Sujet master informatique : outil d'aide à l'alignement de ressources onto-terminologiques : application aux thésaurus des sciences de la Terre et du vivant

1 – Encadrement

Jean-Christophe Desconnets (ESPACE-DEV,IRD), Isabelle Mougenot (ESPACE-DEV, UM)

2 – 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. D'un point de vue sémantique, ces catalogues reposent sur une utilisation très disparate et hétérogène (listes contrôlées de valeurs, ressources onto-terminologiques) qui sont utilisées comme descripteurs des données (contenu, localisations temporelle et spatiale).

Travaux

Notre objectif est d'être en mesure d'aligner les différentes ressources onto-terminologiques disciplinaires de sorte que la découverte et la navigation entre des données issues de différentes disciplines soient possibles. Appliqué à quelques thésaurus existants, notre démarche se veut extensible à d'autres ressources ontologiques.
Les travaux porteront sur :

1) La mise au point d'une méthodologie originale et adaptée à l'alignement des ressources onto-terminologiques. Suite à des travaux préliminaires, nous souhaitons orienter l'étude sur les techniques basées sur les chaînes de caractères, le langage et l'utilisation de ressources linguistiques externes [Jain et al., 2010; Bellahsene et al.,2017 ; Mazuel & Charlet, 2009 ; Jentzsch et al., 2010]. Elles apparaissent adaptées à nos ressources onto-terminologiques (terminologies, des vocabulaires contrôlés et des thésaurus).

2) L'implémentation d'un outil générique qui utilisera et/ou complétera les outils d'alignements existants. Il viendra compléter un service de registre assurant la gestion et l'accès standard (API REST, SPARQL) aux ressources terminologiques des centres de données.
3) Des recommandations seront également attendues pour assurer l'automatisation, la mise à jour des alignements à plus grande échelle.

3 – Résultats attendus

- Etat de l'art et analyse des méthodes d'alignement adaptées aux ressources onto-terminologiques des sciences de la Terre,
- Proposition d'une méthodologie d'alignements,
- Prototype assurant les alignements, leur évaluation et leur exportation pour enrichir les ressources onto-terminologiques existantes,
- Recommandations pour automatiser et gérer la production d'alignements sur de nouvelles ressources.

Les codes sources seront versés à un dépôt GIt et ouverts (open source) à la communauté scientifique sous une licence libre.

4 – 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 Java, Python
- Maîtrise des librairies du web sémantique (Java Jena,OWLReady Python, ...) pour manipuler les ontologies RDF et OWL) et les techniques d'alignements.

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.

Jain, P., Hitzler, P., Sheth, A. P., Verma, K., & Yeh, P. Z. (2010, November). Ontology alignment for linked open data. In *International semantic web conference* (pp. 402-417). Springer, Berlin, Heidelberg.

Zohra Bellahsene, Vincent Emonet, DuyHoa Ngo, Konstantin Todorov: YAM++ Online: A Web Platform for Ontology and Thesaurus Matching and Mapping Validation. ESWC (Satellite Events) 2017: 137-142

Laurent Mazuel, Jean Charlet, SPIM-AlignmentGUI - un logiciel d'aide à la réalisation d'alignements entre ontologies, Proc. IC Poster session, Hammamet (TN), 2009

Anja Jentzsch, Robert Isele, Christian Bizer: Silk - Generating RDF Links while publishing or consuming Linked Data. Poster at the International Semantic Web Conference (ISWC2010), Shanghai, November 2010.

Exemples de ressources onto-terminologiques existantes dans le domaine des sciences de la Terre et du Vivant

GCMD https://gcmdservices.gsfc.nasa.gov/
NERC vocabulary server (http://vocab.nerc.ac.uk/
Thesaurus Eau http://thesaurus.oieau.fr/thesaurus/page/ark:/99160/7af302a6-7518-4a8a-84a6-b8df7b595e14),
SWEET ontology http://sweet.jpl.nasa.gov/ontology/
TaxRef https://taxref.mnhn.fr/taxref-web/accueil
WORMS: http://www.marinespecies.org/