An Approach to Compare Bio-Ontologies Portals

Julien Grosjean, Lina Soualmia, Khedidja Bouarech, Clement Jonquet, Stefan Darmoni

An Approach to Compare Bio-Ontologies Portals

Julien GROSJEAN [a,1], Lina F. SOUALMIA [a,b], Khedidja BOUARECH [c], Clément JONQUET [c] and Stéfan J. DARMONI [a,b]

[a] *CISMeF & TIBS, LITIS EA4108, Rouen University Hospital, France*
[b] *LIMICS, INSERM UMR 1142, Paris, France*
[c] *LIRMM, Université Montpellier 2 & CNRS, Montpellier, France*

Abstract. Background: main biomedical information retrieval systems are based on controlled vocabularies and most specifically on terminologies or ontologies (T/O). These classification structures allow indexing, coding, annotating different kind of documents. Many T/O have been created for different purposes and it became a problem for finding specific concepts in the multitude of existing nomenclatures. The NCBO (National Center for Biomedical Ontologies) BioPortal[2] and the CISMeF (Catalogue et Index des Sites Médicaux de langue Française) HeTOP[3] projects have been developed to tackle this issue. Objective: the present work consists in comparing both portals. Methods: we hereby are proposing a set of criteria to compare bio-ontologies portals in terms of goals, features, technologies and usability. Results: BioPortal and HeTOP have been compared based on the given criteira. While both portals are designed to store and make T/O available to the community and are sharing many basic features, they are also very different mainly because of their basic purposes. Conclusion: thanks to the comparison criteria, we can assume that a merge between BioPortal and HeTOP is possible in terms of functionnalities. The main difficulties will be about merging the data repositories and applying different policies on T/O content.

Keywords. Controlled Vocabulary, Internet, Information Storage and Retrieval, Terminology as subject, Health terminologies and ontologies, Crosslingual access to health information, Portal

Introduction

Because of the constant growth of biomedical data it became mandatory to index or annotate it with controlled and structured vocabularies in order to store it and to retrieve it intelligently. A key aspect in addressing semantic interoperability for biomedical data is the use of terminologies or ontologies (T/O) as a common data structure data [2] [5]. Many T/O have been created in the last decade for different purposes: indexing or annotating documents, organising knowledge, inferring facts, etc.
Several tools have been created to store, search and use multiple T/O simultaneously. Among them, the UMLS (Unified Medical Language System) [1], the EBI Ontology Lookup Service [3], the NCBO BioPortal [9] or the CISMeF HeTOP [6].

[1] Corresponding author: julien.grosjean@chu-rouen.fr

The NCBO project (National Center for Biomedical Ontology) of the university of Stanford and the CISMeF (Catalogue et Index des Sites Médicaux de langue Française, Rouen University Hospital, France) have invested considerable efforts in the development of tools and services based on T/O to assist health professionals to search resources on the Internet and to use T/O. Both groups have developed web portals, respectively BioPortal and HeTOP that offer various services to find, index, browse, visualise and annotate T/O. This article aims to compare BioPortal and HeTOP by defining objective comparison points based on their functionalities and features. Thus, it could be possible to define a strategy to merge both portals into one unique solution, offering the best services for human users and programmes.

1. Material & Methods

1.1. BioPortal - http://bioportal.bioontology.org

Developed by the NCBO, BioPortal is a repository of biomedical ontologies which hosts more than 350 ontologies in different formats [8], [10]. These ontologies are regularly updated by users and accessible via a web site for the humans and web-based services for programmes. BioPortal is a library of community ontologies [4] designed as a "one-stop shop" repository. Users have access to the ontologies with or without restriction (depending on the level of restriction set by the editor) and can access to editing, comment, rating and content adding operations.

1.2. HeTOP - http://www.hetop.org

HeTOP (Health Terminology/Ontology Portal) [6], is a T/O portal developed by the CISMeF team. It hosts more than 55 T/O in several languages. Most of the T/O are international or French national references such as MeSH, ICD-10, or CCAM. These T/O are regularly updated and are accessible via a web site and a web-based service. HeTOP has been designed as reference multi-terminology and multi-lingual portal [7] to help librarians, translators, students and health professionals to retrieve resources and knowledge across a high variety of complex health fields.

1.3. Criteria and Comparison Approach

The work of [4] has inspired us for establishing a criteria list for comparing BioPortal and HeTOP properties (technologies, methodologies, policies, target users, final purposes, features,...) but also from other similar portals. We categorised all the criteria in four groups: (i) Content, (ii) Functions & Tools, (iii) User Interface & Usability, (iv) Methods & Technologies.

2. Results

2.1. Content Comparison

The comparison of content reveals 3 important differences between BioPortal and HeTOP: (i) The volume of data is more important in BioPortal considering the T/O num-

bers or concept numbers (respectively 5,960,457 and 1,951,834). However, since T/O have various number of concepts and terms (terms are preferred terms plus synonyms, in any language) and since HeTOP is dealing with multilingual content, T/O numbers are not suited comparison indicators: the more relevant figures are the number of terms and the total number of relations. Indeed, BioPortal and HeTOP have around the same number of terms (about 6,600,000). Unfortunately, it is not possible to easily calculate the total number of relations in BioPortal. (ii) T/O formats and update frequencies are quite different in BioPortal compared to HeTOP. While BioPortal is focused on ontologies and takes advantage of standard formats and programmes, HeTOP is also hosting heterogeneous representation formats such as Microsoft Excel files, XML files or database dumps. A special work has to be done for every new T/O source: this can not be executed automatically and it implies expertise and development. (iii) One major difference between BioPortal and HeTOP is the expertise brought to every T/O. While BioPortal is automatically importing ontologies and does not change anything, except new automatic mappings and some manual users mappings, each HeTOP hosted T/O undergoes a series of process to leverage its content and meta-data (new translations, synonyms, definitions, relations, mappings, etc.).

2.2. Functions & Tools Comparison

Despite the similarity of basic tools for both portals, some details and tools differ slightly but have a direct consequence for the human users. (i) The BioPortal search engine only searches for exact terms in English (among preferred terms and synonyms) while HeTOP search engine is able to add wildcards to search for terms containing the query, in two languages simultaneously. This has a direct impact on how users can search terms; for instance, if one searches "myopathy" (in English) in the NCIT[4], BioPortal retrieves 5 terms whereas HeTOP 25; because wildcards are managed, HeTOP search engine is actually querying "*myopathy*" and retrieves terms such as "Cardiomyopathy".

2.3. User Interface & Usability Comparison

To compare the user experience using both portals, we compared two functions involving time responses. First, a comparison have been made between search engines performances in terms of response time and result numbers; we picked and performed 10 random queries on both search engines and we measured the user experience time (with the FireBug Mozilla Firefox plug-in) and noted the results. No options have been selected in both portals and the searches have been made among all T/O, in English and in French in HeTOP and only in English in BioPortal (no possible multilingual search). Results are 5.57 sec/19.7 (average response time/number of results) and 3.17 sec/359.4 respectively for BioPortal and HeTOP.
About multilingualism, within BioPortal, T/O in other languages than English are mostly available as "views" of the corresponding English T/O (e.g., the French MeSH is a view of the MeSH) but it is impossible to get the French term while browsing the English one and vice-versa. Within HeTOP, T/O are not language specific (e.g., the MeSH exists only once with the available translated terms). Therefore it is easy from the English term to

[4]National Cancer Institute thesaurus, edited by the National Cancer Institute: `http://ncit.nci.nih.gov/`

get to the French one and vice-versa. The whole user interface is internationalised and the search can be performed per language. Switching from one language to another is context sensitive.

2.4. Methods & Technical Comparison

While BioPortal is based on a RDF triple store data model, HeTOP has a meta-model for T/O which encapsulates specific T/O models into an Oracle 11g r2 relational database. Both portals are coupled to a web-based service (`http://www.bioontology.org/wiki/index.php/Resource_Index_REST_Web_Service_User_Guide`, `http://cispro.chu-rouen.fr/CISMeFhetopservice/`. About code license, APIs are open for BioPortal and proprietary for HeTOP.

3. Conclusion

As described above, BioPortal and HeTOP are sharing many features such as T/O browsing, tab representation and resources access tools. Both portals are valuable tools to support research projects. However, their policies and basic purposes are significantly different. While BioPortal is opened to the community T/O, HeTOP is focusing on reference T/O with experts interventions. Thus, BioPortal only accepts ontologies (at least files in ontology formats) whereas HeTOP provides access to non-ontological sources. BioPortal has been created as a "one-stop shop" repository and can be instantiated in other environments using the virtual appliance. On the contrary, HeTOP is a static server/web site acting such as a platform to help various kinds of users for different goals. Moreover, HeTOP is focusing on T/O content, working hard with experts and curators to leverage lexicons, mappings and other knowledge resources. This has a direct consequence on data volume and restriction policies related to download, versioning or updates.
Several criteria of this comparison study are focusing on usability: a portal dedicated to be used by experts or lay people has to be understandable and usable with less efforts and knowledge. A special work is made on HeTOP to deal with it: many T/O meta-data labels (attributes, relations, ...) are translated in several languages and well defined.
On the other hand, BioPortal is more open than HeTOP: users can upload ontologies and annotate concepts and they have more options. BioPortal proposes persistent concept URL thanks to the REST technology. However, HeTOP is more user-oriented with high quality and multilingual T/O content, it is also faster (search engine and navigation) and maybe more adapted to lay people and students which means that it is more reliable and more friendly to use on a daily basis. Unfortunately, no evaluation on user's satisfaction has been performed on HeTOP nor BioPortal; it is an ongoing work for HeTOP with a set of online questions.
This study helps to understand both portals philosophies and functionalities. Furthermore, it allows to point at BioPortal and HeTOP advantages and drawbacks. Despite the differences, we can assume that a merge is possible. The vast majority of policy points and functionalities could be kept in both approaches to create a single portal. The main difficulty would be the technical choices to merge a RDF data store and a relational database with many specificities. To tackle this, it would be possible to integrate the two data layers in a unique system coupled to a single API. This merge would have a great

cost but it would be a considerable benefit for the biomedical community and T/O users. This work is part of the SIFR (Semantic Indexing of French Biomedical Data Resources) research project[5], in collaboration betweem LIRMM, CISMeF and NCBO.

Acknowledgements

This work is issued from a collaboration between NCBO, CISMeF and the LIRMM. We used the knowledge of the partners, the scientific productions and the technical documentations of BioPortal and HeTOP to define comparison criteria. We thank the NCBO team for its cooperation on technical elements. This work was supported in part by the French National Research Agency under JCJC program, grant ANR-12-JS02-01001, as well as by University Montpellier 2, CNRS and IBC project.

References

[1] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* (2004) PMID: 14681409

[2] O. Bodenreider & R. Stevens, Bio-ontologies: Current Trends and Future Directions. *Briefings in Bioinformatics*, **7** (2006), 256–274

[3] R. Côté, F. Reisinger, L. Martens, H. Barsnes, J.A. Vizcaino, H. Hermjakob, The Ontology Lookup Service: bigger and better. . *Nucleic Acids Res.* (2010) PMID: 20460452

[4] M. d'Aquin & N. F. Noy, Where to Publish and Find Ontologies? A Survey of Ontology Libraries. *journal of Web Semantics*, **11** (2012), 96–111.

[5] D. L Rubin, N. H. Shah & N. F. Noy, Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics*, **9**(2) (2008), 75–90.

[6] J. Grosjean, T. Merabti,B. Dahamna, I. Kergourlay, B. Thirion, L. F. Soualmia & S. J. Darmoni, Health Multi-Terminology Portal: a semantics added-value for patient safety. *Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety volume **166** of Studies in Health Technology and Informatics* (2011), p. 129–138.

[7] J. Grosjean, T. Merabti, N. Griffon, B. Dahamna, L. Soualmia & S. J. Darmoni, Multi-terminology cross-lingual model to create the Health Terminology/Ontology Portal. *American Medical Informatics Association Annual Symposium*, p. 1753 (2012), Chicago, USA.

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

[9] D. L. Rubin, S. E. Lewis, C. J. Mungall, S. Misra, M. Westerfield, M. Ashburner, I. Sim, C. G. Hute, H. Solbrig, M.-A. Storey, B. Smith, J. Day-Richter, N. F. Noy & M. A. Musen, National Center for Biomedical Ontology: Advancing Biomedicine through Structured Organization of Scientific Knowledge. *OMICS A Journal of Integrative Biology*, **10**(2) (2006), 185–198.

[10] P. L Whetzel., N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache & M. 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) (2011), 541–545.

[5] `http://www.lirmm.fr/sifr/`