'informations [4] et sur les relations spatiales [5] . [9 jours]
- Implémentation du programme. [4 semaines]
- Rédaction du rapport final. [4 jours]
- Préparation à la soutenance. [4 jours]

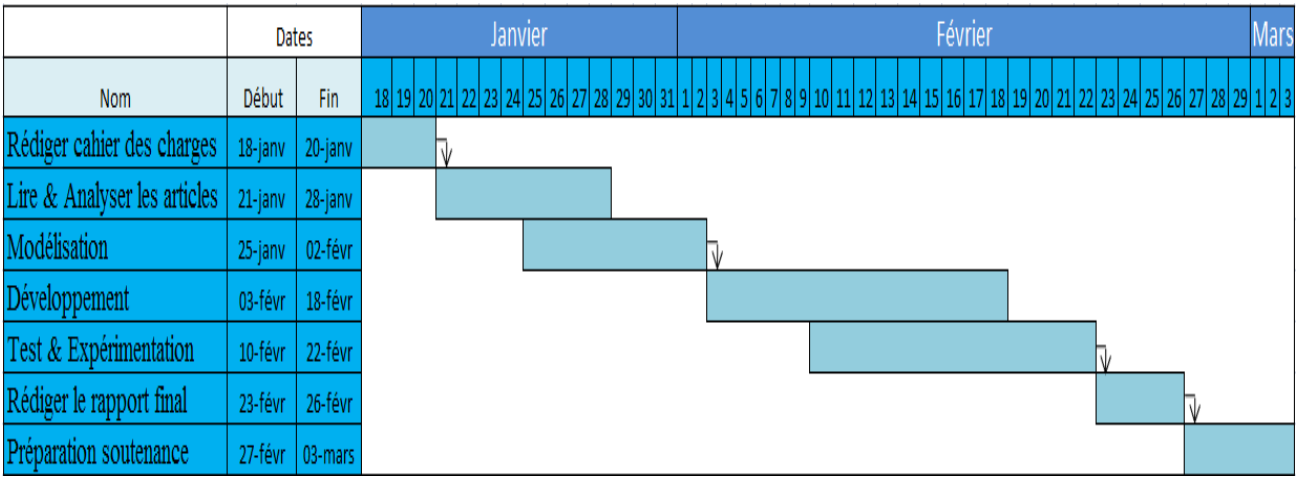

Figure1-Diagramme de Gantt

Références:

[1] Liens vers la définition de topographie Cliquez ici
[2] Lien vers l'outil DBpedia Spotlight Cliquez ici
[3] Liens vers l'outil Alchemy API Cliquez ici
[4] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. Mausam. Open information extraction The second generation. In IJCAI, volume 11, pages 3–10, 2011. Lien vers IJCAI11-OIE.pdf
[5] J. Strötgen, M. Gertz, and P. Popov. Extraction and exploration of spatio-temporal information in documents. In Proceedings of the 6th Workshop on Geographic Information Retrieval, page 16. ACM, 2010. Lien vers GIR10-spatial-extraction.pdf