Visualizing ontology mappings to help ontology engineers identify relevant ontologies for their reuse

Simon Kocbek[1,*] Jin-Dong Kim[1] Jean-Luc Perret[2] and Patricia L. Whetzel[3]

[1] Database Center for Life Science, Research Organization of Information and Systems, Tokyo, Japan
[2] Novartis Animal Health, Centre de Recherches, St-Aubin, Switzerland
[3] Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, USA

ABSTRACT

The importance of ontologies in biomedicine is increasing in the areas such as the standardization of terminology, the verification of data consistency, and the integration of heterogeneous biomedical databases. New ontologies are being built and added to repositories such as BioPortal. These ontologies represent a large network of biomedical concepts where a single ontology connects a group of closely related concepts. When ontology engineers build new ontologies they often search for existing ontologies to avoid redundancy of concepts. When selecting existing ontologies, engineers consider different factors such as ontology domain, the size of the ontology, and also the relations between ontologies and their concepts. In this paper we present a graph that aims to visualize mappings of all BioPortal ontologies. We believe that this graph can help ontology engineers in deciding which ontology to use when selecting existing concepts for building new ontologies.

1 INTRODUCTION

Experts with different backgrounds may define biomedical ontologies from different perspectives. One possible perspective is to describe a set of biomedical ontologies as a web of biomedical concepts where a single biomedical ontology represents a group of closely related concepts. Each ontology has different number of concepts and relations, describes specific domain, and is built by specific experts. Therefore, biomedical ontologies represent a large network with specific properties. One of the important properties of this network is connections between ontologies. When ontologists build new ontologies they often search for existing ontologies to avoid redundancy of concepts as recommended by the OBO Foundry principles (Smith et al., 2005). Therefore, it is important to understand relations between existing concepts. Existing concepts with similar meanings are identified with so called ontology mappings. Identifying ontology mappings is a critical step in integrating data and applications that use different ontologies (Ghazvinian et al., 2009).

Currently, there are 330 biomedical ontologies available through BioPortal, i.e., a library of biomedical ontologies and terminologies which are accessible via the NCBO Web services (Whetzel et al., 2011). The use of BioPortal ontologies is growing and new biomedical ontologies are being added to the repository. As the number of biomedical ontologies grows, the number of mappings between these ontologies also increases. Presenting this information to ontology users and engineers can be a challenging task and visualization techniques can offer an effective solution. There are several tools that are capable of visualizing ontology mappings, however, visualizing mappings between more than two ontologies is not common.

Our group is developing OntoFinder - a web service for searching relevant ontologies and building new ontologies based on existing ones[1]. Currently, OntoFinder ranks existing ontologies based on lexical matching of terms. However, very often this is not enough and we wish to extend this functionality. Therefore, as a part of OntoFinder, we are also developing a tool for visual representation of mappings between ontologies. To help ontologists decide which ontology to use, we generated a graph of all BioPortal mappings. This graph identifies densely connected communities of ontologies, and it visualizes so called "bridging" ontologies, i.e., ontologies that connect many other ontologies/communities.

In this paper we present our visualization and discuss its changes at two different time points. In the next section we review related work. Section 3 describes the BioPortal data. In Section 4 we describe and analyze our visualization techniques and we conclude the paper in Section 5.

2 RELATED WORK

Visualizing ontology mappings can be categorized as a part of ontology visualization techniques. These techniques are often used to visualize ontology structures (i.e., ontology concepts and relations of a single ontology). For example, there are several plugins available for Protégé, i.e., a popular ontology editor and knowledge-based framework (Grosso et al., 2004). The OntoViz plugin[2] is useful for graphically representing small ontologies or parts of ontologies. The Jambalaya plugin (Storey et al., 2001) is another add-on and it uses TreeMap to visualize more complex ontologies. Authors of the NavigOWL tool[3] emphasize the tool's appealing

---

* To whom correspondence should be addressed: simon@dbcls.rois.ac.jp

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

[2] Available at http://protegewiki.stanford.edu/wiki/OntoViz

[3] Available at http://protegewiki.stanford.edu/wiki/NavigOWL

graph layouts that can be applied over the semantic net in order to understand ontology structure.

Besides visualizing ontology structures, some of the tools also support the visualization of mappings between ontolgies. Some examples are OLA - OWL Lite Alignemnt (Euzenat et al., 2004), Alviz (Lanzenberger & Sampson, 2006), CogZ (Falconer & Storey, 2007) and Optima (Kolli & Doshi, 2008). However, these tools are only capable of visualizing mappings between pairs of ontologies. Two ontologies are usually presented side by side (e.g., two graph or two trees). The mapped concepts are then connected with a line (Fig. 1), have the same color, or other patterns are used. The BioMixer tool (Voyloshnikova et al., 2012), on the other hand, supports visualization of mappings between more than two ontologies. It offers the following three different mapping views that differ in their level of detail: 1) the mapping overview visualization, 2) the mapping matrix visualization, and 3) the detailed mapping graph. To our knowledge BioMixer is currently the only tool that is capable of visualizing mappings between more than two ontologies.
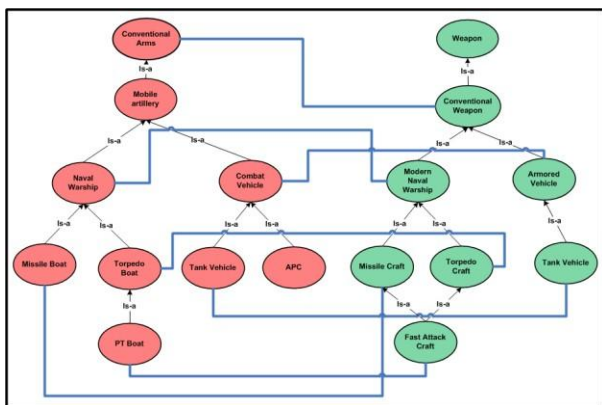

Fig. 1. An example of visualizing mapppings between two ontologies (Optima). Two mapped concepts are connected with a blue line.

Besides the above tools for visualizing mappings, Ghazvinian (Ghazvinian et al., 2009) produced graphs of large subset of biomedical ontologies and their mappings to provide answers to questions such as:

- To what degree are the domains covered by different ontologies connected?
- If you are new to a domain, what are the important or representative ontologies with good coverage?, and
- If you want to build domain-specific tools for creating ontology mappings, what are good ontologies to use for background knowledge?

Ghazvinian's work is similar to our work. However, Ghazvinian focused only on visualizing one type of mappings, while we visualize all three types that are supported in BioPortal (we discuss more about these types in the next section). In addition, our data contains larger number of ontologies and mappings between these ontologies.

3 DATA CHARACTERISTICS

OntoFinder uses BioPortal[4] for searching ontologies. 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. At the time of writing this paper, BioPortal contained 330 ontologies. The ontologies differ in several features such as number of the concepts, number of relations between the concepts, and types of concepts. The ontologies are published by several different groups (e.g., the OBO library, and the Proteomics Standards Initiative), and are grouped in 40 categories (e.g., Anatomy, Cell, and Health).

The following three mapping sources are available in BioPortal:

- NCBO, which uses the LOOM algorithm (Ghazvinian et al., 2009) to automatically calculate mappings. LOOM can identify mappings for terms with close lexical match or mappings for terms with the same URI from different ontologies. Ghazvinian defines mappings as similarity mappings, where two concepts/classes from different ontologies are similar if the meaning of one concept/class is similar or identical to the meaning of the other concept/class. LOOM also uses synonyms when searching for similar concepts.

- Unified Medical Language System (UMLS) links terms with the same UMLS concept unique identifier (CUI) or mappings between terms from the UMLS MRMAP.RRF data.

- Open Biological and Biomedical Ontologies (OBO) represents mappings between ontology terms related by an OBO xref property.

The mappings can be of a different type (e.g., identical, related, or close match). Detailed information about these types is available at the BioPortal's mapping WikiPage. Information about ontology mappings can be collected through a RESTful web service API or a SPARQL endpoint. The following are some mapping properties that can be gathered:

- type of the mapping (automatic/manual),
- information about the source ontology,
- information about the target ontology,

---

[4] Available at http://bioportal.bioontology.org/

- information about source of the mapping (algorithm, authorship, contact information, etc.), and

- direction of the mapping.

4 GRAPH OF BIOPORTAL ONTOLOGY MAPPPINGS

4.1 Visualization and statistics

The goal of the visualization techniques described in this section is to effectively visualize all types of mappings (described in Section 3) of all BioPortal ontologies and to offer a summary of a network of densely connected ontologies. In addition, we wish to recognize "bridging" ontologies, i.e., ontologies that connect many other ontologies/communities. We do not focus on visualizing individual mappings between concepts.

For creating the graph, 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), version of the ontology (e.g., alpha), and the number of mappings from/to the ontology. Initially all BioPortal ontologies were considered. However, in our final analysis we ignored ontologies that:

- have the retired or alpha status,

- contain the keyword test in their full name, and

- are labeled as restricted or private.

We used Gephi [4], an open source tool for graph analysis and visualization, to visualize the data. Gephi was chosen because it provides layout algorithms to draw large graphs and offers node and edge filtering capabilities. In addition, a number of graph and node properties can be calculated. The following three main features of Gephi are important for our work:

- 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 (Blondel *et al.*, 2008). With this feature we classified BioPortal ontologies into groups of highly related ontologies.

- Betweenness centrality [6] is a measure of the frequency of occurrence of a particular node in all shortest paths between any two nodes. The feature was used for identifying "bridging" ontologies.

- Graph Density measures how close the network is to complete. A complete network has all possible edges and density equal to 1. This feature was used to analyze how interconnected ontologies are.

Since BioPortal data often changes (e.g., new ontologies/mappings are defined), we decided to perform visualization of BioPortal data with Gephi through different time points. All of our visualizations are available on our project's Wiki page[5]. We are planning to upload new visualizations at least every few months or when major changes are going to be made to the BioPortal data. Currently two visualizations are available.

The first visualization was created based on the data that was collected in October 2012 (the data contains 284 ontologies). The results were presented as a graph in a poster at the SWAT4LS workshop (Kocbek et al., 2012). The graph contained 30,560 mappings between 254 ontologies. This means that each of these ontologies contained at least one concept mapped to a concept in another ontology. The remaining 30 ontologies had no reference to other ontologies. The majority of the identified mappings were bidirectional (i.e., when a concept $c1$ referred to a concept $c2$ then also the concept $c2$ referred to the concept $c1$). In addition, the majority of the mappings were symmetric (i.e., when an ontology $O1$ referenced an ontology $O2$ with x number of concepts, then also $O2$ referenced $O1$ with the same number of concepts). We count one bidirectional mapping as two mappings. Only 218 asymmetric ontology mapping pairs were found in our data. We obtained a graph density of 0.380 and a modularity of 0.346 which indicate a relatively homogeneous graph with little structure. Nevertheless, the community detection revealed five communities of interconnected nodes which, for two of these communities, clearly discriminate communities of ontologies related to anatomy and clinical terms. The three other identified communities are more heterogeneous.

In the second visualization we illustrate mappings from BioPortal data gathered in February 2013. This time we collected data for 294 ontologies, where 268 ontologies contained at least one mapping, while 26 remaining ontologies did not reference any other ontology. Again, the majority of mappings were bidirectional and symmetric. We calculated the graph density of 0.473 and a modularity of 0.339. Six communities of densely connected ontologies were identified. The result is shown in Fig. 2 where each color represents one community. A node represents an ontology and an edge represents a mapping between two ontologies. Ontologies with no mappings are presented with gray color on the right side of the figure. Due to a very dense graph, not all edges can be shown, and only edges between ontologies with more than 500 mappings are illustrated. Node labels represent ontology name abbreviations (please refer to BioPortal's webpage for full names). Edge thickness is proportional to number of concept mappings, where a thicker line represents a higher number of concept mappings. The node size is

---

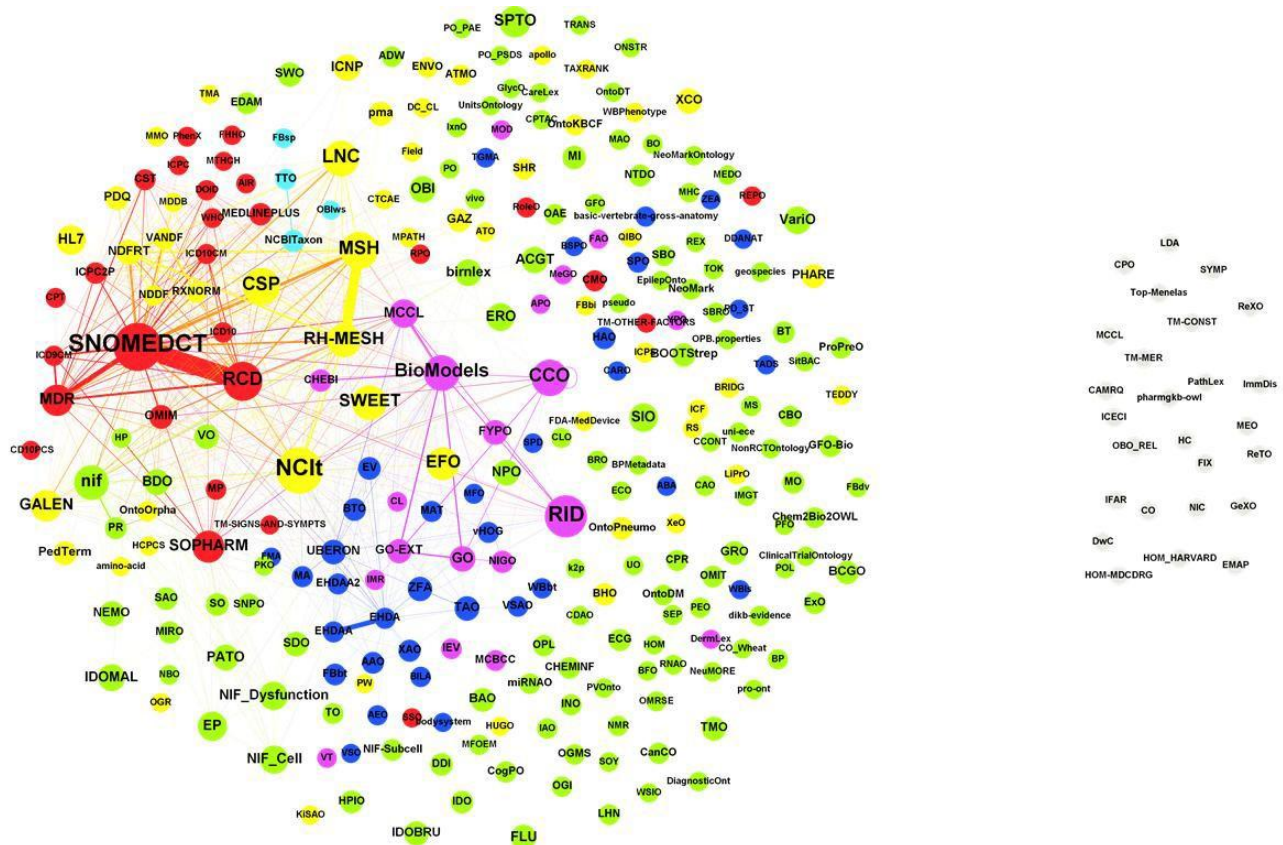[5] Available at http://bionlp.dbcls.jp/redmine/

Fig. 2. Graph illustrates a network of 294 publicly available biomedical ontologies. Six identified communities of densely connected ontologies are presented with six different colors (green, red, dark blue, light blue, purple and yellow). Ontologies with no mappings are presented on the right side of the figure (gray color).

proportional to the betweenness centrality metrics. A larger node represents a "bridging" ontology, i.e., an ontology with higher frequency of occurrence in all shortest paths between any two other ontologies.

Table 1 summarizes statistics for both visualization versions. The columns in the table stand for:

AO – Number of all ontologies;

CO – number of connected ontologies;

IO – number of isolated ontologies;

IC – number of identified communities;

GD – graph density; and

GM – modularity.

Version	AO	CO	IO	IC	GD	GM
Feb2013	294	268	26	6	0.473	0.339
Oct2012	284	254	30	5	0.380	0.346

Table 1. Comparison of statistics for two versions of BioPortal data.

4.2 Analysis of the graph

The graph in Fig. 2 shows that the top five "bridging" ontologies are SNOMEDCT (SNOMED Clinical Terms), NCIt (NCI Thesaurus), RID (RadLex), RCD (Read Codes, Clinical Terms), and MSH (Medical Subject Headings). The same ontologies scored the highest in the betweenness centrality also in the first visualization from October 2012.

As mentioned in the previous section, it is impossible to effectively visualize all edges between BioPortal ontologies with our method. The graph density of 0.473 and the modularity of 0.339 (Table 1) indicate that BioPortal ontologies present a strongly interconnected community of biomedical ontologies. Still, the graph in Fig 2. shows 6 groups of densely connected ontologies. We compared the identified communities with BioPortal categories and Table 2 shows the summary of our results.

Three groups in Table 2 relate to BioPortal's category classification and more or less clearly discriminate communities of ontologies related to anatomy (dark blue color),

health (red color), and taxonomy (light blue color). Other groups are more heterogeneous.

Color	Size (% in graph)	# of cate-gories	Prevailing categories
Green	47.21	37	Health, Biological process
Yellow	20.07	16	Health
Dark blue	13.01	14	Anatomy and its subclasses
Red	10.41	7	Health
Purple	7.43	17	Biological Process, Phenotype
Light blue	1.49	2	Taxonomic Classification

Table 2. Identified communities of ontologies and their prevailing BioPortal categories.

Although one might argue that there is no reason to include ontologies with no mappings (the gray colored nodes) in our visualizations, we wish to offer a complete graph of all BioPortal ontologies. In addition, with including ontologies with no mappings we offer ontology engineers the option of identifying new, recently added ontologies or ontologies that contain parsing errors.

When comparing the Feb2013 visualization with the Oct2012 version (Table 1), we can notice three interesting facts. First, note that the recent version contains 10 more ontologies in total but also less ontologies with no mappings. In addition, graph density increased by 0.093. This means that: first, new publicly available ontologies have been added to BioPortal repository or some alpha/retired ontologies changed their status; and second, some of previously unmapped ontologies have mappings in the new version. The BioPortal data was updated in January 2013 and mappings were re-generated which explains our findings.

Second, we can notice that some ontologies belong to different communities. An example of an ontology that changed the community is SOPHARM (Suggested Ontology for Pharmacogenomics). This ontology was in the purple (heterogeneous) community in the old version and it is in the red community (Health) in the new version. No updated version of SOPHARM was uploaded to BioPortal since our first visualization, therefore the change in community indicates that some ontologies may get more mappings to relevant ontologies through time.

Third, we can notice that modularity of the graph slightly decreased, which indicates that the new graph has less structure than the old version. However, a closer look shows one new community in the recent version. This community is represented with light blue color in Fig. 1 and it consists of the following 3 ontologies that fall into the "Taxonomic classification" BioPortal category: NCBITaxon (NCBI organismal classification), TTO (Teleost taxonomy) and FBsp (Fly taxonomy). The group also includes OBIws (Bioinformatics Web Service Ontology), which is an ontology with only 6 mappings

where 4 mappings reference the TTO ontology. Fig. 3 shows two enlarged graphs for this newly identified group. Colored nodes represent ontologies that are directly connected to the ontology in the focus (in the upper part of the figure, TTO is in the focus, while in the lower part NCBITaxon is in the focus). We can notice that the NCBITaxon and TTO present ontologies with highest Betweenness and they connect the taxonomic ontologies with other groups.
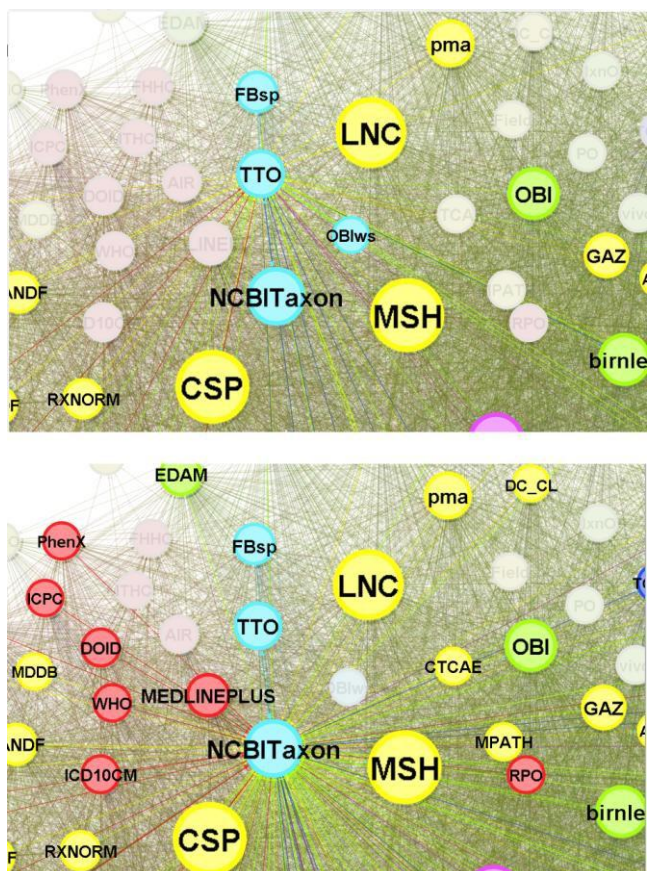

Fig. 3. Enlarged graph of the newly identified community (light blue color) and its neighbors. In the upper part direct mappings to TTO are shown and in the lower part direct mappings to NCBITaxon are the focus of the visualization.

5   CONCLUSION

In this paper we presented a visualization of BioPortal mappings. We collected the mapping data through BioPortal's web service and identified densely connected communities of ontologies and so called "bridging" ontologies, i.e., ontologies that connect many other ontologies/communities.

We believe that the graphs that we presented in this work can be helpful to ontology engineers when they

choose relevant ontologies for their reuse. When deciding which ontologies to reuse, engineers can use our graph for a quick overview of connected ontologies. For example, when building new anatomy ontology, the engineer might find the dark blue group and its connected ontologies from our graph interesting candidates for their reuse. In addition, generating graphs at different time points can identify ontologies that are "popular". For example, when a node becomes bigger comparing to a previous version of the graph, we can infer that the ontology represented by the node gets a lot of attention from other ontologies and it might be important. Size of a node in our graphs also implies which ontologies are general ontologies (bigger nodes) and which ontologies are more specific ontologies (smaller nodes). This is often important when selecting ontologies for their reuse.

The work presented in this paper is still in initial phase and future work is planned. At this point, we assume that mappings are correctly defined and we do not check if mapped concepts are used properly in the target ontology. Next, our technique does not consider size of ontologies. Large ontologies with high amounts of concepts usually contain higher number of mappings compared to smaller ontologies. As a result, large ontologies have advantage and also more chances to be identified as bridging ontologies. However, small, well-constructed ontologies may also contain important and relevant concepts. In the future we will also keep generating new graphs at different time points. In addition, our goal is to integrate these visualizations in the OntoFinder tool. We believe that it would be useful for users to be able to zoom in the graph, select the minimum mapping threshold shown, search for specific ontologies in the graph, etc. Also, we wish to visualize more detailed information about mappings, such as types of mappings (automatic, manual), algorithm used (for automatic mappings), authorship (for manual mappings) and direction of mappings (which ontology is the "source" of the term and which ontology is reusing it). In the future we will focus on how to integrate this information in OntoFinder and search for new ways to visualize mappings. We will also consider feedback from ontology engineers and users to improve our tools.

ACKNOWLEDGEMENTS

REFERENCES

Blondel, V.D., Guillaume J.L., Lambiotte, R., Lefebre, E., Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 10, P1000, (2008).

Euzenat, J., Loup, D., Touzani, M., Valtchev, P.: Ontology alignment with OLA, in: Proceedings of the Third International Workshop on Evaluation of Ontology-based Tools (EON), (2004).

Falconer, S. and Storey, M.-A.: A cognitive support framework for ontology mapping. In: Proceedings of ISWC/ASWC, (2007).

Ghazvinian, A., Noy, N.F., Musen, M.A.: Creating mappings for ontologies in biomedi-cine: simple methods work. AMIA Annu. Symp. proc. 2009:198–202, (2009).

Grosso, W.E., Eriksson, H., Fergerson, R., Gennari, J.H, Tu, S.W, Musen, M.A.: Knowledge modeling at the millennium (the design and evolution of Protégé-2000). Proc 12th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop Canada; 1999:7–4–1 to 7–4–36. Euzénat et al., (2004).

Kocbek, S., Perret J.L., Kim , J.D., Visual presentation of mappings between biomedical ontologies, In Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2012), (2012).

Kolli, R., Doshi, P., OPTIMA: Tool for Ontology alignment with Application to Semantic Reconciliation of Sensor Metadata for Publication in Sensor Map, In Proceedings of 2008 IEEE International Conference on Semantic Computing, 484-485, (2008).

Lanzenberger, M. and Sampson, J.: Alviz - a tool for visual ontology alignment. In Proc. of Intl. Symposium of Visualization of the Semantic Web, (IV06-VSW)10th Intl. Conference Information Visualization. IEEE Computer Society, (2006).

Smith, B., Ashburner, M., Rosse, C., Bard, J., et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 1251-5, (2007).

Storey, M.-A., Mussen, M., Silva, J., Best, C., Ernst, N., Fergerson, R., Noy, N.: Jamba-laya: Interactive visualization to enhance ontology authoring and knowledge acquisition in Protégé. In Proceedings of Workshop on Interactive Tools for Knowledge Capture, K-CAP-2001, Victoria, BC, Canada, http://www.thechiselgroup.org/jambalaya., (2001)

Voyloshnikova, E., Fu, B., Grammel, L., Storey, M.A., BioMixer: Visualizing Mappings of Biomedical Ontologies, The Third International Conference on Biomedical Ontologies (ICBO 2012), (2012).

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).