Comparing BioPortal and HeTOP: towards a unique biomedical ontology portal?

Julien Grosjean, Lina F. Soualmia, Khedidja Bouarech, Clement Jonquet, Stefan J. Darmoni

Comparing BioPortal and HeTOP: towards a unique biomedical ontology portal ?

Julien Grosjean[1], Lina F. Soualmia[1,2], Khedidja Bouarech[3], Clément Jonquet[3], and Stefan J. Darmoni[1,2]

[1]CISMeF & TIBS, LITIS EA4108, Rouen University Hospital, France
[2]LIMICS, INSERM UMR 1142, Paris, France
[3]LIRMM, Université Montpellier 2 & CNRS, Montpellier, France
{julien.grosjean,lina.soualmia,stefan.darmoni}@chu-rouen.fr
{bouarech,jonquet}@lirmm.fr

Abstract. The volume of data in the biomedical field constantly grows. The vast majority of information retrieval systems are based on controlled vocabularies and most specifically on terminologies or ontologies (T/O). These classification structures allow indexing, coding, annotating various types of documents. In Health, many T/O have been created for different purposes and it became a problem to find specific concepts in the multitude of nomenclatures. The NCBO (National Center for Biomedical Ontologies, Stanford University) BioPortal project[1] and the CISMeF (Catalogue et Index des Sites Médicaux de langue Française, Rouen University Hospital) HeTOP[2] portals have been developed to tackle this issue. While both portals are designed to store and make T/O available to the community, they are also very different mainly because of their basic purposes. The present work consists in comparing both portals and in answering the following question: is it possible to merge BioPortal and HeTOP into one unique solution to manage T/O ?

1 Introduction

The biomedical data constantly grows, specially with new technologies and the Internet media. Therefore, it becomes mandatory to index/annotate it with controlled and structured vocabularies in order to store it and to retrieve it, with intelligent methods. A key aspect in addressing semantic interoperability for life sciences data is the use of terminologies or ontologies (T/O) as a common denominator to structure data and make them interoperable [2] [15]. T/O are classifications that describe the knowledge of a specific domain with *concepts* and *relations* between them. Ontologies are more complex than terminologies because they can define *rules* and *functions* to infer and structure the knowledge.
Since T/O are commonly used in Information Systems and specifically in Health, it is a real challenge to manage and consult them into a unique gateway. Indeed, many T/O

[1] http://bioportal.bioontology.org/
[2] http://www.hetop.eu

have been created in the last decade for different purposes: indexing/annotating documents, organizing the knowledge, inferring facts, etc. Some T/O (e.g. MeSH[3], ICD-10[4], CCAM[5]) are commonly used on a daily basis in hospitals or in research labs and are very useful for information retrieval. T/O are not always well structured or defined because a lack of standardization or formatting. Moreover, the semantics and syntactic interoperability between T/O is a great challenge to allow interconnection between systems and knowledge.
Several tools have been created to store, search and use multiple T/O in the same time: among them, the UMLS (Unified Medical Language System) [1], the EBI Ontology Lookup Service [3], the NCBO BioPortal [14] or the CISMeF HeTOP [7].
The NCBO project (National Center for Biomedical Ontology) of the university of Stanford and the CISMeF team (Catalogue et Index des Sites Médicaux de langue Française, Rouen University Hospital, France) have invested considerable effort in the development of tools and services based on T/O to assist health professionals to search documents on the Web and to use T/O. Both groups have developed web portals, respectively BioPortal and HeTOP that offer various services to find, index, browse, visualize and annotate T/O. Despite their specificities, the two portals share an overall common objective which is to provide to the biomedical community a unique interface to explore Health T/O and resources.
Both portals can be very similar at first sight but many functionalities and specificities make them very different in terms of use, focus and technologies. Moreover, for the same T/O included in both portals, the content may differ: this can be an issue for the end users.
This article aims to compare BioPortal and HeTOP 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 end users and machines.

2 Material & Methods

2.1 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 [12], [16]. These ontologies are regularly updated by users and accessible via a web site for the humans or Web services for machines. 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 edition, comment, rating and content addition operations. The NCBO also develops tools and

[3] Medical Subject Headings, edited by the National Library of Medicine: `http://www.nlm.nih.gov/mesh/`
[4] International Classification of Diseases 10th revision, edited by the World Health Organization: `http://www.who.int/classifications/icd/en/`
[5] Classification Commune des Actes Médicaux, edited by the French National Insurance system: `http://www.ameli.fr/accueil-de-la-ccam/index.php`

services to use and exploit ontologies in BioPortal: (i) the **Annotator**[9] performs semantic annotation of biomedical documents, identifying ontology concepts present in the text, (ii) the **Resource Index**[11] uses these annotations to index various biomedical resources and facilitate information retrieval and semantic search, (iii) the **NCBO Recommender** [10] which helps to identify which ontology(ies) is mostly suitable for a specific data corpus.

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

HeTOP (Health Terminology/Ontology Portal) [7], 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 service. HeTOP has been designed as reference multi-terminology and multi-lingual portal [8] to help librarians, translators, students and medical professionals to retrieve resources and knowledge across a high variety of complex medical fields. Like NCBO, CISMeF is developing additional tools and services to use and exploit T/O in HeTOP: (i) the **ECMT**[6] performs document indexing, identifying T/O concepts present in a text [13], (ii) **InfoRoute**[7], which is an info-button service allowing to access many portals from a simple query [6], performing semantic expansion based on the T/O, (iii) **MT@HeTOP**[8] is a service also based on the HeTOP T/O to translate and map terms, (iv) **Doc'CISMeF**[9] [5] which is a search engine for quality web resources about Health manually or automatically indexed by CISMeF curators with the MeSH and other reference T/O.

2.3 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 categorized all the criteria in five groups: (i) Content, (ii) Policy and community, (iii) Functions & Tools, (iv) User Interface & usability, (v) Methods & Technologies.

3 Results

3.1 Content comparison

The comparison of content shown in the Table 1 reveals 3 important differences between BioPortal and HeTOP: (i) The volume of data is more important in BioPortal considering T/O numbers or concept numbers (ratio are respectively 6 and 3). However, since T/O have various number of concepts and terms (terms are preferred terms

[6] http://ecmt.chu-rouen.fr/
[7] http://inforoute.chu-rouen.fr/
[8] http://cispro.chu-rouen.fr/MT_EHTOP/
[9] http://doccismef.chu-rouen.fr

plus synonyms, in any language) and since HeTOP is dealing with multilingual content, T/O numbers are not good comparison indicators: the more relevant figures are the term number and the total relations number. 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 relation number 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 programs, 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. Those processes are automatic, semi-automatic or manual operations and aim to correct or add content: new translations, synonyms, definitions, relations, mappings, etc. Furthermore, since each T/O has a specific model (concept types, constraints, rules, etc.), additional work on models handling can help user to understand how to use T/O. In BioPortal, each concept is an instance of a unique class which does not respect the T/O initial model (specially for terminologies which are based on concept types specificities).

3.2 Policies & community aspects comparison

BioPortal and HeTOP have community members and strong policies about which T/O to host and how they are managed. The Table 2 shows several criteria to compare both approaches. Some points are about T/O policies and others about user interaction.

3.3 Functions & tools comparison

Despite the similarity of basic tools for both portals, some details and tools differ slightly (Table 3) but has a direct consequence for the end users. (i) The BioPortal search engine only searches exact terms in English (among preferred terms and synonyms) while HeTOP search engine is able to add wildcards to search terms containing the query, in two languages at the same time. This has a direct impact on how users can search terms; for example, if one searches “myopathy” (in English) in the NCIT[10], BioPortal retrieves 5 terms while HeTOP matches 25 terms; because wildcard is managed, HeTOP search engine is actually querying “*myopathy*” and retrieves terms such as “Cardiomyopathy”. (ii) Some tools are available in BioPortal but not in HeTOP and *vice versa.*

3.4 User Interface & usability comparison

To compare user experience using both portals, we compared two functions involving time responses (Table 4). First, a comparison have been made between search engines

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

Property	BioPortal	HeTOP
T/O metrics	n(T/O) = 368 n(concepts) = 5 960 457 n(terms) = 6 600 000 n(relations) = ? n(mappings) = 5 000 000	n(T/O) = 56 n(concepts) = 1 951 834 n(terms) = 6 636 654 n(relations) = 8 023 181 n(mappings) = 1 340 855
T/O source	From the UMLS From the OBO Foundry Directly from registered users	From the UMLS From different official sources (see HeTOP terminologies list for details) From other research teams
T/O formats	OWL, OBO, RDF, RRF	No automatic imports
Handling T/O initial model	No	Yes
T/O documentation	Documentation available on specific pages	Depending on the T/O, on a single page
T/O update	Automatic for OBO Foundry and UMLS. Available form for custom T/O	Depending on the T/O and its update frequency, some are automatic (4)
T/O versioning	Yes, with a persistent `ontology_version_id`	No, only the lastest version available
T/O languages	Some "views" of English T/O	n = 23, coverage depending on the T/O
T/O added content	User annotations New automatic mappings	Opened to the community, reviewed by curators New content from automatic/manual tools
T/O organization	T/O are categorized in groups and categories	N/A
T/O quality leverage	No	Yes, on the vast majority of T/O
T/O interoperability	Automatic and manual mappings	Automatic and manual mappings

Table 1. Content comparison between BioPortal and HeTOP (dec. 2013)

Property	BioPortal	HeTOP
T/O choices	UMLS ontologies and community ontologies	Reference T/O and projects T/O
T/O access	All ontologies are public except those set to private by users	22 T/O are freely available, 16 more when registered and the others are restrited to research projects users (e.g. industrial partners T/O)
T/O edition	No edition available but one can add notes to concepts	No edition available (except for curators) but one can suggest translations or synonyms for specific T/O. CISMeF team curators can valid them.
Downloadable content	Yes for public ontologies and public mappings	No
User options	Users can own ontologies and manage versions. Projects creation is possible.	A query history and a selection history are available.
	The user interface is mostly the same for every ontologies.	A special effort is brought to adapt the user interface to the T/O content. Especially for metadata labels (attributes, relations, etc.) and for a multilingual use.
User documentation	Yes, in a Help tab	No
Registered users	n = 3 017	n = 1 518
Traffic	17 500 hits/day	15 000 hits/day (500 users/day)
Research projects	215	HeTOP is a support a research projects and integrates some productions such as VCM icons, mappings, interface terminologies, etc.

Table 2. Policies & community aspects comparison between BioPortal and HeTOP (dec. 2013)

Property	BioPortal	HeTOP
search engine	Exact match terms (concepts preferred terms and synonyms) and advanced options	Exact and partial match terms (concepts preferred terms, synonyms, ids, definitions,...), and advanced options
results display	In a dedicated page, organized by T/O	Integrated into a unique view, limited to 500 results and organized by T/O and concept types
concept view	In a description tab	In a description tab
hierarchy view	Always visible (if provided), no poly-hierarchy	In a dedicated tab, with poly-hierarchy
relations view	Flat, uncommented relation types	Organized and understandable and multilingual relations types
mappings view	In a dedicated tab	Embedded in the relations tab
resources access	Not accessible from the concept page	Dedicated tab to access the InfoRoute tool
cross-lingual navigation	No	Yes by clicking on the flags

Table 3. Functions and tools comparison between BioPortal and HeTOP (dec. 2013)

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 result number (random queries are: “heart failure”, “asthma”, “lung disease”, “pregnancy hypertension”, “pneumonia”, “childhood”, “cell junction”, “surgery”, “gene methylation” and “egg allergy”). 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 multilingual search). We also performed the same queries with a wildcard option (not available in BioPortal).

The third comparison of response times concerns concept view page access times. A random selection of 20 concept identifiers has been performed and we measured the time needed to open the given concept view with the same method that the previous comparison.

About multilingualism, within BioPortal, T/O in another language 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 translated terms available). Therefore it is easy from the English term to get to the French one and vice-versa. The whole user interface is internationalized and search can be performed per language. Switching from one language to another is context sensitive.

3.5 Methods & technologies comparison

In the Table 5, we compared the methods and technologies used in both portals using documentations and scientific productions.

Property	BioPortal	HeTOP
Multilingual display	No	Yes
Contextual links to other portals/browsers	No	Yes, to several other portals (including BioPortal, MeSH, LOINC, OMIM browsers)
Navigation display	Hierarchy and concept links	Hierarchies, concept links and cross-lingual navigation
Search engine results (execution times/number)	5.57 sec. / 19.7	3.17 sec. / 359.4
Search engine results with wildcard	Not applicable on search tab	4.46 sec / 501.3
Concept page access time	4.1 sec.	less than 1 second

Table 4. User Interface & usability comparison between BioPortal and HeTOP (dec. 2013)

Property	BioPortal	HeTOP
Data model	RDF (ontologies are loaded in a triple store)	Meta-model for T/O which encapsulate specific T/O models
Database implementation	4Store triple store	Oracle 11g r2, with partionning & domain indexes options
Web services	Accessible in REST (`http://www.bioontology.org/wiki/index.php/Resource_Index_REST_Web_Service_User_Guide`)	Accessible in SOAP and REST (`http://cispro.chu-rouen.fr/CISMeFhetopservice/`)
Other API	CTS2 (OMG Standard)	Yes but not accessible
Technical documentation	NCBO wiki: full support documentation about technologies and developments	Yes but not accessible
Web site technologies	Ruby on Rails, Javascript, Spring/Hibernate, Protégé, LexGrid, Rainbow	Vaadin, JEE, Infinispan, CXF
License	open source	proprietary
Reusability	NCBO Virtual appliance is available to install and run its own version of BioPortal locally	N/A

Table 5. Methods & technologies comparison between BioPortal and HeTOP (dec. 2013)

4 Discussion

As described in the result tables, BioPortal and HeTOP are sharing many features such as T/O browsing, tab representation or 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 ontology formats) whereas HeTOP provides access to non-ontological sources. BioPortal has been created as a "one-stop shop" repository and can be instantiate in other environments using the virtual appliance. On the contrary, HeTOP is a static server/web site acting like 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 another hand, BioPortal is more open than HeTOP: users can upload ontologies and annotate concepts and they have more users 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; this work is currently in preparation for HeTOP with a set of online questions.
Despite those 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.

5 Conclusion

In this study, BioPortal and HeTOP have been compared, based on many criteria. This helps to understand both portals philosophies and functionalities. Furthermore, it allows to point at BioPortal and HeTOP advantages and drawbacks. As a conclusion, a merge of those portals can be possible to create a better tool for end-users and machines. However, such developments would be a hard task to carry out and the inconsistency of T/O content (depending on the source and CISMeF experts capacity to leverage specific T/O) would be huge challenge to meet. This work is part of the SIFR (Semantic

Indexing of French Biomedical Data Resources) research project[11], 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. BODENREIDER O. (2004). *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic Acids Res. PMID: 14681409
2. BODENREIDER O. & STEVENS R. (2006). *Bio-ontologies: Current Trends and Future Directions Briefings in Bioinformatics*, **7**, 256–274
3. CÔTÉ R., REISINGER F., MARTENS L., BARSNES H., VIZCAINO J. A., HERMJAKOB H. (2010). *The Ontology Lookup Service: bigger and better.* Nucleic Acids Res. PMID: 20460452
4. D'AQUIN M. & NOY N. F. (2012). Where to Publish and Find Ontologies? A Survey of Ontology Libraries. *journal of Web Semantics*, **11**, 96–111.
5. DARMONI S. J., THIRION B., LEROY J.-P., DOUY M., LACOSTE B., GODARD C., RIGOLLE I., BRISOU M., VIDEAU S., GOUPY E., PIOT J., QUÉRÉ M., OUAZIR S & ABDULRAB H. (2001). Doc'CISMEF: a search tool based on "encapsulated" MeSH thesaurus. In *10th World Congress on Medical Informatics*, p. 314–318, London, UK.
6. DARMONI S. J., PEREIRA S., NÉVÉOL A., MASSARI P., DAHAMNA B., LETORD C., KERDELHUÉ G., PIOT J., DERVILLE A. & THIRION B. (2008). French Infobutton: an academic and business perspective. In *American Medical Informatics Association Annual Symposium*, p. 920, Washington DC, USA.
7. GROSJEAN J., MERABTI T., DAHAMNA B., KERGOURLAY I., THIRION B., SOUALMIA L. F. & DARMONI S. J. (2011). Health Multi-Terminology Portal: a semantics added-value for patient safety. In *Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety*, volume 166 of *Studies in Health Technology and Informatics*, p. 129–138.
8. GROSJEAN J., MERABTI T., GRIFFON N., DAHAMNA B., SOUALMIA L. & DARMONI S. J. (2012). Multi-terminology cross-lingual model to create the Health Terminology/Ontology Portal. In *American Medical Informatics Association Annual Symposium*, p. 1753, Chicago, USA.
9. JONQUET C., SHAH N. H. & MUSEN M. A. (2009). The Open Biomedical Annotator. In *American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09*, p. 56–60, San Francisco, USA.
10. JONQUET C., MUSEN M. A. & SHAH N. H. (2010). Building a Biomedical Ontology Recommender Web Service. In *Biomedical Semantics*, **1**

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

11. JONQUET C., LEPENDU P., FALCONER S., COULET A., NOY N. F., MUSEN M. A. & SHAH N. H. (2011). NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources. In *Web Semantics*, p. 316–324.
12. NOY N. F., SHAH N. H., WHETZEL P. L., DAI B., DORF M., GRIFFITH N. B., JONQUET C., RUBIN D. L., STOREY M.-A., CHUTE C. G. & MUSEN M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, **37** (web server), 170–173.
13. PEREIRA S., NÉVÉOL A., KERDELHUÉ G., SERROT E., JOUBERT M. & DARMONI S. J. (2008). Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue. In *American Medical Informatics Association Annual Symposium*, p. 586–590, Washington DC, USA.
14. RUBIN D. L., LEWIS S. E., MUNGALL C. J., MISRA S., WESTERFIELD M., ASHBURNER M., SIM I., HUTE C. G., SOLBRIG H., STOREY M.-A., SMITH B., DAY-RICHTER J., NOY N. F. & MUSEN M. A. (2006). National Center for Biomedical Ontology: Advancing Biomedicine through Structured Organization of Scientific Knowledge. *OMICS A Journal of Integrative Biology*, **10**(2), 185–198.
15. RUBIN D. L., SHAH N. H. & NOY N. F. (2008). Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics*, **9**(2), 75–90.
16. WHETZEL P. L., NOY N. F., SHAH N. H., ALEXANDER P. R., NYULAS C., TUDORACHE T. & MUSEN M. A. (2011). 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.