Provenance and Linked Data in Biological Data Webs

Jun Zhao, Graham Klyne and David Shotton
Image Bioinformatics Research Group
Department of Zoology
University of Oxford
Oxford OX1 3PS, UK
{jun.zhao,graham.klyne,david.shotton}@zoo.ox.ac.uk

ABSTRACT

To created a linked data web of heterogeneous biological data resources, we need not only to define and create the alignment between related data resources but also to express the knowledge about why data items from different sources are linked with each other and how each data link has evolved, so that scientists can trust the data links provided by the data web. This paper highlights the importance of keeping provenance information about the links between data items from different sources, and proposes the use of named graphs to make a provenance statement about each pair of linked data items and each release of a data web.

Categories and Subject Descriptors

D.2.12 [**Software Engineering**]: Interoperability—*Data mapping*; H.3.5 [**Information Storage And Retrieval**]: Online Information Services

General Terms

Language, Reliability

Keywords

Data Web, Named Graphs, Provenance, RDF, Semantic Web, Trust

1. INTRODUCTION

The number of biology databases available has increased rapidly in the recent years [4]. To obtain knowledge about a gene or protein from this sea of data, biologists often need to go through an information gathering process, navigating between the public genomic and publication databases. These resources are scattered around the world and present data in heterogeneous formats. Scientists have to rely on their domain knowledge in order to identify how data resources are linked with each other.

To simplify this process, the Image Bioinformatics Research Group (IBRG)[1] of the University of Oxford proposes the use of *subject-specific* data webs, which use the Web as the native platform upon which to integrate access to datasets relating to particular subjects [7]. Within each data web, data resources are integrated using loosely coupled software tools that permit both information discovery and links back to the original data. With this approach, the data linked into the data webs are neither required to be semantically coordinated nor constrained to conform to a single imposed model. Furthermore, copyright and access control issues remain the concern of the data sources, not of the data web that unites them. These data sources maintain their unique characters and continue independent publication of their holdings.

The first demonstrator data web being developed by IBRG is FlyWeb[2], which will integrate the heterogeneous data resources concerning research on fruit fly *Drosophila melanogaster*. These data resources include FlyTED[3] (our local research image repository concerning gene expression in the testis of fruit flies), BDGP[4] (the Berkeley *Drosophila* Genome Project database concerning gene expression in the *Drosophila* embryos), FlyBase[5] (the global database of genomics information concerning *Drosophila*), and online research publications on *Drosophila* gene expressions. The goal of FlyWeb is to allow biologists to obtain information about a *Drosophila* gene, including the gene expression images of its testis and embryos, without having to hop between the *Drosophila* data islands on the Web.

To build FlyWeb, we need not only to define and implement the alignment between *Drosophila* data resources, but also to *maintain* the data links between related data items from different sources. This position paper focuses on the second issue, and will analyze the motivation for keeping provenance of the links between related data items and present our proposed solution.

2. SEMANTIC WEB AND FLYWEB

The initial development goals of the FlyWeb Project include understanding the distributed *Drosophila* data resources; creating the alignment between them; and creating a query service to access the integrated data resources. At the time

[1] http://ibrg.zoo.ox.ac.uk/
[2] http://imageweb.zoo.ox.ac.uk/wiki/index.php/FlyWeb_project
[3] http://www.fly-ted.org/
[4] http://www.fruitfly.org/
[5] http://flybase.org/

of the writing, concentrating first on linking the FlyTED and BDGP databases, we have achieved:

- Describing the *Drosophila* testis images in FlyTED using an extension to the Fly Anatomy Ontology [1], which is also used by BDGP to describe its *Drosophila* embryo gene expression images.
- Publishing FlyTED, the *Drosophila* testis gene expression image database, through a SPARQL endpoint [9], the same interface used by BDGP for publishing its gene expression images and annotations [5].
- Identifying the relationships between FlyTED and BDGP using the genomic knowledge, particularly the gene names, captured in FlyBase.

These initial works provide the foundation that permits us to align the two data resources and build a lightweight data web. However, the evolving nature of biological databases has motivated us to further consider how to manage the links in FlyWeb, once they have been established.

3. MOTIVATION

Data items from different *Drosophila* data resources are integrated into FlyWeb using references to the original data. Related data items are linked together in FlyWeb using biological knowledge from public genomic databases. Biological knowledge is growing rapidly, and genomic databases are frequently updated. By referencing back to these evolving databases, FlyWeb can synchronize with advances in biological knowledge. However, with each update of such an external resource, some of the links between data items recorded locally within FlyWeb may become obsolete or need to be updated with more links to related data items. We need to provide additional metadata about these data links in order to maintain consistency between FlyWeb and the advancing biological knowledge. This will allow scientists to:

- Trust that the data links established in FlyWeb are valid;
- Trust that the data referenced in FlyWeb are consistent with the latest release of the public databases;
- Trace back the data links established by FlyWeb using previous releases of the public databases, which may previously have been used by the scientists to annotate their own local data.

Thus, for each data link to a pair of data items, we need to record the following *provenance metadata*:

- The evidence of the link;
- When this link was created, by whom, using which version of which database;
- When this link was updated or deprecated;
- Whether there were any previous links between this pair of data items;
- What previous links between data items became obsolete, and why.

To express this provenance of data links, we propose to use named graphs.

4. NAMED GRAPHS

An RDF graph contains a collection of RDF triples. A named graph is an RDF graph which is assigned a name in the form of a URI [3]. It provides a way to group RDF statements into sub-graphs that may be asserted separately, and it also provides names for such graphs. By grouping and naming RDF statements as a named graph, applications can state access control rights, copyright, or provenance information about these RDF statements as a whole. Thus, named graphs provide a mechanism for establishing trust within the Semantic Web. More generally, this mechanism allows us to make statements about the content of the graph without asserting that the statements contained in the graph are true.

In order to provide information about why a pair of related data items are linked together in FlyWeb, or why/when they become no longer linked, we create a named graph for each pair of linked data items. In this position paper, we only consider two types of links between data items, *i.e.* either they are same as or different from each other. There may be other types of data links in FlyWeb. But the provenance model introduced in this paper is not yet designed for describing all different types of data links.

FlyWeb will be updated whenever a major release of the linked-in *Drosophila* databases is announced. To provide information about each FlyWeb release and the versions of the public databases upon which each release is based, we will also create a named graph for each release of the FlyWeb.

In this position paper we use TriG as the notation to define named graphs. "TriG is a variation of Turtle [2] which extends that notation by using '{' and '}' to group triples into multiple graphs, and to precede each by the name of that graph" [3].

4.1 A Named Graph for Each FlyWeb Release

FlyWeb integrates several *Drosophila* data sources, noted as a, b, c, etc. Each data source is associated with version information. Thus a_x indicates version x of data source a.

Each release of FlyWeb (fw_g, fw_k, etc) will contain a collection of data items, i_m, i_n, etc, from different *Drosophila* data sources. A data item from data source a of version x should be uniquely identified as $i_m(a_x)$. In fw_g, $i_m(a_x)$ will be described by all the metedata from its original data source, as well as by FlyWeb statements about whether it is related to another data item $i_n(b_x)$ from data source b_x, or whether $i_m(a_x)$ had previously been linked with $i_n(b_x)$ in a previous release of FlyWeb.

Each release of FlyWeb itself is a named graph, which is associated with information about when it was released, by whom, using which versions of which databases. An example of two such named graphs is given below.

The following two examples show two graphs (see Figure 1). Example 1 tells information about FlyWeb version 1.0 (`<dwi:flyweb_r1>`) that it was released on "2007-12-19" and it was built using FlyTED version 1.0, the BDGP database version "2007-03-09" and FlyBase version 4.3. It also contains a single statement about data items, *i.e.* the gene `<flyted:gene_g1>` from FlyTED is the same as the gene `<bdgp:gene_g2>`[6] from BDGP. Example 2 defines a

[6]The `bdgp` namespace might not be the actual namespace used by the BDGP SPARQL endpoint. Due to technical maintenance, its server was unreachable when the paper was written.

named graph of FlyWeb version 1.1, which was built using the same versions of FlyTED and BDGP as FlyWeb version 1.0, but a different version of FlyBase. Because of this update, gene `<flyted:gene_g1>` is no longer the same as `<bdgp:gene_g2>`.

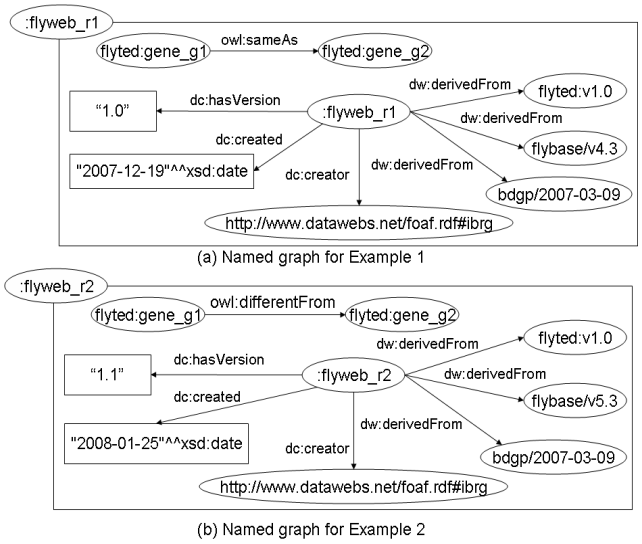

Figure 1: Two named graphs for Example 1 and 2.

EXAMPLE 1. *Named graph for FlyWeb release 1.0*

```
@prefix dw: <http://www.datawebs.net/> .
@prefix flyted: <http://id.fly-ted.org/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix bdgp: <http://www.fruitfly.org/> .
@prefix dwi: <http://id.datawebs.net/> .
@prefix : <http://id.datawebs.net/> .

:flyweb_r1 {
  flyted:gene_g1  owl:sameAs bdgp:gene_g2 .
  :flyweb_r1  dc:created  "2007-12-19"^^xsd:date;
              dc:hasVersion "1.0"   ;
              dc:creator
        <http://www.datawebs.net/foaf.rdf#ibrg> ;
              dw:derivedFrom flyted:v1.0 ;
              dw:derivedFrom <bdgp/2007-03-09> ;
              dw:derivedFrom <flybase/v4.3> .
}
```

EXAMPLE 2. *Named graph for FlyWeb release 1.1*

```
@prefix dw: <http://id.datawebs.net/> .
@prefix flyted: <http://id.fly-ted.org/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix bdgp: <http://www.fruitfly.org/> .
@prefix dwi: <http://id.datawebs.net/> .
@prefix : <http://id.datawebs.net/> .

:flyweb_r2 {
  flyted:gene_g1 owl:differentFrom bdgp:gene_g2 .
  :flyweb_r2  dc:created "2008-01-25"^^xsd:date;
              dc:hasVersion "1.1"   ;
              dc:creator
        <http://www.datawebs.net/foaf.rdf#ibrg> ;
              dw:derivedFrom flyted:v1.0 ;
              dw:derivedFrom <bdgp/2007-03-09> ;
              dw:derivedFrom <flybase/v5.3> .
}
```

4.2 A Named Graph for Each Data Link

In each FlyWeb named graph, a collection of named graphs are also created for the data links between pairs of related data items. Each such named graph states:

- Why a pair of data items should be or should no longer be linked;
- When the link was made or released, and by whom;
- Which previous links had been created between this pair of data items;
- What the type the link is: a `MappingRelation`, either a `SameRelation` or a `DifferentRelation`. The last two concepts will be defined in a data web ontology using the `owl:sameAs` and `owl:differentFrom` properties.

Example 3 (see Figure 2) shows a named graph `<dwi:mapping_m1>` that defines an abstract relationship between the gene from FlyTED (`<flyted:gene_g1>`) and the gene from BDGP (`<bdgp:gene_g2>`) and traces this relationship by its two children, both of which are themselves named graphs and define the actual relationships between these two genes built in different releases of FlyWeb.

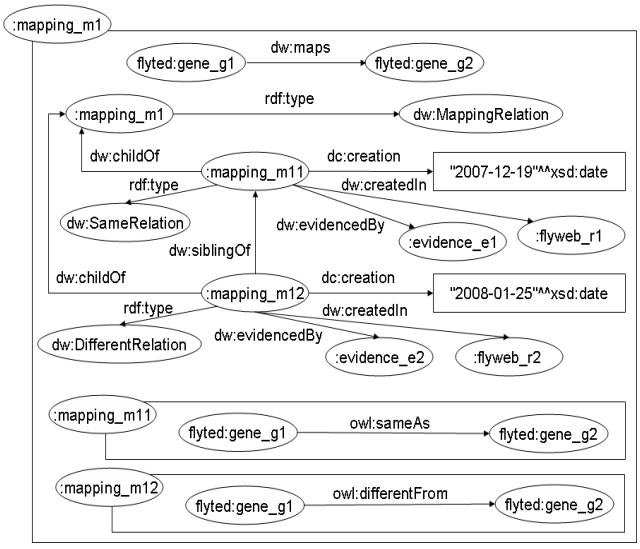

Figure 2: The named graph for Example 3.

The first child `<dwi:mapping_m11>` defines that the two genes are synonyms given the evidence of `<dwi:evidence_e1>` and that this link was created on "2007-12-19" within the release of `<dwi:flyweb_r1>`. The second child `<dwi:mapping_m12>` states that the two genes are not the same given the evidence of `<dwi:evidence_e2>`, and that this link was created on "2008-01-25" within the release of `<dwi:flyweb_r2>`.

The dw:childOf property links <dwi:mapping_m11> and <dwi:mapping_m12> with the graph <dwi:mapping_m1>, and they are linked together by the property dw:siblingOf. These properties enable us to trace a lineage of the data links between a pair of data items.

EXAMPLE 3. *Named graph for a data link*

```
@prefix dw: <http://www.datawebs.net/>           .
@prefix flyted: <http://id.fly-ted.org/>         .
@prefix dc: <http://purl.org/dc/elements/1.1/>   .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#>    .
@prefix bdgp: <http://www.fruitfly.org/>         .
@prefix dwi: <http://id.datawebs.net/>           .
@prefix : <http://id.datawebs.net/>              .
@prefix rdf:
   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .


:mapping_m1 {
   :mapping_m1     rdf:type   dw:MappingRelation .
   flyted:gene_g1  dw:maps         bdgp:gene_g2  .
   # the first child
   :mapping_m11  dw:childOf      :mapping_m1     ;
                 dw:evidencedBy  :evidence_e1    ;
                 dw:createdIn    :flyweb_r1      ;
                 rdf:type        dw:SameRelation ;
                 dc:creation
                          "2007-12-19"^^xsd:date .
   :mapping_m11 {
         flyted:gene_g1  owl:sameAs bdgp:gene_g2 .
   }

   # the second child
   :mapping_m12  dw:childOf     :mapping_m1      ;
                 dw:evidencedBy :evidence_e2     ;
                 dw:siblingOf   :mapping_m11     ;
                 dw:createdIn   :flyweb_r2       ;
                 rdf:type
                           dw:DifferentRelation ;
                 dc:creation
                          "2008-01-25"^^xsd:date .
   :mapping_m12 {
   flyted:gene_g1  owl:differentFrom bdgp:gene_g2 .
   }
}
```

5. SCENARIOS

This section uses the above example datasets to walk through three scenarios to show how the named graphs could help us to manage the data links in FlyWeb in a manner that promotes trust.

5.1 Links in a Previous Release

The first scenario shows how FlyWeb can help users to find out which data items in FlyWeb are linked to their gene, which is annotated using information from FlyBase release 4.3.

Many biology data compilations are maintained locally by research groups and might not always be kept up-to-date with successive releases of the genomic database FlyBase due to the ending of the projects that funded them. Such local legacy data will have been annotated using information from a now out-of-date version of the public database. Subsequent releases of the public database might have annotated its gene records using different gene names. Occasionally, new biological evidence shows that a particular DNA sequence, formerly thought to be a single gene and given a single gene name, in fact encodes two distinct genes that are then given different names.

Without provenance data, users would not be able to find in FlyWeb any data relating to their locally recorded former gene names, because the genes are now annotated with new names. In order to prevent this situation in the future, we provide provenance information for each release of FlyWeb, to state which versions of the public databases it links to. This provides the flexibility for the scientists to trace data links for their legacy data. A SPARQL query [6] for this scenario is shown below, which will search for all the data items that are linked to the gene <flyted:gene_g1> in the release of FlyWeb that was built using FlyBase version 4.3.

```
SELECT *
WHERE { ?g dw:derivedFrom <flybase/v4.3>
        graph ?g {
            { flyted:gene_g1 ?p ?data }
            UNION
            { ?data1 ?p1 flyted:gene_g1 } }
}
```

5.2 All Links in the Latest Release

This scenario shows how users can navigate information about a *Drosophila* gene in the latest release of FlyWeb using the version information and the creation date associated with the named graph of each release of FlyWeb. The following SPARQL query will retrieve all the data links from the v1.1. release of FlyWeb.

```
SELECT *
WHERE { ?g dc:hasVersion "1.1" .
        graph ?g {?gene1 ?p ?gene2 } }
```

5.3 Explaining Conflicts

One way of allowing users to trace the data links between a pair of related data items is to keep a history of all the data links that have ever existed between them. This means that conflicting statements about the relationship between the same pair of data items might exist in different releases of FlyWeb. In order to explain these conflicts, we provide the evidence information for the data links.

Example 1 and Example 2, describing release1.0 and 1.1 of FlyWeb, contain conflicting statements about the relationships between <flyted:gene_g1> and <bdgp:gene_g2>. In order to explain this conflict, we need to take the following steps:

- Retrieve all the statements about the data link between <flyted:gene_g1> and <bdgp:gene_g2> from different releases of FlyWeb. This will return all the statements about the graphs <dwi:mapping_m11> and <dwi:mapping_m12> that define the relationships between the two gene names;
- Compare the statements about these two graphs in order to find out the differences between the two versions of relationships between <flyted:gene_g1> and <bdgp:gene_g2>;

- Present the differences resulting from the above comparison step to the users, including their creation date, in which release of FlyWeb they were created, as well as the evidence for explaining why each different relationship existed between `<flyted:gene_g1>` and `<bdgp:gene_g2>`.

A SPARQL query for the first step would be:

```
CONSTRUCT {?cg ?p ?o}
WHERE {
   graph ?g {flyted:gene_g1 ?p1 bdgp:gene_g2 .
             ?g rdf:type dw:MappingRelation .
             ?cg  dw:childOf ?g .
             ?cg    ?p     ?o}
}
```

6. CONCLUSIONS

In this position paper we have analyzed how recording the provenance of data links can help us both maintaining the links between related data items and bringing trust to the data web, by providing evidence for links, or tracing how the data links have been updated and maintained. We have shown the potential of named graphs for expressing this provenance information. The flexibility of RDF named graphs and the RDF query language SPARQL provide the capability for us to query and filter the data links on behalf of the data web users, *e.g.* by presenting only those links newly created since the previous release of FlyWeb, or those present in a particular earlier release of FlyWeb.

When defining this conceptual provenance model, we have adopted existing vocabulary as much as possible, such as the properties of `dc:creation` and `dc:creator` from the Dublin Core Metadata Element Set[7]. We have also used the `dw` namespace (`http://www.datawebs.net/`) to specify the following properties of our own:

- `dw:derivedFrom`
- `dw:evidencedBy`
- `dw:childOf`
- `dw:siblingOf`
- `dw:createdIn`
- `dw:maps`

We are planning to include these conceptual properties in a data web provenance ontology, that will include other existing vocabularies about provenance and trust.

In this conceptual model we associated with each data link a `dw:evidencedBy` property to provide the information about why particular statements were asserted. This will bring trust to the linked data for the scientists, so that they can verify that the links are consistent with scientific knowledge. However, we are still investigating how much information should be provided as evidence for each data link: whether it should contain the actual heuristic used for building the links or a textual description of this heuristic; and how we can make this evidence information more comprehensible for biological researchers.

There is a separate provenance issue that is not discussed in this position paper, namely the provenance of the data items themselves. We discussed neither the provenance information for telling where each data item came from nor the provenance information that might be associated with a data item from the individual data resource. These are key research topics for Semantic Web and provenance for life sciences [3, 8]. The datasets published by BDGP through their SPARQL endpoint have been annotated with some provenance and evidence information [5]. Those data provenance statements will be integrated into FlyWeb along with all other descriptions concerning the data. We need to research how this provenance of data can best be incorporated into FlyWeb, together with the provenance of the data links.

[7] `http://www.dublincore.org/documents/dces/`

7. ACKNOWLEDGEMENT

This work is supported by funding from the JISC (FlyWeb Project to Dr David Shotton; `http://imageweb.zoo.ox.ac.uk/wiki/index.php/FlyWeb_project`) and from BBSRC (Grant `BB/E018068/1`, The FlyData Project: Decision Support and Semantic Organisation of Laboratory Data in *Drosophila* Gene Expression Experiments, to Drs David Shotton and Helen White-Cooper). The FlyTED Database was developed with funding from the UK's BBSRC (Grant `BB/C503903/1`, Gene Expression in the *Drosophila* testis, to Drs Helen White-Cooper and David Shotton). Preliminary data web requirements analysis was supported by a JISC grant to Dr David Shotton (Defining Image Access Project; `http://imageweb.zoo.ox.ac.uk/wiki/index.php/DefiningImageAccess`).

8. REFERENCES

[1] M. Ashburner and *et al.* A structured controlled vocabulary of the anatomy of *Drosophila melanogaster*. `http://obofoundry.org/cgi-bin/detail.cgi?id=fly_anatomy`.

[2] D. Beckett. Turtle - Terse RDF Triple Language, 2007. `http://www.dajobe.org/2004/01/turtle/`.

[3] J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named graphs, provenance and trust. In *Proc. of the 14th International World Wide Web Conference*, pages 613–622, Chiba, Japan, 2005. `http://doi.acm.org/10.1145/1060745.1060835`.

[4] M. Y. Galperin. The molecular biology database collection: 2008 update. *Nucleic Acids Research*, 36(Database issue):2–4, 2008. `doi:10.1093/nar/gkm1037`.

[5] C. Mungall. A SPARQL endpoint for a database of annotated gene expression. `http://www.bioontology.org/wiki/index.php/OBD:SPARQL-InSitu`.

[6] E. Prud'hommeaux and A. Seaborne. SPARQL query language for RDF, January 2008. W3C Recommendation. `http://www.w3.org/TR/rdf-sparql-query/`.

[7] D. Shotton. *World Wide Science: Promises, Threats and Realities*, chapter Data webs for image repositories. Oxford University Press, 2008. in press.

[8] J. Zhao, C. Goble, R. Stevens, and D. Turi. Mining Taverna's Semantic Web of Provenance. *Journal of Concurrency and Computation:Practice and Experience*, 2007. `doi:10.1002/cpe.1231`.

[9] J. Zhao, G. Klyne, and D. Shotton. Building a semantic web image repository for biological research images. In *Proc. of the 5th European Semantic Web Conference*, Tenerife, Spain, 2008. accepted.