Laboratoire d'Informatique de Robotique et de Microélectroniq de Montpellier
LIRMM

SIFR project

Internship position
Master in computer science
2014-2015

Title:

Multilingualism in an ontology repository: the case of BioPortal

Information :

Superviser:	Clement Jonquet (LIRMM, UM2) – jonquet@lirmm.fr
Profile:	Computer science or informatics master students
Context:	Project SIFR (Semantic Indexing of French Biomedical Data Resources)
Where:	University of Montpellier (ex UM2), Laboratory of Informatics, Robotics, & Microelectronics of Montpellier (LIRMM)
When:	2nd semester 2014-2015

Keywords:

BioPortal, multilingual semantic Web, biomedical ontologies, knowledge representation, multilingual ontology alignments, web application, databases.

Technologies:

Semantic Web technologies (RDF, OWL, SKOS, SPARQL), Web application technologies (Java, JEE, Ruby, Ruby on Rails, Web services API)

French Abstract:

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 a développé BioPortal (http://bioportal.bioontology.org), une plateforme web sémantique 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, le portail ne gère pas les aspects multilingues. Gérer le multilinguisme dans un portail d'ontologies ne se limite bien sûr pas à offrir l'interface graphique dans plusieurs langues. Il faut se poser les questions de la représentation multilingue des données du portail (ontologies, alignements) et de leur valorisation dans les services offerts (recherche, indexation, annotation) ; le tout en utilisant les standards du web sémantique.

Le travail du stage consiste dans un premier temps à faire un état de l'art sur la gestion du multilinguisme dans les applications web sémantique et de faire des propositions pour la plateforme BioPortal qui seront implémentées dans une instance locale du portail déployée au LIRMM. Dans un second temps, il faudra 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.

Context (eng):

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 [5]. The *Stanford Center for Biomedical Informatics Research* (BMIR) group at Stanford University (http://bmir.stanford.edu) has 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 group has developed a Web-based portal, the *NCBO Bioportal* [8] that offer a variety of services to search or index biomedical data as well as searching, exploring, annotating and visualizing the available standards ontologies. The portal currently only deals with English ontologies and does not deal with multilingualism.

We have already discussed a set of propositions to deal with multilingualism within BioPortal and to represent multilingual mappings [6]. Our main objective is to handle multilingualism in a proper semantically rich and consistent manner enabling BioPortal users to use ontologies independently of the language and therefore enabling cross lingual search or annotation with ontologies and mining of data indexed with ontologies.

The internship aims first to review the current solutions and approach to deal with multilingualism in semantic web applications at the light of recent proposition that have been done in the community. Then the intern will 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 our local instance of the BioPortal platform.

Presentation:

State of the art in the domain will include reviews of approaches such as: GOLD [3], DOOR [1], SKOS-XL, LEMON [7] and its predecessors (Lexvo, Lingvoj, etc.) as well as the recent LEMON translation module [4].

Proposition for representation of multilingual content in BioPortal will include:

- Classifying multilingual ontologies;

- Review multilingual ontologies in BioPortal;

- Representation of natural language property for an ontology;

- Representation of relation between ontologies in different languages;

- Representation of multilingual mappings;

- BioPortal internationalization (content and interface).

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.

- From the UMLS Metathesaurus [2] 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 clarified.

- From other unilingual mappings existing between ontologies (eng-eng or fr-fr). BioPortal includes large number of mappings between the ontologies.

- From other multilingual dictionary (alignment of terms) that would be available online e.g., WordNet, GoogleTranslate.

You will systematically represent the mappings using the semantic web standard (i.e., RDF) and use the appropriate URIs provided by BioPortal (local instance and original one). Finally, you will implement the procedures to upload the mappings into the local BioPortal mappings repository via the BioPortal REST web service API.

References:

[1] Carlo Allocca, Mathieu d'Aquin, and Enrico Motta. Towards a Formalization of Ontology Relations in the Context of Ontology Repositories. In A. Fred, J.L.G. Dietz, K. Liu, and J. Filipe, editors,

Knowledge Discovery, Knowlege Engineering and Knowledge Management, volume 128 of *Communications in Computer and Information Science*, pages 164–176. Springer, 2011.

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

[3] Scott Farrar and Terence Langendoen. A linguistic ontology for the semantic web. *Glot International*, 7(3):97–100, 2003.

[4] Jorge Gracia, Elena Montiel-Ponsoda, Daniel Vila-Suero, and Guadalupe Aguado de Cea. Enabling Language Resources to Expose Translations as Linked Data on the Web. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *9th International Conference on Language Resources and Evaluation, LREC'14*, pages 409–4013, Reykjavik, Iceland, May 2014. European Language Resources Association.

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

[6] Clement Jonquet and Mark A. Musen. Gestion du multilinguisme dans un portail d'ontologies: étude de cas pour le NCBO BioPortal. In C. Roche, R. Costa, and E. Coudyzer, editors, *Terminology & Ontology : Theories and applications Workshop, TOTh'14*, page 2, Brussels, Belgium, Dec. 2014.

[7] John McCrae, Dennis Spohr, and Philipp Cimiano. Linking lexical resources and ontologies on the semantic web with lemon. In G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. DeLeenheer, and J.Z. Pan, editors, *8th Extended Semantic Web Conference, ESWC'11*, number 6643 in Lecture Notes in Computer Science, pages 245–259, Heraklion, Crete, Greece, May 2011.

[8] 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.

Expected profile:

- Computer science or informatics master degree students.
- Experience with semantic Web technologies.
- Good English oral and writing skills. Good knowledge of French or another EU language is desirable.
- Excellent writing skills as reports, documentations, and technical notes will always be necessary.
- Autonomy and initiative, take on technical decisions within the project and justify choices.
- Friendly person to join a small research team in Montpellier.

Application:

For more information about this position, please contact Clement Jonquet (jonquet@lirmm.fr). To apply, please send an email including links to (NO ATTACHED DOCUMENTS) the following:

- a motivation letter describing an explanation of YOUR interest for the position;
- a curriculum vitae describing your experience and the matches with the expected profile;
- copies of diplomas and other relevant grade certificates;
- names and contact details of referees.