Capturing Provenance for a Linkset of Convenience

Simon Jupp[1], James Malone[1], and Alasdair J G Gray[2]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, United Kingdom
[2] Department of Computer Science, Heriot-Watt University, Edinburgh, United Kingdom

Abstract. Biological interactions such as those between genes and proteins are complex and require intricate OWL models. However, direct links between biological entities can support search and data integration. In this paper we introduce *linksets of convenience* that capture these direct links. We show the provenance statements required to track the derivation of such linksets; linking them back to the full biological justification.

Keywords: Data linking, Provenance, VoID

1 Introduction

Investigating biological systems, such as those implicated in disease, necessitates the connection of many levels of biology; gene, gene variation, gene expression, protein structure, signalling pathways, phenotypic, epidemiological data and so on. The ability to integrate data across these levels relies on links that can be formed between biological entities, for example, going from a gene to proteins or proteins to pathways. For each of these links there is some biological justification that may involve several steps (see Section 2 for details). To support tasks such as search and data integration it is convenient to provide additional shortcuts in the form of a direct link, e.g. genes to pathways.

Modeling the true nature of the links using semantic web technologies such as OWL removes ambiguity when working with data by giving it a well defined and precise semantics. However it increases the complexity of interacting with the data as the OWL model needs to capture the full intricacies of the biological interactions. As we move to publish biological data as linked open data, there is an opportunity to describe direct links between different types of biological entities as a shortcut to be made between entities which feature in common queries, such as gene to protein; capturing the way that biologists often discuss the domain and enable novel integrations of the data. These direct links provide a working notion that cuts through the biology but which does not necessitate capturing (or recapturing) the complex multivariate relationships that can hold between the two entities. Such linksets are already used to support the Open

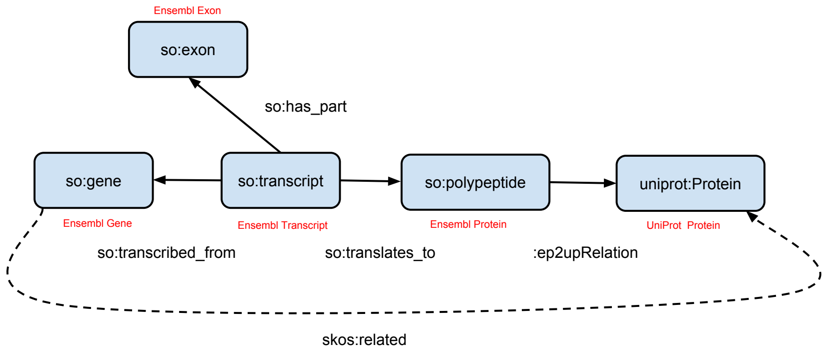

Fig. 1. Linking an Ensembl gene with its UniProt protein. Solid lines show the full semantic modelling required while the dashed line represents the linkset of convenience.

PHACTS Discovery Platform [1], although those linksets do not have adequate provenance.

In this paper we propose a mechanism to model these *links of convenience* using a combination of VoID linksets [2] and PROV [3]. We avoid misrepresenting links by applying semantically weaker relationships together with additional provenance which represents the underlying complexity. We illustrate the model with an example using data from two popular biological databases.

2 Linking genes to proteins use case.

We motivate our work with an example mapping between Ensembl [4] (a database of genome annotation) and Uniprot [5] (a database of protein sequences). These databases already contain cross-references between an Ensembl Gene (EG) and a Uniprot Protein (UP). However to understand how this mapping is generated you currently need to discover the correct publications and online documentation; they are not directly discoverable from the data.

Biological theory tells us that a gene encodes for a protein, although this biological relation only truly holds for the link between the EG and the Ensembl Protein (EP) entity. There are in fact multiple types of UP to EP mappings, for instance they can be derived from an exact sequence identity or they might be based on a percentage sequence identity. Figure 1 illustrates how we model EG to EP using terminology defined in the Sequence Ontology, and for illustration we include a superproperty of the all the EP to UP mappings that we call `ep2upRelation`[3]. We introduce a link of convenience (dashed line) that links the EG to UP that is there to support queries using the semantically weak `skos:related` relation. This schema lacks the provenance to assert that the related link of convenience is derived from the longer chain of semantically richer links that hold from a gene to protein.

[3] UniProt are currently extending their vocabulary to define these relations.

```
# define the ensembl protein partition
:ensembl void:classPartition :EPpartition .
:EPpartition void:class so:Polypeptide .

# define the Uniprot protein partition
:uniprot void:classPartition :UPpartition .
:UPpartition void:class uniprot:Protein .

# define the linkset that links the two partitions
:ensemblProteinToUniprotProteinLinkset a void:Linkset ;
    void:linkPredicate :ep2upRelation ;

# define partitions for ensembl gene, gene transcript and
# transcript protein
:ensembl  void:classPartition :ensemblGenePartition ;
    void:propertyPartition :ensemblGeneTranscriptPartition ;
    void:propertyPartition :ensemblTranscriptProteinPartition ;
:ensemblGenePartition void:class so:gene .
:ensemblGeneTranscriptPartition void:property so:transcribed_from .
:ensemblTranscriptProteinPartition void:property so:translates_to .

# define the linkset that links the two partitions,
# including the dataset description that contains the triples that
# are used to derive this linkset
:ensemblGeneToUniprotProteinLinkset a void:Linkset ;
    void:linkPredicate skos:related ;
    void:subjectsTarget :ensemblGenePartition;
    void:objectsTarget :UPpartition;
    prov:wasDerivedFrom :ensemblGeneTranscriptPartition,
        :ensemblTranscriptProteinPartition,
        :ensemblProteinToUniprotProteinLinkset
```

Fig. 2. Description of the linkset of convenience between Ensembl Gene and UniProt Protein which includes the provenance derivation.

3 Describing Linksets

The model outlined in Figure 1 can be decorated with provenance that captures additional information about how the link of convenience between EG and UP is derived. The resulting linkset description is shown in Figure 2. In the following we describe the blocks of RDF.

The VoID vocabulary of linked datasets allows the description of RDF links between datasets using VoID linksets. A linkset allows us to describe the links, captured as a set of triples, between two datasets. We can use VoID to describe relevant partitions of the datasets based on individual properties or classes, these form new subsets that can participate in multiple linksets. In our scenario we need to capture two crucial linksets; the first is the EP to UP linkset, and the second is the more convenient EG to UP linkset.

The EP-UP linkset captures the `:ep2upRelation` link between types of EP in the Ensembl dataset, and types of UP in the UniProt dataset (lines 10-11). We describe two further subsets; the EP partition of all entities that are of type `so:Polypeptide` in the Ensembl dataset (lines 2-3) and the UniProt subset of all entities that are of type `uniprot:Protein` (lines 6-7).

The EG to UP link of convenience needs a similar linkset description based on an EG partition and the previous UP partition, although this time the relation is `skos:related` (lines 25-26). We also want to capture that the triples in this linkset are derived from another set of triples. This captures that the `skos:related` is a shortcut relation for a more complex path through the RDF graph. Again we can use VoID partitioning, but this time using a property based partition to identify the EG to Ensembl Transcript (ET) and ET to EP links (lines 15-20) . Finally we use the `prov:wasDerivedFrom` relation to link the convenience linkset to the linksets that describe the full path of relations that the shortcut represents (line 28-30).

4 Discusion

It is always important to try and model your data as accurately as possible, and publishing data with RDF and OWL is well suited for this task. The VoID vocabulary already provides a mechanism to define and attach provenance to linksets between datasets, and we are proposing the use of PROV to connect linksets that are derived from other linksets. As a Web of linked biological data emerges, there is a need to identify links that are there for convenience, and expose how they relate back to the core biological (OWL) model. In cases where a link of convenience is derived from a series of other linksets, it is desirable to be able to spot this and unpack the convenience links using common queries. The model proposed supports this task but questions remain as to whether VoID and PROV are enough, so we hope this preliminary work can help motivtate the discussion.

Acknowledgements

EBI contribution supported by EU FP7 BioMedBridges Grant 284209.

References

1. Gray, A.J.G., Groth, P., Loizou, A., Askjaer, S., Brenninkmeijer, C.Y.A., Burger, K., Chichester, C., Evelo, C.T., Goble, C.A., Harland, L., Pettifer, S., Thompson, M., Waagmeester, A., Williams, A.J.: Applying linked data approaches to pharmacology: Architectural decisions and implementation. Semant. Web **5** (2014) 101–113
2. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets with the VoID Vocabulary. Note, W3C (March 2011)

3. Lebo, T., Sahoo, S.S., Mcguinness, D.: PROV-O: The PROV Ontology. Technical report, W3C Recommendation (2013) `http://www.w3.org/TR/prov-o/`.
4. Flicek, P., Amode, M.R., Barrell, D., et al: Ensembl 2014. Nucleic acids research **42** (2014) D749–D755 doi: 10.1093/nar/gkt1196.
5. The UniProt Consortium: Activities at the universal protein resource (UniProt). Nucleic acids research **42** (2014) D191–D198 doi: 10.1093/nar/gkt1140.