Enhancing the Semantics of Links and Paths in Life Science Sources

S. Heymann, F. Naumann and P. Rieger
Humboldt-Universität zu Berlin
Berlin, Ger many
{naumann}@informatik.hu-berlin.de

L. Raschid
University of Maryland
College Park, Maryland
louiqa@umiacs.umd.edu

1 Introduction

An abundance of Web-accessible bio-molecular data sources contain data about scientific entities such as genes, sequences, proteins and citations. The sources are diverse in content, they are richly interconnected to each other, and their contents have varying levels of overlap. Experiment protocols to retrieve relevant data objects require data integration queries that explore multiple sources. To answer such queries, biologists or the query engines that they use must traverse both the links and the paths (informally concatenations of links) through these sources. While such navigational queries that traverse links and paths are critical to scientific exploration, they also pose significant limitations and challenges, since the links do not capture explicit semantics.

The links are inherently poor with respect to both syntactic representation and semantic knowledge. The links are syntactically poor because the source and the target of a link are usually specified at a high level of granularity, i.e., at the level of data objects (or data entries) in sources. However, upon further study, it is clear that the "real" source and target of a link should potentially occur at a finer level of granularity, and correspond to particular sub-elements within these objects. At present, links cannot represent this knowledge. The greater limitation is that the links are semantically poor and carry no explicit meaning other than the fact that the data entries are somehow "related". Scientists who examine the linked objects are usually able to infer the meaning of the link, but this knowledge cannot be exploited by query engines that evaluate queries for life scientists.

These limitations at the level of representation and meaning make it difficult for links to be explored meaningfully when answering queries. In this research, we propose a methodology and tools to assist scientists in exploring and exploiting the knowledge captured in these sources and their interconnections. In order to do so we must accomplish the following:

- Develop a data model that can represent sources, data objects and the enhanced *e-link*s between data objects. The data model will be augmented with a semantics to compose *e-link*s links into meaningful paths, *e-path*s. We must also develop techniques to generate and label existing links.
- Develop a query language and query evaluation engine for scientists to meaningfully explore these *e-link*s and *e-path*s.

The three major repositories, NCBI, DDBJ and EBI have made significant efforts recently to provide integrated access to the many entries and links between entries that exist in the sources that they manage. Examples include RefSeq and LocusLink at NCBI. However, they do not attempt to enhance the representation and the semantics of links so that they can be exploited by a query language, as will be described in this project. There are other projects that target specific links. For example, the PDBSProtEC project [1, 7] is a resource to link PDB chains with SwissProt codes and EC numbers. We expect that there will be many such efforts to enhance specific links, all of which can be exploited in our research.

In this extended abstract, we provide several examples of semantics associated with links and then outline a simple model for *e-link*s.

2 Modeling Life Science Sources

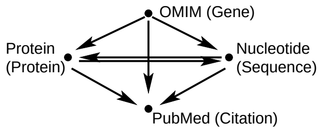


Figure 1: A Source Graph for NCBI Data Sources (and Corresponding Scientific Entities)

Life science sources may be modeled at three levels: the physical level, the object level and the logical level. The physical level corresponds to the actual data sources and the links that exist between them. An example of data sources and links is shown in Figure 1. The sources are a subset of sources at the National Center for Biotechnology Information (NCBI) and can be accessed at `http://www.ncbi.nlm.nih.gov`.

The physical level is modeled by a directed *Source Graph*, where nodes represent data sources and edges represent a physical link between two data sources. A data object in one data source may have a link to one or more data objects in another data source, e.g., a gene associated with a disease in OMIM links to multiple citations in PUBMED. An *Object Graph* as shown in Figure 2 represents the data objects of the sources and the object links between the objects. Thus, each link in the *Source Graph* corresponds to a collection of object links of the *Object Graph*, each going from a data object in one source to another object, in the same or a different source. Note that links in *Source Graph* can be bi-directional (though not always symmetric) and *Source Graph* may be cyclic.

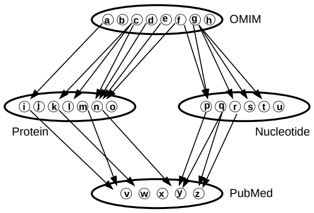


Figure 2: An Object Graph for NCBI Data Sources with Data Entries (Objects) and Links

The logical level consists of classes (entity classes, concepts or ontology classes) that are implemented by one or more physical data sources or possibly parts of data sources of *Source Graph*. For example, the class *Citation* may be implemented by the data source PUBMED. A source of *Source Graph* typically provides a unique identifier for each of the entities or objects in *Object Graph* and includes attribute values that characterize them. Table 1 provides a mapping from the logical classes to some physical data sources of some *Source Graph*.

CLASS	DATA SOURCE
Sequence (s)	NCBI Nucleotide database EMBL Nucleotide Sequence database DDBJ
Protein (p)	NCBI Protein database Swiss-Prot
Citation (c)	NCBI PubMed

Table 1: A Possible Mapping from Logical Classes to Physical Data Sources of *Source Graph*

Figure 1 illustrated a simple *Source Graph*. However, there are thousands of sources, and this number is increasing. These sources are also richly interconnected. Figure 3 [5] illustrates the multiplicity of sources and links between sources supported by the SRS (version 8.0) data integration system.

3 Enhancing Links Among Data Entries

Physical links between the objects or data entries in the physical sources are created for many different reasons. Biologists insert them when they discover a certain relationship following an experiment or study. Data curators add links to augment, to 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 physical links that are in use today are insufficient to represent such subtle and diverse relationships.

Figure 4 illustrates a specific labeling of links with some semantics or meaning associated with each link [5]. We emphasize that today, physical links between Web-accessible data sources *do not support meaningful labeled links* as illustrated in this figure. Such a labeling of the links between data sources is overlaid on the physical links in the SRS (version 8.0) data integration system. We are interested in enhancing the representation and semantics captured by *e-links* beyond the labeling described in Figure 4 and in providing a data model and query language that can represent and exploit *e-links*.

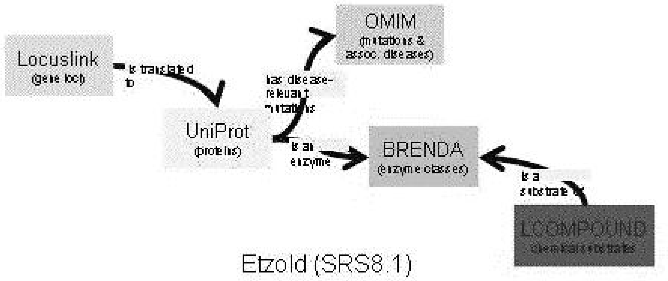
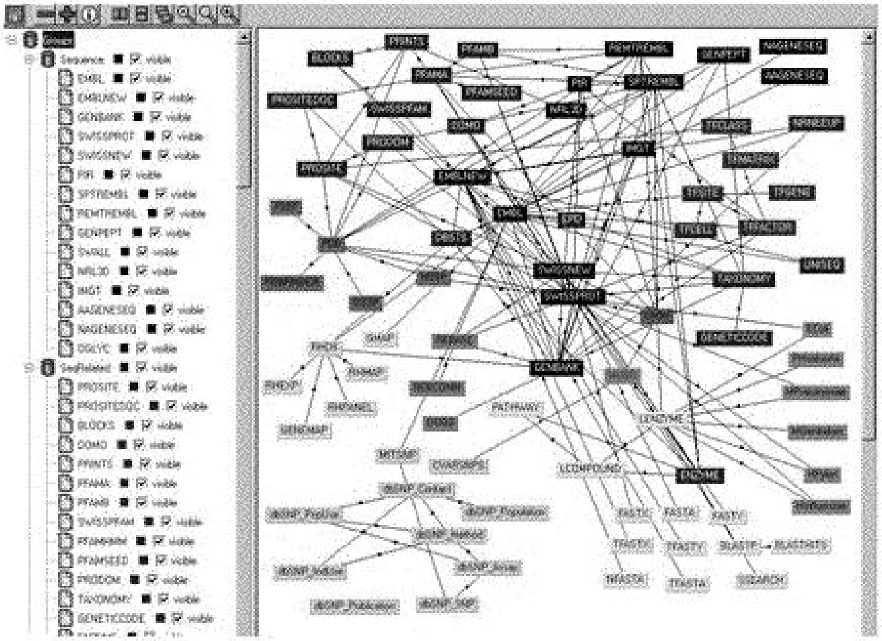


Figure 4: A Particular Labeling of Links [5]

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 by embedding an OMIM ID as a top-level attribute of the entry, and the entry may include an HTML hyperlink to the OMIM entry. Such a link 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 examining the SWISSPROT entry rely on their experience and can infer



- Sequence ■ ☑ visible
- EMBL ■ ☑ visible
- EMBLNEW ■ ☑ visible
- GENBANK ■ ☑ visible
- SWISSPROT ■ ☑ visible
- SWISSNEW ■ ☑ visible
- PIR ■ ☑ visible
- SPTREMBL ■ ☑ visible
- REMTREMBL ■ ☑ visible
- GENPEPT ■ ☑ visible
- SWALL ■ ☑ visible
- NRL3D ■ ☑ visible
- IMGT ■ ☑ visible
- AAGENESEQ ■ ☑ visible
- NAGENESEQ ■ ☑ visible
- DBLYC ■ ☑ visible
- SeqRelated ■ ☑ visible
- PROSITE ■ ☑ visible
- PROSITEDOC ■ ☑ visible
- BLOCKS ■ ☑ visible
- DOMO ■ ☑ visible
- PRINTS ■ ☑ visible
- PFAMA ■ ☑ visible
- PFAMB ■ ☑ visible
- SWISSPFAM ■ ☑ visible
- PFAMHMM ■ ☑ visible
- PFAMSEED ■ ☑ visible
- PRODOM ■ ☑ visible
- TAXONOMY ■ ☑ visible
- GENETICCODE ■ ☑ visible

Figure 3: The Multiplicity of Sources and Links Supported by SRS 8.0 [5]

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 ”real” 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* pointed to by the *e-link*.

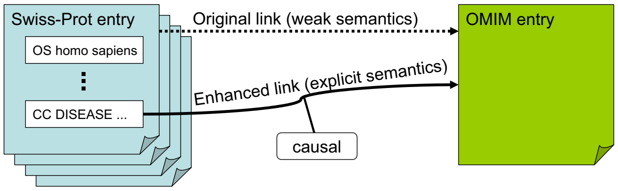

Swiss-Prot entry
OS homo sapiens
⋮
CC DISEASE ...
Original link (weak semantics)
Enhanced link (explicit semantics)
causal
OMIM entry

Figure 5: The Enhanced *e-link* from Swiss-Prot to OMIM

We note that determining the semantics and labels of *e-links* for some physical link between two sources may not be straightforward, and scientists may not always reach a consensus as to the desired semantics. Nevertheless, we believe that the significant activity related to ontologies for the life sciences, and the resulting advances in establishing controlled vocabularies to describe functionality and relationships among concepts, e.g., the GO Ontology [2] and GOA [3] will contribute towards the success of our research.

4 Further Examples of *e-links*

We consider several examples that represent more complex situations to illustrate that enhancing links is a challenge.

Consider the physical link from the origin source UniProt to the target source OMIM. The physical link instances between UniProt and OMIM entries actually corresponds to two distinct *e-links* with different semantics. Both *e-links* originate in the same subelement 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 subelement in Omim is MAPPING. In this example, the original physical link of the *SOurce Graph* is classified as two *e-links*, whose target sub-element in Omim is different, and where the two *e-links* have different meaning.

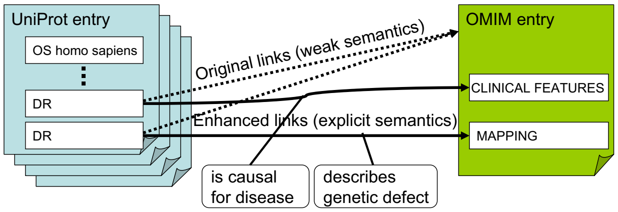


Figure 6: Enhancing a Link from UniProt to Omim to Produce Two *e-links* with Different Target Sub-Elements in Omim

Next, consider the link from the origin source UniProt to the target source GO. This physical link captures three *e-link*s, where the origin sub-element and the meaning of the three *e-link*s is different; it is illustrated in Figure 7. The target is the GO entry. The first *e-link* has meaning *has (sub)cellular location* and the origin sub-element in UniProt is SUBCELLULAR. The second *e-link* has the meaning *has molecular function* and the origin sub-element in UniProt is MOLFUNC and the third *e-link* has meaning *participates in biological process* and the origin sub-element in UniProt is BIOLPROC.

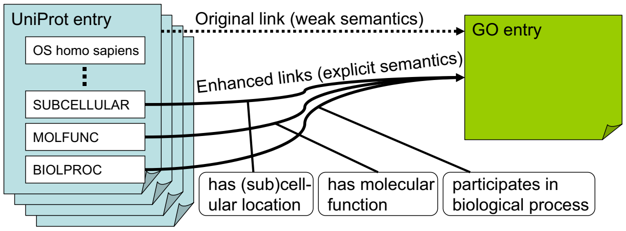

UniProt entry
OS homo sapiens
⋮
SUBCELLULAR
MOLFUNC
BIOLPROC
Original link (weak semantics)
GO entry
Enhanced links (explicit semantics)
has (sub)cellular location
has molecular function
participates in biological process

Figure 7: Enhancing a Link from UniProt to GO to Produce Three *e-links* with Different Origin Sub-Elements in UniProt

Finally, we consider the case where different physical links between different data sources appear to have the same meaning. There are six physical links in the *Source Graph*, each originating in the same sub-element of UniProt . The target of each link is a data entry in one of six different protein data sources, InterPro, Pfam, SMART, PROSITE, PRINTS, and TIGRFAMS; the links are illustrated in Figure 8. While the physical links are between different sources, they each have the meaning *contains a sequence signature.* This example of six physical links producing potentially six *e-link*s, all of which are equivalent with respect to meaning is a frequent occurrence in life science sources since the contents of sources overlap and the sources are richly interconnected. This motivates our research on a data model that can specify the equivalence of *e-link*s that are of the same link type.

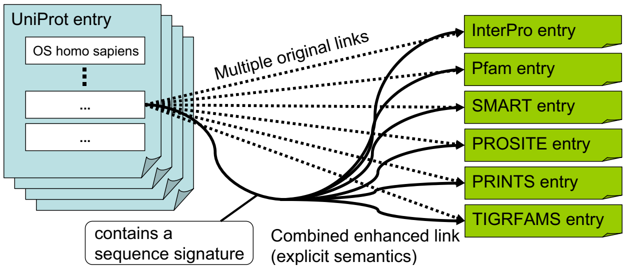


Figure 8: Enhancing a Link from UniProt to Six Protein Data Sources to Produce Six *e-link*s of the same Link Tpye

5 A Data Model to Exploit *e-link*s

We present a structure to represent *e-link*s and then briefly outline a data model and query language.

5.1 A Structure for *e-link*s

We propose to enhance the current link implementation in life science sources to include semantic labels, link types, and a more precise identification of the link's source and target elements within the data entry. An *e-link* is a 5-tuple $\{l, P\_1, P\_2, S, lt\}$, as follows:

- l is the original link, either represented as an ID or as a hyperlink, or both.
- P_1 is a set of navigational paths that define the origin of the link in relation to the data entry storing l.
- P_2 is a set of navigational paths that define the target of the link in relation to the data entry storing l.
- S is a word or label from some controlled vocabulary, e.g., GO terms.
- lt is an element from a pre-defined set of link types LT.

5.2 Some Example *e-link*s from a Uni-Prot Entry

Consider the portion of a Uni-Prot entry MEFV_HUMAN with accession number O15553 in Figure 9.

```
      ID    MEFV_HUMAN STANDARD; PRT; 781 AA.
      AC    O15553; Q96PN4; Q96PN5;
      DT    16-OCT-2001 (Rel.  40, Created)
      DT    16-OCT-2001 (Rel.  40, Last sequence update)
      DT    29-MAR-2004 (Rel.  43, Last annotation update)
      DE    Pyrin (Marenostrin).
      GN    MEFV OR MEF.
      ...   ...
      OS    Homo sapiens (Human).
→1    OX    NCBI_TaxID=9606;
      RN    [1]
      ...   ...
      CC    -!- DISEASE: DEFECTS IN MEFV ARE THE
            CAUSE OF FAMILIAL MEDITERRANEAN
      CC        FEVER (FMF) [MIM:249100]...
      ...   ...
→2    DR    MIM; 608107;
→3    DR    MIM; 249100;
      ...   ...
→4    FT    VARIANT 694 694 M -> I (in FMF).
      ...   ...
```

Figure 9: Portions of the Uni-Prot entry O15553

The four lines marked with '→' result in four *e-link*s. We describe two of the *e-link*s in detail. We note that these links were generated manually and we briefly discuss techniques to support this process.

$\rightarrow^1$ The entry OX NCBI_TaxID=9606 will be enhanced to produce the 5-tuple
{-"- ; ./OS ; ./; is causal for disease; LinkType-ID-1}.
The first element -"- denotes the original physical link. The second element (./OS) shows that the origin of this link is not the Uni-Prot entry, but its OS attribute, i.e., the term Homo sapiens (Human). The third element (./) shows that the target of the link is unchanged. The fourth element (is causal for disease) is a label capturing the semantics of the link and finally, the fifth element LinkType-ID-1 identifies the link tpye in a pre-defined set of link types LT.

$\rightarrow^4$ The entry FT VARIANT 694 694 M -> I (in FMF). is enhanced to produce {-"- ;./FT[27]; ./AV.0002[1]; genetic background; LinkType-ID-4}.
In this example, the third link element (./AV.0002[1]) refines the link target to point to a certain part of the target element, namely the second allelic variant (AV). The link label is genetic background and the link type is LinkType-ID-4.

5.3 Next Steps in Exploring Enhanced Links

We briefly outline the data model, query language and evaluation engine for *e-links* and *e-paths*. We enhance the original data model presented in Section 2; it included logical classes, the *Source Graph*, the mapping from the logical classes to the *Source Graph* of Table 1 and the *Object Graph*. The data model will now include LT, a pre-defined set of link types. LS represents a set of links between the sources in *Source Graph*, and each element of LS will be of some link type in LT. LO, the set of object links between data entries in the *Object Graph*, will be enhanced and represented by *e-links*.

We will develop a semantics to compose link types in LT so that *e-links* can be traversed to produce meaningful paths, *e-paths*. We will also specify equivalences among *e-links* which may lead to equivalent *e-paths*. A query language will allow scientists to specify a query as a *labeled regular expression* [8] against the logical classes and the link labels of LT. The query language will also allow scientists to identify specific physical links of LS of interest, to be included or excluded during query evaluation.

We will develop a query evaluation engine which will first return a set of (uninstantiated) *e-paths* through the *Source Graph*, and then a set of (instantiated) *e-paths* through the *Object Graph* that retrieves participating *e-links* among the data entries.

Clearly, a data model alone is not enough since there already exist huge numbers of links that are not enriched. For example, the links among the NCBI data sources alone require several gigabytes of storage [4]. Hence, a next step is to explore methods to automatically or semi-automatically perform this enrichment by analyzing the linked data entries and by soliciting information from biologists via a user-friendly tool. An example of link generation and extraction of *e-links* capturing "marker semantics" from PubMed to the Human Genome ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ is described in [6]. We will also exploit existing efforts in ontology development to create an inventory of link semantics and link types that will list a set of possible semantic labels for links in LT, together with potential domains for the origin and target entries.

The benefits of *e-links* are numerous. First, the enhanced representation encodes useful knowledge. Second, biologists will be able to express properties such as the equivalence of *e-links*. Further, *e-links* will allow biologists to share semantics. Finally, query engines will be able to use *e-links* in answering queries and the enrichment will allow biologists to determine how relevant a particular *e-link* or *e-path* is to their own research or experiment.

Acknowledgements: This research is partially supported by NSF grants EIA0130422, IIS0222847 and IIS0430915. We thank Zoé Lacroix (Arizona State University), Alex Lash (Memorial Sloan Kettering Cancer Center), Ulf Leser (Humboldt-Universität zu Berlin), George Mihaila (IBM Laboratories) and Maria-Esther Vidal (Universidad Simón Bolívar)).

References

[1] *http://www.bioinf.org.uk/pdbsprotec/.*

[2] *http://www.geneontology.org/.*

[3] *http://www.ebi.ac.uk/GOA/.*

[4] *Private Communication with E. Yaschenko, NCBI*, 2004.

[5] T. Etzold. The bioinformatics data integration tunnel: Do we see the light yet? *Keynote Presentation, DILS*, 2004.

[6] A. Lash, L. Raschid, and A. Lee. A protocol to extract and generate links capturing marker semantics. *In preparation*, 2004.

[7] A. Martin. Pdbsprotec: A web-accessible database linking pdb chains to ec numbers via swissprot. *Bioinformatics*, 20(6):986–988, 2004.

[8] G. Mihaila, L. Raschid, and M.E. Vidal. A data model and labeled regular expression query language for enhanced navigational queries in life science sources. *In preparation*, 2004.