HAL
open science

Gestion du multilinguisme dans un portail d'ontologies: étude de cas pour le NCBO BioPortal

Clement Jonquet, Mark A. Musen

To cite this version:

Clement Jonquet, Mark A. Musen. Gestion du multilinguisme dans un portail d'ontologies: étude de cas pour le NCBO BioPortal. TOTh Workshop, Dec 2014, Bruxelles, Belgique. lirmm-01099882

HAL Id: lirmm-01099882

https://hal-lirmm.ccsd.cnrs.fr/lirmm-01099882v1

Submitted on 5 Jan 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TOTh workshop:
Multilingual Thesaurus and Terminology

December 5th 2014 at the Cinquantenaire Museum, Brussels, Belgium
http://porphyre.org/workshop-toth/2014-en

Clement Jonquet
Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier
jonquet@lirmm.fr

Mark A. Musen
Stanford Center for Biomedical Research (BMIR) – Stanford University
musen@stanford.edu

Gestion du multilinguisme dans un portail d'ontologies: étude de cas pour le NCBO BioPortal

Les terminologies et les ontologies jouent un rôle central en sciences de la vie pour structurer les données biomédicales et les rendre interopérables [2]. L'utilisation d'ontologies pour indexer et intégrer les ressources de données est un moyen de valoriser la connaissance en facilitant la recherche et la fouille de données. Cependant, les découvertes qui pourraient être réalisées sont souvent limitées par la disponibilité et le traitement des données dans une langue seulement, le plus souvent l'anglais, pour laquelle il existe le plus d'ontologies et d'outils. Dans le cadre du projet *Indexation sémantique de ressources biomédicales francophones* (SIFR - http://www.lirmm.fr/sifr), nous nous intéressons à la gestion du multilinguisme dans la plateforme BioPortal (http://bioportal.bioontology.org) du Centre National pour les Ontologies Biomédicales (NCBO). BioPortal [7] permet d'accéder, visualiser, rechercher et commenter plus de 350 ontologies ou terminologies (principalement en anglais) de différent domaine en biologie ou médecine. Les ontologies peuvent être utilisées pour annoter automatiquement des données textuelles et le portail offre également un index sémantique de plusieurs jeux de données biomédicales annotées avec les ontologies du portail. Les utilisateurs ont accès à la plateforme soit via une application Web, soit via une interface de service Web. Dans la suite, nous présentons les choix qui permettront au portail de gérer les ontologies multilingues et les traductions d'ontologies de manière consistante sémantiquement et transparente pour les utilisateurs. 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).

Vocabulaire

Dans la suite nous parlerons seulement d'ontologie (incluant ainsi la notion de terminologie). Nous parlons d'*ontologies multilingues* quand elles offrent des labels dans différentes *langues naturelles* (e.g., Orphanet) et utilisent les mécanismes standards pour distinguer les labels (e.g., rdfs:label et xmlang) ou une représentation lexicale riche (e.g., SKOS, SKOS-XL, LEMON [6], Lexvo [3], Lingvoj). Nous parlons d'*ontologies monolingues* quand elles offrent des labels dans une seule langue qui sert, en général, de référence à la conceptualisation. Ces ontologies sont soient proposées de manière originale dans un langage donné (e.g., MeSH) soit sont le résultats d'une traduction d'une ontologie dans un autre langage (e.g., MeSH-fr). Nous parlerons également d'*ontologies multilingues partielles* pour identifier les ontologies multilingues qui n'offrent pas systématiquement tous les labels dans toutes les langues, ce qui rend leur exploitation plus difficile. Une *traduction* est vue comme la relation entre deux ontologies monolingues de langue différente qui représente principalement le même contenu (domaine, classes et relations). Nous parlerons *d'alignement multilingue*, pour un alignement 1-à-1 entre deux concepts d'ontologies monolingues et *d'alignement multilingue de traduction*, lorsque les deux ontologies concernées sont une traduction l'une de l'autre. Par exemple, le terme Mesh-

fr/mélanome est un alignement multilingue de traduction de Mesh/melanoma mais n'est seulement qu'un alignement multilingue de DOID/melanoma.

Statut de la gestion du multilinguisme dans BioPortal aujourd'hui

A l'heure actuelle, BioPortal est principalement 'anglais' et ne gère pas le multilinguisme. Le portail accepte et traite les ontologies multilingues et monolingues mais n'est pas encore capable respectivement de mettre en valeur la richesse sémantique des unes et de gérer les alignements multilingues des autres. Les ontologies monolingues autres qu'anglaises, sont parfois représentées sous forme de vues (i.e., un mécanisme qui permet de manipuler un sous ensemble ou une sous partie d'une ontologie), mais ce n'est pas systématique car BioPortal n'a pas de solution pour représenter les relations entre ontologies. En outre, les ontologies monolingues ne sont pas systématiquement excluent de certains service de la plateforme comme par exemple l'annotation automatique, ce qui peut créer des résultats complètement inappropriés si des ontologies d'une autre langue sont utilisées pour annoter des données dans une langue différente. Finalement, le portail n'est pas du tout internationalisé, ni au niveau du contenu, ni au niveau des interfaces.

Représentation de la propriété langage naturel d'une ontologie

Pour représenter la/les langue(s) d'une ontologie, nous proposons d'utiliser la propriété omv:naturalLanguage inclue dans l'ontologie OMV (http://omv2.sourceforge.net) qui est elle-même utilisée dans l'ontologie des métadonnées de BioPortal, Metadata, qui représente les propriétés d'une ontologie ainsi que d'autres éléments de BioPortal (projets, alignements, etc.). Cette propriété utilise l'ISO-639-3 pour ses valeurs.

Représentation de la distinction linguistique des ontologies

Pour affiner la représentation du type d'une ontologie, nous pouvons étendre OMV pour distinguer les 2 types d'ontologies précédemment mentionnés. Ainsi, il faut créer les classes : meta:MultilingualOntology rdfs:subClassOf omv:Ontology avec comme contrainte omv:naturalLanguage some Literal et meta:LanguageSpecificOntology rdfs:subClassOf omv:Ontology avec comme contrainte omv:naturalLanguage exactly 1 literal. Cependant, cette solution ne permet pas de distinguer le cas des ontologies multilingues partielles.

Représentation des relations entre ontologies

Pour représenter la relation de 'traduction' entre les ontologies nous proposons d'utiliser l'ontologie DOOR [1] qui à ce jour est la seule ontologie connue pour représenter des relations entre ontologies. Nous devons cependant étendre cette ontologie avec une nouvelle relation pour représenter qu'une ontologie est une évolution spécifique d'une autre ontologie avec une syntaxe différente, soit une ontologie équivalente mais dans une autre langue. L'extension de DOOR peut être faite dans Metadata, ce qui donne : meta:isTranslationOf subPropertyOf door:explanationEvolutionOf, subPropertyOf door:syntactically EquivalentTo.

Représentation des alignements multilingues pour les ontologies monolingues

Pour garder un modèle simple et unique, nous suggérons d'utiliser le même moyen que pour représenter les autres alignements (i.e., 1-à-1) dans BioPortal mais avec une relation spécifique empruntée à l'ontologie GOLD (http://linguistics-ontology.org) [5] : gold:translation, qui permet de représenter que deux expressions ont globalement le même sens. Cette propriété peut être affinée par gold:freeTranslation, si les expressions on exactement le même sens et gold:literalTranslation si la traduction est mot à mot. Cette méthode ne nécessite pas de créer de relation spécifique e.g., frenchToEnglishTranslationOf car la propriété omv:naturalLanguage des deux ontologies concernées fournissent cette information.

Réconciliation des alignements multilingues

Quand nous disposons d'au moins deux ontologies monolingues qui sont la 'traduction' l'une de l'autre, nous devons réconcilier les alignements 1-à-1 entre ces deux ontologies dans la base d'alignements de BioPortal. Cette tache peut être plus ou moins difficile suivant le niveau d'information inclue dans l'ontologie traduite. Du plus facile au plus compliqué, par exemples : si les termes utilisent les mêmes codes (e.g., Mesh et Mesh-fr) ; si il existe un métathesaurus dans lequel

ces deux ontologies ont été intégrées (e.g., UMLS, CISMeF) ; si les alignements ont déjà étaient créés dans d'autres bases de connaissances ; si les alignements peuvent être induit d'autres alignements monolingues ; si il existe des base de données multilingues desquelles des alignements peuvent être extraits de traductions ; n'importe qu'elle approche complexe d'alignements ontologiques [4]. Dans tous les cas, ce travail doit être fait en post-traitement automatique (via l'API de service web) après l'inclusion des deux ontologies dans BioPortal.

Internationalisation de BioPortal

Une fois tous les aspects précédents gérés, *l'internationalisation* du portail devient possible. Avant tout, au niveau contenu, le portail (UI et service web) permet de passer de manière automatique d'une ontologie monolingue à une traduction. Pour chaque concept, les labels dans d'autres langues et/ou les alignements vers d'autres ontologies de langue différente sont disponibles. Ensuite, l'internationalisation de l'interface (menu, documentation, etc.) est la dernière étape comme pour n'importe quelle application web multilingue.

Conclusion et perspectives

L'enjeu de la gestion du multilinguisme dans le domaine des ontologies biomédicales est selon nous primordial et dépasse les aspects linguistiques. En effet, l'intégration multilingue de jeux de données permettra des études translationnelles sur des données relatives à des populations différentes. Et donc de pouvoir affiner les recherches médicales sur les des domaines tels que la pharmacogénomique, l'étude du rôle de l'environnement sur l'expression des gênes, ou les relations gènes-maladies. Dans les prochains mois, nous travaillerons à mettre en place les propositions précédentes au sein d'une instance locale de BioPortal déployée au LIRMM dans le contexte de SIFR.

Remerciements

Ce travail est financé principalement par l'Agence Nationale de la Recherche, projet ANR-12-JS02-01001, ainsi que par l'Université de Montpellier, le CNRS et l'IBC de Montpellier.

Références

[1] C. Allocca, M. d'Aquin, and E. Motta. Towards a Formalization of Ontology Relations in the Context of Ontology Repositories. In A. Fred, J. 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] O. Bodenreider and R. Stevens. Bio-ontologies: Current Trends and Future Directions. *Briefings in Bioinformatics*, 7(3):256–274, August 2006.

[3] G. de Melo. Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud. *Semantic Web*, page 7, July 2013.

[4] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Berlin Heidelberg, DE, 2007.

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

[6] J. McCrae, D. Spohr, and P. 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. 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. Springer.

[7] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. B. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, and M. A. Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37((web server)):170–173, May 2009.