Fiche de présentation de TER
2012-2013

Titre :

Extraction et réconciliation d’alignements multi-langue dans des ontologies biomédicales.

Information :

Encadrant :	Clement Jonquet (LIRMM, UM2) – jonquet@lirmm.fr
Spécialités :	DECOL, AIGLE
Nombre d’étudiants :	2-3
Contexte:	Projet SIFR (Semantic Indexing of French Biomedical Data Resources)
Ou:	LIRMM, SMILE & TEXTE research team
Quand:	2nd semestre 2012-2013

Mots clés :

Web application, ontologies, knowledge representation, semantic Web, database

Technologies :

OWL, SKOS, Java, JEE, MySQL, Tomcat, ResTful web services (RestLet), XML, RDF, BioPortal Web services API

Résumé :

Les terminologies et ontologies biomédicales jouent un rôle clé dans l’interopérabilité sémantique des données des sciences du vivant en servant de dénominateur commun. Pour construire des applications cliniques, médicales ou industrielles, il est crucial que les chercheurs convergent vers un ensemble de méthodes et de formats interopérables pour le traitement des données. L’Université de Stanford et le CHU de Rouen ont développé des portails pour les ontologies/terminologies biomédicales (e.g., édition, navigation, visualisation, annotation de données, indexation, etc.) qui assistent les professionnels de santé et les chercheurs en médecine dans la construction de système à base de connaissances qui utilisent les ontologies. Cependant, les terminologies, lorsque décrites dans des langages naturel différents (e.g., anglais pour Stanford, français pour Rouen) ne sont pas liées les unes aux autres de façon formelle et standard. Le travail du TER consiste à extraire des alignements multilingues (c’est-à-dire des alignements entre concepts définis dans des ontologies similaires mais de langues différentes) à partir de diverses sources de données (OWL, SQL), en utilisant diverses méthodes d’extraction automatique et à réconcilier automatiquement ces alignements dans la plateforme BioPortal via l’API REST.

Contexte :

A key aspect in addressing semantic interoperability for life sciences is the use of terminologies and ontologies as a common denominator to structure biomedical data and make them interoperable. Ontologies formalize the knowledge of a domain by means of concepts, relations and rules that apply to that domain [4].

The *Stanford Center for Biomedical Informatics Research* (BMIR) group at Stanford University (http://bmir.stanford.edu) and the *Catalogue et Index des Sites Médicaux de langue Française* (CISMeF) group at Rouen University Hospital (http://www.cismef.org) have both invested lot of efforts in developing terminology/ontology-based tools and services to assist health professionals and users in their search for electronic information available on the Web and in the use of ontologies. The two groups have developed Web-based portals, respectively the *NCBO Bioportal* [5] and the *CISMEF Health Multi-Terminology Portal [2]*, that offer a variety of services to search or index biomedical data as well as searching, exploring, annotating and visualizing the available standards ontologies.

Both portals aim to deal with multilingualism. This goes from simply enabling switching between multilingual ontologies to enabling the information retrieval or data mining algorithms to leverage the information available in another language. However, none of the portals presently use a formal and standard model to represent multi-lingual mappings existing between the same concepts in different languages ontologies [3]. We have already discussed a common format to represent multi-lingual mappings within BioPortal which is the platform that hosts ontologies in different languages.

The TER aims to implement several methods to extracts the multilingual mappings from different sources of data and then reconcile the extracted mappings into a unique repository hosted by the BioPortal platform and accessible via REST web services.

Présentation du sujet :

You will extract multilingual mappings from several data sources and using several approaches (sorted hereafter from simplest to harder):

- From label descriptions within an ontology description file e.g., OWL. Indeed, some ontologies provides multilingual labels using the xml:lang property or another specific syntax. This is a simplest case.

- From the UMLS Metathesaurus [1] which is set of terminologies which are manually integrated and distributed by the United States NLM. Indeed, UMLS include a few French terminologies.

- From the CISMEF information system and the HMTP portal which is the biggest source of French-English mappings for biomedical terms. Format to be clarify.

- From other unilingual mappings existing between ontologies (eng-eng or fr-fr). Indeed, both HMTP and BioPortal include large number of mappings between the ontologies they host.

- From our own multilingual dictionary (alignment of terms) built along the execution of the previous approaches or other public dictionary that would be available online e.g., WordNet, GoogleTranslate.

We will systematically represent the mappings using the semantic web standard (i.e., RDF) and use the appropriate URIs provided by BioPortal. You will store the extracted mappings locally (relational DB hosted at LIRMM) and develop a simple curation web application with user interface that external curators or evaluators would use to validate/evaluate the extracted multilingual mappings.

We will be highly concerned by reuse the methods to process new ontologies in the future and automatically generate multilingual alignments with already existing ontologies.

Finally, you will implement the procedures to upload the mappings into the BioPortal mappings repository via the BioPortal REST web service API described at: http://www.bioontology.org/wiki/index.php/BioPortal_REST_services.

Références :

[1] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270, 2004.

[2] Stéfan J. Darmoni, Suzanne Pereira, Saoussen Sakji, Tayeb Merabti, Élise Prieur, Michel Joubert, and Benoit Thirion. Multiple Terminologies in a Health Portal: Automatic Indexing and Information Retrieval. In C. Combi, Y. Shahar, and A. Abu-Hanna, editors, *12th Conference on Artificial Intelligence in Medicine, AIME'09*, number 5651 in Lecture Notes in Computer Science, pages 255–259, Verona, Italy, June 2009. Springer.

[3] Bo Fu, Rob Brennan, and Declan O'Sullivan. Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web. In Buitelaar P., Cimiano P., and Montiel-Ponsoda E., editors, *1st Workshop on the Multilingual Semantic Web*, volume 571 of *CEUR*, pages 13–20, Raleigh, NC, USA, April 2010.

[4] Tom R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, June 1993.

[5] Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, Clement Jonquet, Daniel L. Rubin, Margaret-Anne Storey, Christopher G. Chute, and Mark A.

Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37((web server)):170–173, May 2009.

[6] Patricia L. Whetzel, Natalya F. Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39((web server)):541–545, June 2011.