Stage de M2 2020-2021 – Cirad **Montpellier**.

Titre. Extraction d'entités nommées géographiques et exploitation dans le cadre du Web de données

Mots-clés. Web sémantique, Web de données, extraction d'entités nommées, traitement automatique des langues, indexation, descripteurs géographiques.

Encadrants. **Anne Toulet**, Cirad (anne.toulet@cirad.fr), **Andon Tchechmedjiev**, IMT Mines Alès (andon.tchechmedjiev@mines-ales.fr)

Lieu du stage. **CIRAD**, Avenue Agropolis, 34398 **Montpellier** Cedex – France.

Durée du stage. 6 mois

Description du stage.

Ce stage s'inscrit dans le cadre du projet Issa (Indexation Sémantique d'une archive scientifique et Services Associés pour la science ouverte), lauréat de l'appel à projet CollEx-Persée 2019/2020. Le projet Issa est multipartenaire (Cirad Montpellier, IMT Mines Alès, Inria/Cnrs Sophia Antipolis). Dans ce contexte, le/la stagiaire sera amené.e à échanger et coopérer avec les deux autres partenaires du projet, IMT Mines Alès et Inria Sophia Antipolis, qui accueilleront également deux stagiaires M2 sur des sujets connexes. Des déplacements seront éventuellement à prévoir entre les différents sites. Ainsi, ce stage bénéficiera d'une forte dynamique de groupe.

Contexte scientifique

La science ouverte est un mouvement international qui cherche à rendre la recherche scientifique et les données qu'elle produit accessibles à tous. Dans cet objectif, les archives ouvertes – bases de données documentaires accessibles librement et gratuitement sur internet contenant des documents issus de la recherche scientifique – accentuent leurs efforts pour accroître l'accessibilité aux ressources dont elles disposent.

L'objectif de ce stage est de permettre un accès et une interopérabilité accrus à des publications scientifiques proposées par une archive ouverte en adoptant des techniques d'indexation[1] sémantique adossée à des référentiels terminologiques standards. **Le recours aux techniques du Web sémantique et du traitement naturel des langues sera privilégié**.

Dans ce travail, nous nous intéresserons en particulier à la question de l'**indexation par des mots-clés géographiques**.

Agritrop[2], l'archive ouverte des publications du Cirad, servira de cas d'utilisation tout au long du stage. Ce portail propose essentiellement des publications scientifiques mais aussi un fonds de cartes et de documents anciens. Chaque document est décrit par des métadonnées riches parmi lesquelles se trouvent des descripteurs thématiques et géographiques issus du thesaurus **Agrovoc**[3]. Ainsi, Agritrop permet de disposer d'un corpus annoté de plus de 10.000 publications (titre, résumé, mots clés, etc.).

Actuellement, les descripteurs géographiques sont choisis dans la hiérarchie des concepts d'Agrovoc mais ils ne satisfont pas à un certain nombre d'exigences, en particulier en terme de précision spatio-temporelle. Le **besoin de s'adosser à des terminologies géographiques plus spécifiques**, prenant en charge les différents aspects de **géoréférencement**, de **toponymie**, de **désambiguïsation des termes** est un point crucial dans cette réflexion. Le recours à un vocabulaire standardisé de descripteurs

[1] L'indexation permet de préciser le contenu d'un document à travers des mots-clés et ainsi de retrouver dans un catalogue tous les documents qui traitent d'un sujet donné quel que soit le support.
[2] https://agritrop.cirad.fr/
[3] Thésaurus développé et maintenu par la FAO depuis le début des années 80 (http://aims.fao.org/fr/agrovoc)

géographiques doit permettre de répondre à des besoins aussi variés que la recherche de documents, leur analyse manuelle ou automatique et leur représentation graphique sous forme de data visualisation.

Tâches à accomplir

1. États de l'art
   - Dresser un inventaire détaillé des référentiels géographiques existants en comparant leurs différentes approches. On s'intéressera particulièrement à GeoNames.
   - Étudier et comparer les différents outils d'extraction d'entités nommées existants, en particulier pour les entités géographiques.
2. Conception et tests d'une chaîne de traitement
   Développer et appliquer une chaîne d'extraction, de désambiguïsation et de liage d'entités nommées géographiques dans des publications scientifiques sur la base de méthodes étudiées dans l'état de l'art, en particulier celles basées sur les techniques d'apprentissage profond ou Entity Linking. Cette chaîne de traitement sera testée sur un jeu de publications scientifiques issues d'Agritrop.
3. Exploitation de l'indexation sémantique
   Exploiter les descripteurs géographiques obtenus en utilisant les technologies du Web sémantique et le Web de données, par exemple en permettant la visualisation géographique (cartographie) et/ou l'enrichissement encyclopédique à partir des mots-clés géographiques obtenus précédemment.

Le/la stagiaire devra appliquer les règles de bonne pratique de développement et d'industrialisation d'une application. Un soin particulier sera apporté à la réalisation d'une application web "packagée" comme un produit fini, facilement déployable et configurable dans d'autres environnements.

Références.

- Arnaud, J. Cataloguer, rechercher des cartes. Le référencement géographique en question. Documentaliste-Sciences de l'Information, vol. 51(3), 68-79., 2014. https://doi.org/10.3917/docsi.513.0068
- J. Li, A. Sun, J. Han and C. Li, "A Survey on Deep Learning for Named Entity Recognition," in *IEEE Transactions on Knowledge and Data Engineering*, https://doi.org/10.1109/TKDE.2020.2981314
- Named Entity Recognition : une application pratique du NLP. Katy Fokou. 06/12/2019 https://www.smalsresearch.be/named-entity-recognition-une-application-du-nlp-utile/
- Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. REL: An Entity Linker Standing on the Shoulders of Giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '20*). Association for Computing Machinery, New York, NY, USA, 2197–2200. DOI:https://doi.org/10.1145/3397271.3401416
- A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. Vikas Yadav, Steven Bethard. Proceedings of the 27th International Conference on Computational Linguistics. 2018. https://www.aclweb.org/anthology/C18-1182
- Hengchen, S., van Hooland, S., Verborgh, R. & De Wilde, M. (2015). L'extraction d'entités nommées : une opportunité pour le secteur culturel ?. I2D – Information, données & documents, volume 52(2), 70-79. https://doi.org/10.3917/i2d.152.0070
- GIS and Named Entity Recognition: Identifying Geographic Locations in Text. Mark Altaweel | February 13, 2019 | Spatial Analysis. https://www.gislounge.com/gis-and-named-entity-recognition-identifying-geographic-locations-in-text/ Voir aussi : https://github.com/geovista/GeoTxt
- Index géographique : avant le géocodage, le gazetteer. Frédéric Rodrigo. 2019. https://makina-corpus.com/blog/metier/2019/index-geographique-avant-le-geocodage-le-gazetteer