Linking Bioinformatics Research data and Publications through Metadata and Knowledge Organization Systems

Jian Qin
School of Information Studies
Syracuse University
Syracuse, NY, USA
jqin@syr.edu

Marcia Lei Zeng
School of Library and Information Science
Kent State University
Kent, OH, USA
mzeng@kent.edu

Bioinformatics is a field that has many well established knowledge organization systems (KOS) [1]. Typical bioinformatics data types include DNA and protein sequences, macromolecular structures, genomes, and gene expressions (Luscombe, Greenbaum, & Gerstein, 2001).

KOS in bioinformatics serves two primary functions: 1) organizing knowledge of organisms through applying scientific taxonomy and nomenclature in order to identify, name, and classify them in bioinformatics data as well as the metadata that describes such data; and 2) organizing information and knowledge contained in research publications, that is, scholarly output from studying the organisms, as well as in regulation and guideline documents.

Conventionally, KOS resources are characterized by their structure or function. Examples of hierarchically structured KOS include the National Center for Biological Information (NCBI)'s *NCBI Taxonomy* and *NCBI Organismal Classification*. KOS such as thesauri and subject heading lists primarily arrange the terms representing concepts according to a known order (such as alphabetical) while using attributes to show terms related to a concept and reveal relationships between the concept and other immediately related concepts. The *Medical Subject Headings (MeSH)* and *NCI Thesaurus (NCIt)* are two examples of this type of KOS. Many KOS recourses have employed both of such primary structures with different quality levels in regard to the logic and semantics embedded in the vocabularies. Emerging ontologies during last two decades (such as *Gene Ontology* and *Cell Ontology*) that are also widely used in bioinformatics databases have been built on the existing KOS structures but with stronger formality.

Looking closely, subtle distinctions exist between the KOS used for representing the organisms in research data and those used for representing topics *about* the organisms in research publications:

- Taxonomies have traditionally been used to describe research data in bioinformatics. For example, each entry in the *NCBI Taxonomy* (http://www.ncbi.nlm.nih.gov/taxonomy) identifies an organism by an ID, inherited blast name, rank, genetic code, other names, type material (if any), and full lineage. When a taxon is assigned to a genetic sequence, it connects the sequence data to the organism represented by the taxonomy. In general, hierarchical KOS semantically identifies an organism, which is used in the metadata that documents the origin, discovery, geospatial location, and other features related to the organism.
- Subject headings, thesauri, term lists, and other controlled vocabularies are intended to represent the knowledge in scholarly publications about the organisms rather than to document the organisms and their hierarchical relationships.

Although the distinction as presented above does exist, there is also an essential connection between the KOS serving the two purposes mentioned above. The topics in scholarly publications are produced based on research data (including data about organisms). Linking research data to publications is becoming an increasingly common practice among research data repositories. Dryad Digital Repository (http://datadryad.org/), for instance, is a "curated resource that makes the data underlying scientific publications discoverable, freely reusable, and citable" (Dryad, 2016). GenBank, one of the largest international data repositories for DNA sequences, has been providing PubMed links in the metadata records that describe genetic sequences. Nevertheless, the pathway between different KOS is not readily available for knowledge discovery from bioinformatics data to scholarly publications.

This presentation will use two cases to show how bioinformatics data and research publications might benefit from interoperable KOS resources for more effective knowledge discovery. A key point we are trying to make is that converting KOS into a Linked Open Data format is not simply a matter of using SKOS or OWL to transform natural language taxonomies and thesauri into RDF triple stores, but rather, it is a process of remodeling and redesigning of KOS in the context of knowledge representation networks.

The first case is an ongoing study that investigates the knowledge nodes in precision medicine publications. By deriving the sources, types, and relations of knowledge nodes from these publications, we aim to develop a knowledge network scheme that will enable easy incorporation and linking of different types of KOS for domain metadata.

Another case focuses on the types of interrelationships between the KOS that have been traditionally used to represent organisms and the KOS that have been primarily used to describe research publications. While there have been some mappings between KOS vocabularies in bioinformatics, e.g., most concepts in the *Gene Ontology* (http://geneontology.org/) has been mapped to those in *NCBI Taxonomy*, concepts in hierarchically structured KOS are not always directly map-able to those in the KOS that created in non-hierarchical manner . Analyzing what common foundations exist and shared by the two types of KOS will be helpful for establishing valid and interoperable relationships between KOS vocabularies.

References

Dryad. (2016). The organization: Overview. Web resource. http://datadryad.org/pages/organization

Luscombe, N.M., Greenbaum, D. & Gerstein, M. (2001). What is bioinformatics? An introduction and overview. Yearbook of Medical Informatics, pp. 83-99. https://www.ebi.ac.uk/luscombe/docs/imia_review.pdf

Note: [1] In Bioportal repository [http://bioportal.bioontology.org/], there are 683 registered KOS vocabularies as of Sept. 5, 2016, including 535 ontologies and 138 other KOS vocabularies that are in ontology views.