Mark Tomko

Translating Biological Data Sets Into Linked Data

Introduction

Modern biology researchers generate vast quantities of data which are collected and made available online. Sites such as the United States National Center for Biotechnology Information[1] (NCBI), the European Bioinformatics Institute[2] (EBI), and the Wellcome Trust Sanger Institute[3] now provide gigabytes of information about genes, genomic expression, proteins, and the taxonomy and phylogeny of organisms. Although much of the data are interrelated, different data providers make their data available in a plethora of special-purpose data formats. Many institutions supply software tools to help researchers discover, understand, annotate, and contextualize the information they provide.

However, the diversity of data formats reduces interoperability between software tools and data hosted by different providers. These artificial, though unintentional, boundaries impede researchers' abilities to study multiple data sets in conjunction. Linked data standards present an opportunity for data formats to converge, thereby enabling software tools to more effectively make use of the entire range of data available.

Translation of existing data sets into linked data knowledge organization schemes has the potential to significantly increase the value and utility of existing knowledge. I have developed a preliminary translation program that expresses Pfam, a large database of protein families, as a SKOS graph. The knowledge organization scheme resulting from

[1] http://www.ncbi.nlm.nih.gov/
[2] http://www.ebi.ac.uk/
[3] http://www.sanger.ac.uk/

this project and the program used to generate it are available online[4]. This translation serves as a demonstration of the ability of existing linked data vocabularies to represent the collected knowledge of current research, which may improve the ability of researchers in the biological sciences to share and process scientific data.

Knowledge Organization Schemes In The Biological Sciences

A number of projects have established knowledge organization schemes for well-known bioinformatics data sets. BioThesaurus maps the names of genes and proteins to protein sequences maintained in the UniProt database (Liu, Hu, Zhang, & Wu, 2005). The PRO protein ontology provides a description of relationships between proteins and protein classes. PRO is itself a component of a larger collection of biological ontologies maintained by the Open Biomedical Ontologies[5] (OBO) consortium (Natale et al., 2007).

The OBO consortium consists of over 60 biological ontologies, most of which are expressed in a file format created for OBO. However, efforts are underway to provide translations of OBO ontologies into the OWL linked data standard (Smith et al., 2007). The existence of these high-quality knowledge organization systems demonstrates the value of encoding biological data sets according to linked data standards and making them available for wider use.

Pfam And UniProt

Computational biologists, evolutionary biologists, and structural biologists frequently study protein families and protein sequences (Finn et al., 2009, p. 1). A protein family is a group of proteins exhibiting a high degree of similarity in the primary

[4] http://web.simmons.edu/~tomko/pfam/
[5] http://www.obofoundry.org/

sequence of amino acids (Nelson & Cox, 2000, p. 188). The biochemistry underlying cellular function depends on combinatorial interactions involving proteins. Elucidating structural and functional similarities between proteins helps researchers understand both biochemical processes that occur within cells as well as evolutionary relationships between organisms.

Pfam is a publicly available database of protein families, which uses multiple sequence alignments and Hidden Markov Models (HMMs) to identify each family (Finn et al., 2009, p. 1). The most recent release of Pfam, version 25.0, contains over 12,000 protein families (Finn, 2011). The proteins that make up Pfam families derive from UniProt (Finn et al., 2009, p. 3), an extensive database of protein sequences developed by a consortium of European and American institutions (Jain et al., 2009; The UniProt Consortium, 2010). Together, Pfam and UniProt comprise a large and richly structured data set linking information about proteins, protein structures, and similarities between proteins in a variety of organisms.

Both Pfam and UniProt provide tools to search and browse their data sets online. Although UniProt provides its data in a variety of formats including RDF/XML, the Pfam database is available only in a textual description known as Stockholm format. Some software tools exist that can process Stockholm files, but the use of the format is generally restricted to the Pfam data set.

Translating Pfam Into SKOS

To design the linked data representation of the Pfam data set, I began by identifying key concepts expressed in the data. Based on these, I developed a mapping between the Pfam concepts and the SKOS vocabulary. This conceptual design led to the

development of a computer program capable of translating the Pfam data files from the original Stockholm representation into a SKOS graph.

Developing the computer program presented a number of challenges. For instance, the program needed to track relationships between proteins and protein families, potentially requiring a significant amount of memory. Therefore, a key design goal for the program was to minimize the in-memory state it maintained. The translation program was implemented in the Scala programming language (École Polytechnique Fédérale de Lausanne, 2011; Odersky et al., 2006). The complete source code for the translation program is available at the author's GitHub site[6].

Conclusion

This proof-of-concept system for translating an existing large data into a linked data knowledge organization scheme demonstrates that linked data techniques hold value for expressing shared biological data. By integrating data from Pfam with the UniProt protein database, it provides an example of how large data sets expressed using open standards may be used in conjunction to improve access to modern scientific knowledge. Enabling the use of shared data sets together maximizes their value to the scientific community.

References

École Polytechnique Fédérale de Lausanne. (2011). The Scala Programming Language. Retrieved June 24, 2011, from http://www.scala-lang.org/

[6] http://github.com/mtomko/pfamskos/

Finn, R. D. (2011, April 1). No, seriously, we've made a release. *Xfam Blog*. Retrieved June 24, 2011, from http://xfam.wordpress.com/2011/04/01/no-seriously-weve-made-a-release/

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., et al. (2009). The Pfam protein families database. *Nucleic Acids Research*, *38*(Database), D211-D222. doi:10.1093/nar/gkp985

Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B., Martin, M., et al. (2009). Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics*, *10*(1), 136. doi:10.1186/1471-2105-10-136

Liu, H., Hu, Z.-Z., Zhang, J., & Wu, C. (2005). BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, *22*(1), 103 -105. doi:10.1093/bioinformatics/bti749

Natale, D. A., Arighi, C. N., Barker, W. C., Blake, J., Chang, T.-C., Hu, Z.-Z., Liu, H., et al. (2007). Framework for a Protein Ontology. *BMC Bioinformatics*, *8*(Suppl 9), S1. doi:10.1186/1471-2105-8-S9-S1

Nelson, D. L., & Cox, M. M. (2000). *Lehninger Principles of Biochemistry* (3rd ed.). New York: Worth Publishers.

Odersky, M., Altherr, P., Cremet, V., Dragos, I., Dubochet, G., Emir, B., McDirmid, S., et al. (2006). *An Overview of the Scala Programming Language*. Lausanne, Switzerland: École Polytechnique Fédérale de Lausanne.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support

biomedical data integration. *Nature Biotechnology*, *25*(11), 1251-1255. doi:10.1038/nbt1346

The UniProt Consortium. (2010). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research*, *39*(Database), D214-D219. doi:10.1093/nar/gkq1020