Proceedings of the Fifth International Workshop on Data Integration in the Life Sciences (DILS 2008).
Berlin Heidelberg New York: Springer; 2008. p. 1-4.

Ontologies and Data Integration in Biomedicine: Success Stories and Challenging Issues

Olivier Bodenreider

Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
olivier@nlm.nih.gov

Abstract. In this presentation, we review some examples of successful biomedical data integration projects in which ontologies play an important role, including the integration of genomic data based on Gene Ontology annotations, the cancer Biomedical Informatics Grid (caBIG) project, and semantic mashups created by the Semantic Web for Health Care and Life Sciences community.

1 Introduction

The promise of translational medicine, hinges upon bridging basic research and clinical practice [1]. One key element to the integration of the research and clinical communities is the integration of the information sources and data used in these communities. In practice, bridges need to be created both across domains (e.g., between genotypic and phenotypic information sources) and across knowledge bases within a domain (e.g., between genomic and pathway resources). Biomedical ontologies play an important role in data integration [2]. They support data integration in two different ways, corresponding to two different approaches to data integration: warehousing and mediation [3]. One the one hand, by providing a controlled vocabulary in a given domain, ontology support the standardization required from *warehousing approaches* to data integration, in which the sources to be integrated are transformed into a common format and converted to a common vocabulary. On the other hand, *mediation-based approaches* use ontologies for defining a global schema (in reference to which queries are made) and mapping between the global schema and local schemas (the schemas of the sources to be integrated).

We review examples in which ontologies have been used successfully for integrating biomedical data, including the integration of genomic data based on Gene Ontology annotations, the cancer Biomedical Informatics Grid (caBIG) project, and semantic mashups created by the Semantic Web for Health Care and Life Sciences community. Barriers to integration are discussed next.

2 Gene Ontology

The Gene Ontology (GO) [4] is a controlled vocabulary for the functional annotation of gene products across species [5]. In less than a decade, GO has been adopted by

A. Bairoch, S. Cohen-Boulakia, and C. Froidevaux (Eds.): DILS 2008, LNBI 5109, pp. 1–4, 2008.

several dozen model organism communities (e.g., Mouse Genome Informatics [6]) and has become a *de facto* standard for functional annotation. In addition to standardizing annotations across species, GO asserts relations among terms, which also facilitates data integration. GO is an enabling resource for comparative genomics, because it allows researchers to compare and contrast the functions of genes and gene products across multiple organisms [7]. Annotations repositories can be integrated not only with other annotation repositories, but also with a variety of data, including gene expression profiles (microarray data).

3 Cancer Biomedical Informatics Grid (caBIG)

The cancer Biomedical Informatics Grid (caBIG) of the National Cancer Institute (NCI) establishes a common infrastructure used to share data and applications across institutions to support cancer research efforts [8] in a grid environment [9]. Ontological resources such as the NCI Thesaurus [10] and the Cancer Data Standards Repository (caDSR) [11], a metadata registry for common data elements, are key resources of the common infrastructure for cancer informatics [12]. The data services currently available include, for example, caArray [13], a microarray data repository and gridPIR [14], a proteomic information resource based on UniProt and other databases from the Protein Information Resource (PIR). The Cancer Translational Research Informatics Platform (caTRIP) [15] takes a mediator-based approach to integrating a number of caBIG data services. Common data elements (CDEs) from the caDSR are used to join and merge data from the various repositories. CaBIG completed a 4-year pilot phase in 2007, involving 1,000 individuals from almost 200 organizations. In the next phase, caBIG tools and infrastructure will be made deployed to NCI-designated cancer centers.

4 Semantic Web for Health Care and Life Sciences

For the past two years, the World Wide Web Consortium (W3C) Health Care and Life Sciences Interest Group (HCLSIG) [16] has investigated the use of Semantic Web technologies in biomedicine. Ontologies play a central role in the Semantic Web [17], especially in biomedicine for which a large number of ontologies have been developed. This group advocates the use of Semantic Web technologies for supporting translational research [18] and has demonstrated the feasibility of integrating disparate resources in the domain of neurosciences, including Entrez Gene, Gene Ontology Annotations, the Allen Brain Atlas, PubMed/MEDLINE, and MeSH [19]. Other such "mashups" (integrative applications) have been developed since (e.g., [20]). Similar approaches have been used to integrate genotype and phenotype information [21], pathway and disease information [22], and to create drug-target networks [23]. Biomedical ontologies are crucial to these integration projects.

5 Challenging Issues

Freely and publicly available – preferably in several popular formats, easily discoverable and widely distributed ontologies are enabling resources for data integration, especially

when they are embraced by active communities, used as a *de facto* standard in major data repositories and can interoperate with other ontologies. Integration is further facilitated by the availability of tools developed for and interfaces to these ontologies. This scenario essentially characterizes the Gene Ontology and explains in part its success.

There are, however, many obstacles preventing ontologies from being used efficiently for data integration. Despite the existence of repositories such as the National Center for Biomedical Ontology's BioPortal [24] and the Unified Medical Language System (UMLS) [25], not all ontologies can be accessed easily. Furthermore, some ontologies in the UMLS are subject to intellectual property restrictions and the UMLS cannot be used without first signing a license agreement. While OBO and OWL are popular formalisms for representing ontologies, many ontologies are only available in proprietary formats.

There is no authoritative mechanism for creating unique identifiers for biomedical entities. As a result, the same entity is often present under different identifiers in multiple ontologies, impeding integration. *Post hoc* mappings across ontologies such as those created by the UMLS somewhat alleviate this problem, but do not provide a complete solution. Additionally, in the Semantic Web, there is a need for a standard way of representing identifiers (e.g., URIs), as well as for services bridging identifiers across namespaces.

Differences in the granularity of annotations across datasets are also an issue, partially compensated by the use of aggregation strategies, such as the GO Slims [26] and the use of semantic similarity metrics [27]. Finally, not all datasets are directly amenable to integration. For example, metadata elements describing gene expression data in microarray repositories and fields in genome-wide association studies (e.g., Framingham Heart Study) are often in free text, not annotated to any ontology. Such datasets need to be preprocessed and encoded to an ontology prior to being integrated with other datasets.

References

1. Marincola, F.: Translational Medicine: A two-way road. Journal of Translational Medicine 1,1 (2003)
2. Bodenreider, O.: Biomedical ontologies in action: role in knowledge management, data integration and decision support. IMIA Yearbook of Medical Informatics, 67–79 (2008)
3. Hernandez, T., Kambhampati, S.: Integration of biological sources: Current systems and challenges ahead. Sigmod Record 33, 51–60 (2004)
4. Gene Ontology, http://www.geneontology.org/
5. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000)
6. Mouse Genome Informatics, http://www.informatics.jax.org/
7. Blake, J.A., Bult, C.J.: Beyond the data deluge: data integration and bio-ontologies. J. Biomed. Inform. 39, 314–320 (2006)
8. caBIG Strategic Planning Workspace: The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community. Medinfo. 12, 330–334 (2007)

9. Oster, S., Langella, S., Hastings, S., Ervin, D., Madduri, R., Phillips, J., Kurc, T., Siebenlist, F., Covitz, P., Shanbhag, K., Foster, I., Saltz, J.: caGrid 1.0: an enterprise Grid infrastructure for biomedical research. J. Am. Med. Inform. Assoc. 15, 138–149 (2008)
10. NCI Thesaurus, http://www.nci.nih.gov/cancerinfo/terminologyresources
11. Cancer Data Standards Repository (caDSR), http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr
12. Komatsoulis, G.A., Warzel, D.B., Hartel, F.W., Shanbhag, K., Chilukuri, R., Fragoso, G., Coronado, S., Reeves, D.M., Hadfield, J.B., Ludet, C., Covitz, P.A.: caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. J. Biomed. Inform. 41, 106–123 (2008)
13. caArray, http://caarray.nci.nih.gov/
14. gridPIR, https://cabig.nci.nih.gov/tools/PIR
15. caTRIP, https://cabig.nci.nih.gov/tools/caTRIP
16. Health Care and Life Sciences Interest Group, http://www.w3.org/2001/sw/hcls/
17. Schroeder, M., Neumann, E.: Semantic web for life sciences. Web Semantics: Science, Services and Agents on the World Wide Web 4, 167–167 (2006)
18. Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, M.S., Ogbuji, C., Rees, J., Stephens, S., Wong, G.T., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.H.: Advancing translational research with the Semantic Web. BMC Bioinformatics 8, Suppl. 3, S2 (2007)
19. HCLS Banff 2007 demo, http://esw.w3.org/topic/HCLS/Banff2007Demo
20. Sahoo, S.S., Bodenreider, O., Rutter, J.L., Skinner, K.J., Sheth, A.P.: An ontology-driven semantic mash-up of gene and biological pathway information: Application to the domain of nicotine dependence. Journal of Biomedical Informatics (2008), doi:10.1016/j.jbi.2008.1002.1006
21. Butte, A.J., Kohane, I.S.: Creation and implications of a phenome-genome network. Nat. Biotechnol. 24, 55–62 (2006)
22. Chabalier, J., Mosser, J., Burgun, A.: Integrating biological pathways in disease ontologies. Medinfo. 12, 791–795 (2007)
23. Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L., Vidal, M.: Drug-target network. Nat. Biotechnol. 25, 1119–1126 (2007)
24. BioPortal, http://www.bioontology.org/tools/portal/bioportal.html
25. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32, 267–270 (2004)
26. GO Slim, http://www.geneontology.org/GO.slims.shtml
27. Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A.: Semantic similarity measures as tools for exploring the gene ontology. In: Pac. Symp.Biocomput., pp. 601–612 (2003)