Labeling and Enhancing Life Sciences Links

Stephan Heymann
Humboldt-Universitat zu Berlin
heyman@dbis.informatik.hu-berlin.de

Felix Naumann
Humboldt-Universitat zu Berlin
naumann@informatik.hu-berlin.de

Louiqa Raschid
University of Maryland
louiqa@umiacs.umd.edu

Peter Rieger
Humboldt-Universitat zu Berlin
rieger@dbis.informatik.hu-berlin.de

1 Introduction

Life sciences data sources contain data about scientific objects such as genes and sequences that are richly interconnected, i.e., a gene object may have links to sequences, proteins, SNPs, citations, etc. Scientific knowledge is enhanced by exploration of relationships between scientific objects, requiring traversal of both links and paths (informally concatenations of links). There are significant limitations and challenges of such exploration, because the links are inherently poor with respect to syntactic representation and semantic knowledge. The links are syntactically poor because the source and the target of the link are typically specified at the level of data objects (data entries). However, most scientists understand that the source and the target of a link are potentially at a finer level of granularity, and may correspond to specific subelements or fields within these data entries. The links are semantically poor because they carry no explicit meaning beyond that the data entries are somehow "related". The lack of syntactic and semantic knowledge prevent the development of tools that can assist scientists to fully explore data sources and interconnections. In this extended abstract, we provide examples of semantically enhanced links and discuss our research to develop a methodology to enhance the structure and the meaning associated with links.

2 Enhancing Links

Links are added to data entries for many different reasons. Biologists insert them when they discover a certain relationship following an experiment or study. Data curators add links to augment, complete or to make consistent, the knowledge captured among multiple sources. For example, a result reported in a paper in PubMed may lead a curator to insert a link from a data entry in say Omim to this citation in PubMed. Algorithms insert links automatically when discovering similarities among two data items, e.g., to represent sequence similarity following a BLAST search. Thus, the simple unlabeled links that are in use today are insufficient to represent such subtle and diverse relationships.

Figure 1 illustrates a specific labeling of links with some semantics or meaning associated with each link [1]. We emphasize that today, physical links between Web-accessible data sources *do not support meaningful labeled links* as illustrated in this figure. Such a labeled view of the links between data sources is overlaid on the physical links in the SRS (version 8.0) data integration system. One of the objectives of our research is to provide a data model and query language that can represent and exploit labeled links, *enhanced links* or *e-links*, so that the labeled view supported by SRS can be exploited by scientists as they explore and integrate data. However, we are interested in enhancing the links beyond the labeling described in Figure 1.

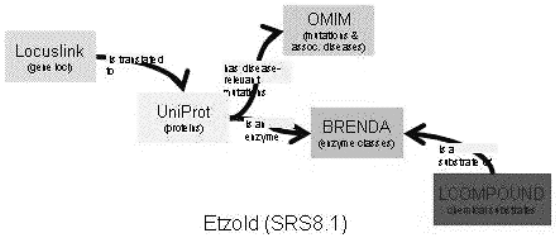

Figure 1. A Particular Labeling of Links

Consider a SwissProt entry with a link to an Omim entry with a certain ID. In the flat structure of the SwissProt entry, this link is represented as a top-level attribute in the form of an Omim ID, and it sometimes has an HTML hyperlink to a data source stor-

ing that particular OMIM entry. A link in that form neither represents the sub-element of the SWISSPROT entry to which the link refers, nor the sub-element of the OMIM entry to which the link points, nor does it represent the reason to insert this link. Biologists regarding the SWISSPROT entry rely on their experience and can sometimes infer these link properties after a time-consuming examination. Machines and algorithms cannot perform such analysis at the necessary level of detail and precision. In this particular case, the *e-link* should not originate from the SWISSPROT entry; instead the origin is the CC-DISEASE attribute within that entry. The *e-link* should also not represent a generic relationship; it should be labeled as a *causal* relationship, telling humans and machines that the protein in question is known to *cause the disease*.

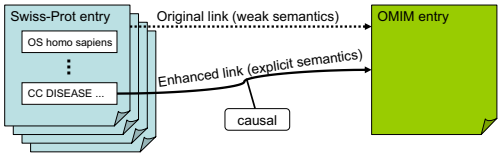

Figure 2. The *e-link* from Swiss-Prot to OMIM

We consider more complex situations to illustrate that enhancing links is a challenge. Consider the physical link from origin source UniProt to target source OMIM. This physical link actually corresponds to two *e-links*. Both *e-links* originate in the same sub-element of UNIPROT. One *e-link* has the meaning *is causal for disease* and the target sub-element in OMIM is CLINICALFEATURES. The second *e-link* has the meaning *describes genetic defects* and the target sub-element in OMIM is MAPPING. In this example, the original links represents two *e-links*, whose target sub-element in OMIM is different, and where the links have different meaning.

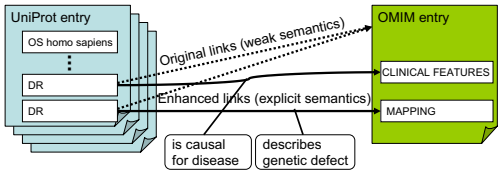

Figure 3. Enhancing a Link from UNIPROT to OMIM to Produce Two *e-links* with Different Target Sub-Elements in OMIM

Finally, we consider the link from origin source UNIPROT to target source GO. This link captures 3 *e-links*, where the origin sub-element and the meaning for the 3 *e-links* is different; it is illustrated in Figure 4.

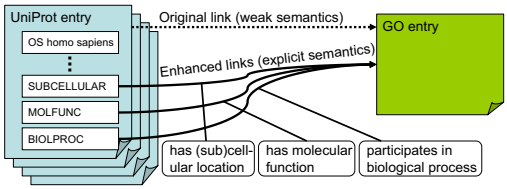

Figure 4. Enhancing a Link from UNIPROT to GO to Produce Three *e-links* with Different Origin Sub-Elements in UNIPROT

3 Methodology to Explore *e-links*

We propose to enhance the current link implementation, so as to support more meaningful queries over *e-links*. Enrichment should include semantic labels, descriptors, and a more precise identification of the link's source and target elements within the data entry. One can then traverse paths and perform a comparison of paths that is meaningful to the biologist.

A semantically enhanced link *e-link* will comprise the following:

- A link identifier corresponding to the link category.
- A set of navigational paths to specify the origin of the link with respect to the parent data entry containing the origin.
- A set of navigational paths to specify the target of the link with respect to the parent data entry containing this target.
- A link label or category and a link descriptor; typically they will refer to some ontology.

Our research will develop the methodology to utilize *e-links* in exploration. To this end, we are developing a data model based on *e-links* to capture the syntactic representation and semantic knowledge associated with a link. Additionally, a concatenation operator can prescribe when it is meaningful to concatenate links to form paths, and when links and paths are equivalent. Finally, a query language and interpreter will allow scientists to meaningfully explore the links and paths.

Acknowledgements. This research is partially supported by NSF Grants IIS0219909 and EIA0130422.

References

[1] T. Etzold. The bioinformatics data integration tunnel: Do we see the light yet? invited presentation. In *DILS*, 2004.