Visual presentation of mappings between biomedical ontologies

Simon Kocbek[1], Jean-Luc Perret[2], Jin-Dong Kim[1]

[1]Database Center for Life Science, Research Organization of Information and Systems, Tokyo, Japan
{simon, jdkim}@dbcls.rois.ac.jp
[2]Novartis Animal Health, Centre de Recherches, St-Aubin, Switzerland
jean-luc.perret@novartis.com

Abstract
Ontology mapping focuses on finding correspondences between concepts from different ontologies. While the amount of available ontologies is increasing, also the number of mappings between them is getting higher. Visualization techniques can be used to help researchers in forming a picture of this information. In the paper we present a visual presentation of mappings between BioPortal ontologies. We present results in the form of a graph where identified communities of tightly connected ontologies are shown.We use metrics such as Betweenness Centrality and Community Detection.

Keywords: ontology, ontology mapping, visualization

1 Introduction

Creating mappings among ontologies by identifying concepts with similar meanings is a critical step in integrating data and applications that use different ontologies [1].

BioPortal is a web portal developed by The National Center for Biomedical Ontology (NCBO) that provides access to a library of biomedical ontologies and terminologies [2]. The ontologies are published by several different groups (e.g., the OBO library, and the Proteomics Standards Initiative) and grouped in 40 categories (e.g., Anatomy, Cell, and Health). Concepts in BioPortal ontologies often overlap and information about mappings between ontologies is available. Two ontologies are mapped when they contain at least one pair of concepts with similar meaning (i.e., the concept *c1* from the ontology *O1* has similar meaning as the concept *c2* from the ontology *O2)*. Our analysis showed more than 30,000 BioPortal mappings . It is hard for humans to understand and form a picture of so many connected ontologies. In addition, ontology mappings are often considered in activities such as data integration, ontology ranking and recommendation [3], or ontology reuse. The later is a also one of the interests in our group where we are developing the OntoFinder/Factory

tool[1] which uses BioPortal ontologies. As a result, we believe that it would be useful to provide visualization of mappings between biomedical ontologies from BioPortal in a form of a graph where each node would present an ontology and edges would present mappings between the ontologies. This kind of graph can provide a macro view of related biomedical ontologies for researchers who are interested in them.

In the next section we describe our visual analysis of mappings between BioPortal ontologies. We conclude the paper in Section 3 where we also provide guidelines for future work.

2 Visualization of BioPortal Mapping Data

For each BioPortal ontology, we collected the following data through the BioPortal web services: the ontology's full name (e.g., Gene Ontology), the ontology's name abbreviation (e.g., GO), status of the ontology (e.g., production), the number of classes in the ontology, and the number of mappings from/to the ontology. Initially, the data for more than 320 ontologies was collected. However, this number was reduced to 284 since we filtered out ontologies that: (1) have the *retired* or *alpha* status, (2) contain the keyword *test* in their name, and (3) are labelled as *restricted* or *private.*

After collecting the data, we identified 30,560 mappings between 254 ontologies (i.e., each of these ontologies contained at least one concept mapped to another ontology). The remaining 30 ontologies had no reference to other ontologies. The majority of the identified mappings were bidirectional and symmetric. This means that when an ontology *O1* referenced an ontology *O2* with *x* number of concepts, then also *O2* referenced *O1* with the same number of concepts. Only 218 asymmetric ontology pairs were found in our data.

We used Gephi (i.e., an open source software for graph analysis and visualization) [4] to visualize our data. Gephi provides layout algorithms to draw large graphs as well as node and edge filtering capabilities. In addition, a number of graph and node properties can be calculated with Gephi. For the scope of this paper the following two main features were used:

- Modularity Analysis (or Community Detection) is a measure of structure in graphs. Graphs with high modularity have separate communities of densely connected nodes inside the communities and sparse connection across communities [5].
- Betweenness centrality [6] is a measure of the frequency of occurrence of a particular node in all shortest paths between any two nodes.

Figure 1 illustrates our visual representation of BioPortal ontology mappings with Gephi. Each node represents an ontology and an edge represents a mapping between two ontologies. Sadly, due to high number of mappings between BioPortal ontologies, not all edges can be shown. Node labels represent ontology name abbreviations (please refer to the BioPortal webpage for full names). Edge thickness is proportional to number of related concepts between two ontologies where a thicker line represents

[1] http://ontofinder.dbcls.jp/

a higher number of related concepts. The node size is proportional to the betweenness centrality metrics. The node colour represents membership to one of the communities detected by modularity analysis.

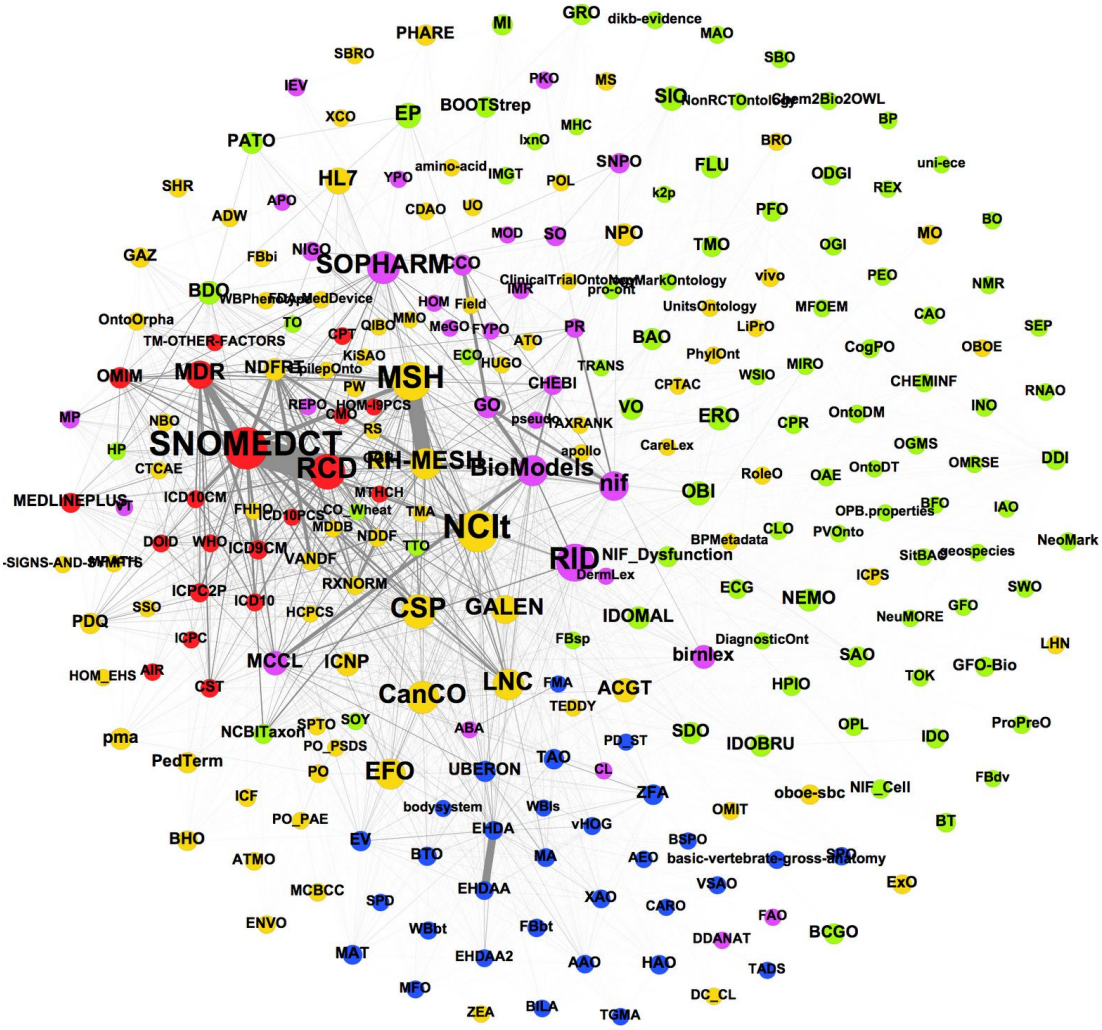

Figure 1: Graph of BioPortal ontology mappings (254 ontologies with at least one mapping, 30,560 edges).

We obtained a graph density of 0.38 and a modularity of 0.346 which indicate a relatively homogeneous graph with little structure. Nevertheless, the community de-

tection revealed five communities of interconnected nodes. Two of these communities, clearly discriminate communities of ontologies related to anatomy and clinical terms. These two communities also relate to BioPortal's category classification since majority of ontologies in each community belong to the same or related categories. The three other identified communities are more heterogeneous. The graph also shows the top three ontologies in term of betweenness centrality are SNOMEDCT, NCIt and MSH.

3 Conclusion

This work was our first attempt to visualize mapping data from BioPortal and as such opens additional research questions and opportunities. Our graph implies that clinical terms and anatomy related ontologies seem to map their terms much more than ontologies in other topics. It is difficult to interpret this observation and could be very much related to the way ontologies in the different domains were built and the needs of applications in fields like pathology. In fields where the number of mappings is large the present analysis may be useful to learn about ontologies in a particular context, especially if one of the ontologies of interest is known. In the future, we would like to analyse internal structure of ontologies and see if there is any connection between the most important terms and identified communities. Since there are many plugins available for Gephi, we would also like to experiment with different add-ons and see whether we can visualize edges in a better way. In addition, a large portion of the mappings in BioPortal are automatically calculated. It would be interesting to see how the visualization changes while methods for automatically ontology mapping and alignment improve.

Acknowledgments:
The work in this paper was inspired at the BioHackathon 2012 event.

References

1. Ghazvinian, A., Noy, N.F., Musen, M.A.: Creating mappings for ontologies in biomedicine: simple methods work. AMIA Annu. Symp. proc. 2009:198–202, (2009)
2. Whetzel, P.L., Noy, N.F., Shah, N.H., et al.: BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res;39:W541–5, (2011)
3. Jonquet, C., Musen, M.A., Shah, N.H.: Building a biomedical ontology recommender web service. J. Biomed. Sem.;1 Suppl 1:S1, (2010)
4. Bastian, M., Heymann, S., Jacomy, M. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. (2009)
5. Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebre Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 10, P1000, (2008)
6. Freeman, L.: "A set of measures of centrality based upon betweenness". Sociometry 40: 35–41. (1977)