Sujet master informatique : Outil d'annotation sémantique pour des données massives et hétérogènes : application aux données spatio-temporelles du système Terre

1 – Encadrement

Jean-Christophe Desconnets (ESPACE-DEV,IRD)

1 – Description du sujet

Contexte

À l'échelle globale, les océans, l'atmosphère et la biosphère sont l'objet de changements majeurs d'une rapidité sans précédent. Les enjeux associés à ces changements appellent à un développement de connaissances sur le système Terre. Ces connaissances sont construites par l'utilisation conjointe des données issues des observations satellites, de terrain ou encore des sorties de modèle de simulation des phénomènes étudiés. Ces divers systèmes génèrent des volumes de données considérables dans divers formats, hébergés par de nombreux centres de données et de calcul. L'étape de découverte des données est un défi de premier ordre pour connaître leur disponibilité, assurer leur réutilisation et/ou leur combinaison pour de nouvelles analyses.

Problématique

L'approche actuelle est de fédérer les bases de données existantes pour en fournir une vue complète et unifiée en vue de permettre leur interrogation. La volumétrie des données nous imposent de baser nos interrogations sur les métadonnées. La transversalité des enjeux scientifiques nous demande de pouvoir rendre découvrables les données au delà d'une discipline. Pour cela, nous avons choisi de décrire les données en utilisant une ontologie disciplinairement neutre, basée sur le paradigme d'observation [Beretta et al., 2020].
Actuellement, les données sont décrites dans les catalogues des systèmes d'observation. Ils sont construits sur des annotations sémantiques faiblement standardisés, incomplètes, voire imprécises. A ce stade, elles ne permettent pas de mettre en oeuvre une indexation efficace sur ces grandes masses de données.

Objectifs

Pour cela, notre objectif est de transformer et enrichir les catalogues existants sur la base de l'ontologie d'observation et des ressources onto-terminologiques disciplinaires. Du fait de l'incomplétude, de l'imprécision et de l'hétérogénéité des métadonnées, il est proposé d'explorer l'apport des techniques de classification par apprentissage pour automatiser la standardisation des annotations existantes (sémantique et syntaxique) et la classification des métadonnées en s'appuyant sur notre ontologie métier. Plusieurs grands jeux de métadonnées venant des catalogues d'observatoires seront mis à disposition.
Il s'agit de proposer

1) une méthodologie originale faisant appel aux techniques d'apprentissage existantes pour standardiser et classifier les métadonnées,
2) l'implémentation d'un prototype qui permettra de mettre en oeuvre les opérations de transformation, standardisation et de classification des métadonnées.

3) proposer des métriques d'évaluation qui permettront de juger de l'adéquation de la méthode aux métadonnées traitées.

Références

V. Beretta, J-C Desconnets, I. Mougenot, M. Arslan, J. Barde & V. Chaffard (2020) : A user-centric metadata model to foster sharing and reuse of multidisciplinary datasets in environmental and life sciences. submitted Computers and Geoscience journal.

C. ROUSSEY,S. BERNARD, G. ANDRÉ, D. BOFFETY. Weather Data Publication on the LOD using SOSA/SSN Ontology. Semantic Web Journal, 2019. http://www.semantic-web-journal.net/content/weather-data-publication-lodusing-sosassn-ontology0

2 – Résultats attendus

- Prototype permettant 1) d'assurer la transformation des données 2) l'annotation sémantique de données spatio-temporelles et 3) l'évaluation de la qualité des différentes approches d'apprentissage proposées
- Etude comparative des méthodes d'apprentissage proposées pour leur sélection en fonction des caractéristiques des jeux de données à annoter.

3 – Prérequis

- Bonne maîtrise des concepts, méthodes et outils liés à la modélisation de données et de connaissances.
- Connaissance des technologies du web sémantique (concepts, langages).
- Maîtrise d'outils de construction, d'alignements ou d'agrégation d'ontologies.
- Bonne maîtrise d'un langage de programmation à l'exemple de Java ou Python, et des librairies associées pour manipuler les représentations de données sous forme de graphe (RDF) et les techniques d'apprentissage (machine learning, deep learning)