Semantic Data Integration for *Francisella tularensis novicida* Proteomic and Genomic Data

Nadia Anwar[1], Ela Hunt[2], Walter Kolch[3] and Andrew Pitt[1]

[1]Division of Integrative and Systems Biology
Faculty of Biomedical and Life Sciences
University of Glasgow, Glasgow G12 8QQ
[2]Department of Computer and Information Sciences,
University of Strathclyde, Glasgow G1 1XB
[3] Beatson Institute for Cancer Research
Glasgow, G61 1BD

Abstract. This paper summarises the lessons and experiences gained from a case study of the application of semantic web technologies to the integration of data from the bacterial species *Francisella tularensis novicida* (*Fn*). *Fn* data sources are disparate and heterogeneous, as multiple laboratories across the world, using multiple technologies, perform experiments to understand the mechanism of virulence. It is hard to integrate such data, and this work examines the role of explicitly provided data semantics in data integration. We test whether the semantic web technologies could be used to reveal previously unknown connections across the available *Fn* datasets. We combined this data with genome data and with public domain annotations within GO, KEGG and the SUPERFAMILY database. Through this connected graph of database cross references, we extended the annotations of an experimental data set by superimposing onto it the annotation graph. Identifiers used in the experimental data automatically resolved and the data acquired annotations in the rest of the RDF graph. This happened without the expensive manual annotation that would normally be required to produce these links. Other lessons learnt and future challenges that result from this work are also presented in detail.

1 Introduction

The ultimate goal in biology is to understand how genes translate to phenotypes, i.e., to understand the complex relationship between genes, messenger RNAs, proteins and tissues. Current research methodologies study these components individually through genomic, transcriptomic, proteomic and metabolomic experiments. Since these technologies require specific expertise for data generation and analysis, it is quite rare to find experiments that are performed using all of these technologies on the same sample. Also, attempts at correlating data across these technologies have so far been disappointing, indicating that measures of transcript level in microarray experiments is not a good indicator for protein abundance [1]. Yet, fundamentally, we know that genes, transcripts, proteins and all the processes performed in cells form a complex system that requires each of these and many other components to function collaboratively. There are significant benefits to be gained by combining data across experiments and individual technologies [3, 4]. It is a long term goal in biology to understand the system as a whole, and, in the short term, combined access to these data can corroborate the predictions and validate discoveries made by each technology.

The post genomic technologies such as transcriptomics [5] and proteomics [6] have strengths and weaknesses [7, 4]. Given the central dogma, in theory, when these data are used together, false discovery rates inherent within those weaknesses could be reduced. The addition of corroborating data can be used to validate predictions that are on the edge of the statistical thresholds. For example, in proteomics experiments, a protein with three or more peptide hits is used as a reliable threshold for identification [8], and many proteins identified by the presence of a single peptide hit are discarded. These singleton hits can be supported with transcription abundance data or metabolomic data if the peptide hit falls into a pathway where metabolites reliant on the protein of interest have been identified. In another example, identification of quantitative trait loci (QTLs) through linkage analysis can identify regions in the genome attributed to a particular polygenic trait (disease phenotype): these regions will contain many genes, some of which are responsible for the trait and some which are not and expression QTL (eQTL) [9] analysis reduces the number of candidate genes in a QTL interval through the addition of gene expression profiles [10]. It has successfully identified genes associated with complex human diseases such as asthma [11].

Combining data from various technologies is very difficult and is very often done by hand. The source data that are produced and stored independently of each other. Thus, gathering data on a particular pathway, organism or disease generated from these technologies requires collecting and combining these data into spreadsheets, or using database software. Once these data are gathered from individual data sources, subsequent downstream analysis may be required, such as statistical tests for clustering or correlating data, or specialist algorithms that can compare the data [1]. In order for these data to be used more effectively the data at each level of analysis need to be readily accessible and easily combined. Data integration, however, is not trivial and requires resolving syntactic, structural and semantic differences across the data sources. The heterogeneity with respect to syntactic differences includes the differences in the data models such as relational databases, object stores, XML stores, flat files or spreadsheets. Structural differences lie in the data schemas that each source specifies and the query languages that they support. Semantic differences are expressed in the terminologies (vocabularies) they recognise. The methodologies that are employed to overcome these problems have so far proved to be difficult to reproduce on alternative data sets and they remain to be difficult to maintain and automate. Also, since database heterogeneity is unavoidable, and a single data model for all biomedical data is neither probable nor possible, we require a mechanism to integrate data in an automated, scalable and flexible way.

In this paper a new solution to flexible data integration is being examined. Semantic integration based on RDF [2] is being tested on omics data generated for the organism *Francisella tularensis novicida (Fn).* Fn is a gram negative bacterium that causes the plague like disease tularemia. In the most highly virulent subspecies, *Francisella tularensis tularensis* (Ft) only 10-50 bacteria are required to cause infection in humans. Ft is much more infectious than the bacterium *Bacillus anthracis*, anthrax, which has been used as a biological weapon. The ability to cause severe disease, the low infectious dose and the bioterrorism concerns posed by this organism have led to the availability of increased research funds [12]. While highly infectious in mice, *Francisella tularensis novicida* (Fn) strain U112 is a less virulent subspecies infecting only immunocompromised humans. This has allowed this subspecies to be well studied in the laboratory. The genome of all four subspecies have been sequenced and compared [13]. Additionally, numerous transcriptomic and proteomic experiments have been performed on Fn.

Of particular interest is the Francisella pathogenicity island (FPI) and the MglA transcriptional regulator. The FPI is a 30Kb region containing 16-19 genes whose functions remain unknown and are essential for growth within macrophage cells. Macrophages are free floating cells within the vascular system and are a part of the innate immune response. Their role is to engulf and digest pathogens in a phagolysosome, an organelle containing digestive enzymes. Normally these cells are a hostile environment for pathogens such as Francisella. However, Francisella is able to survive and replicate in macrophages by escaping the phagolysosome into the macrophage cytosol where they can replicate and ultimately escape, causing cell death. Experimental evidence shows that escape from the phagolysosome is reliant on genes encoded within the FPI [14].

In addition to the FPI, research has focused on a spontaneous mutant that is unable to disrupt the phagolysosome and replicate in the cytosol. The gene that was disrupted in this mutant was named MglA (Macrophage growth locus A). The product of this gene regulates the transcription of genes within the FPI and approximately 90 other genes. In an attempt to understand how MglA controls the transcription of virulence factors, many proteomic [16] and transcriptomic [17] experiments have been performed.

In the majority of published studies experimental data sources are analysed manually and data elements are manually linked to online data sources. More efficient analysis can be performed by the biologists if available online data could be easily integrated with experimental data. However, annotating every experiment, to the same extent as a genome, is very rarely performed due to time constraints. Biologist are therefore working with only part of the picture. We propose here a semantic data integration solution that would facilitate integration of online Fn data sources with individual experimental data sets in a simple and efficient manner.

The rest of this paper proceeds as follows. In Section 2 we provide some background on data integration methods and in Section 3 we broaden this with selected biological applications of data integration. Section 4 presents our solution which combines the graphs of linked data, and in Section 5 we discuss our results and observations. Then we conclude.

2 Data Integration

The goal of a data integration system is to provide uniform access to a set of heterogeneous data sources, and to free the user from the knowledge about how data are structured at the sources and how they are to be reconciled in order to answer queries. Data integration is most commonly achieved using one of three approaches: application integration (mediation), database federation and data warehousing [18].

Application integration involves writing special purpose software agents [19] that can query individual data sources via a single interface and then combine and return the results to the user. However, these applications can be fragile and expensive to maintain. Since integration is coded into the applications that are initially inexpensive and simple to build, these systems are notoriously fragile and susceptible to changes in the underlying systems that are integrated. Adding new data sources often requires the application to be completely rewritten. Very little integration is actually achieved through this approach. The data sources remain autonomous, queries are performed locally and the results that are gathered are combined and returned to the user. Therefore, if analysis or comparison of the data received is required this needs to be coded into the application. Portals offer another approach that is similar to application integration [20]. Usually, portals use web services to facilitate cross-database queries [21]. In these systems a query is captured by a mediating script (wrapper) which translates the query to the various data sources and returns the results to the user. Portals usually collect the data but do not integrate, rather, the data from the different sources are displayed separately within the portal interface.

The major advantage of mediation is that the application/portal delivers up-to-date data. Each source is mapped and the query mechanism is coded into a wrapper that is hidden from the user. The user accesses each source through a uniform query interface. The disadvantage to this approach is that only the queries supported by each individual system can be wrapped into the application/portal.

A more robust approach to data integration uses database federation (or mediation carried out by the database engine). Database federation describes a particular architecture where a relational database management system provides uniform access to a number of heterogeneous data sources. The data sources are federated, since they are linked together by the database management system. Database federation is an effective approach to the integration of heterogeneous data sources when the data can not be materialised into a data warehouse.

Data integration using a data warehouse approach, where data from the data sources are physically combined into one structure, is a very mature solution. The biggest drawback to developing a data warehouse is the scale of the resource required to integrate source data, and such data integration is usually performed piecemeal in data warehouses. Also, the integration performed by data warehouses is rarely reusable between projects. Each new project, therefore, has to perform its own data integration from scratch. Data warehouses are notoriously difficult to build, expensive to maintain and inflexible to changes in the questions that can be asked. This is largely because they require a copy to be made of data from all of the underlying data sources in a synchronized extraction, transformation and loading (ETL) process. Data not extracted into the warehouse cannot be queried conveniently, and changing the data that are selected involves considerable redesign work. This places a large upfront design burden on the warehouse schema and the ETL process. Biological data integration requires a more flexible technology that is amenable to the ever changing landscape of biological data.

3 Biological Data Integration

Initial solutions used to interoperate across bioinformatics databases used precomputed cross-references or Linkouts [22]. These database cross references are used in sequence databases to link to functional annotations within other databases. For example, EMBL nucleotide database cross links to protein sequence database Uniprot, protein function databases such as Prosite and Interpro, protein structure databases, enzyme and pathway databases and the literature database Pubmed. These links are based on identifiers and are calculated using sequence analysis tools. Sequence databases deliver data to users via flatfile downloads and are indexed in systems such as SRS [23] and Entrez[24]. Cross references in the databases enable users to move seamlessly from one database to another. However, the databases are linked together rather than integrated.

The increased complexity of biological data and the analyses performed on these data led to the development of more complex data integration solutions. Application integration for the interoperation of data and applications became the mechanism of choice when technologies such as CORBA became popular [25]. There are also examples of federated systems, such as BioKleisli [26] which used a query language to query and manipulate data that were maintained in different formats and DiscoveryLink from IBM [27] which provides users with a virtual database which can be accessed using SQL queries. Several data warehouse solutions have also been described [28–30]. None of these integration system can be easily extended or adapted to alternative data sets. This is mostly due to the underlying weaknesses of the technologies that were used to build the systems. Biological integration is not a solved problem. As new technologies become avaiable, the bioinformatics community exploits these with varied success. For example, semantic data integration is now in vogue [31–35] as it offers a solution to data integration that is more flexible and powerful. The advantages of semantic web technologies make it a very attractive alternative to traditional integration. This research project has aimed to understand how semantic data integration can be used effectively for biological data. A proof of concept exercise was performed to integrate data sets from laboratories studying the bacteria *Francisella tularensis* using multiple functional genomics technologies. Initially we focused on integration as a means to extend data annotations.

4 Proof of Concept - Combining RDF data

Rather than data integration in the traditional sense where overlapping data elements are resolved into one structure, genomic, transcriptomic and proteomic data need to be linked together using a scaffold that represents their relatedness. Semantic web technologies offer exactly this scaffold. Since genomics provides data on genes, transcriptomic experiments provide data on the transcription of genes (in particular tissues or under specific conditions), and proteomics provides identified peptides, this is not a simple case of resolving different data types and data formats. In this situation there are no common data elements between the data sets. As we deal here with data relationships which do not involve equality but different degrees of similarity or physical overlap, it is clear that traditional integration methods can not match these data in a simple manner. However, since these data are mutually related, integration can be achieved by using semantic web technologies. Data are combined in a standard data model using RDF and RDF-S. An ontology then maps the relationships between the entities within the RDF. The rich semantics within an ontology allows the definition of detailed relationships between concepts, whereas a database schema defines only the allowed structure of a set of relations. This makes it easier to merge ontologies, or to map them to one another. Ontologies form just one layer in the semantic web stack. Full benefits of data integration can be achieved when the semantic web technologies are layered together. The use of unique identification and XML exchange standards (see discussion) will greatly improve the level of integration that can be achieved. The base technology, where data are combined, is RDF, the Resource Description Framework [2].

The basic tenet of RDF is, everything is a resource that can be connected to other resources via properties [36]. The basic information unit is an RDF statement. A statement comprises of a triple: a subject, a property and an object. A set of triples can jointly form a directed labeled graph that can in theory model most, if not all, domain knowledge. As a graph, the RDF model is oblivious to both syntax and semantics, which makes it ideal for combining data. In theory, RDF can be used to model almost any data. RDF-S is the vocabulary definition language for RDF. The inter-relationships between the properties and objects in RDF are defined in RDF-S. RDF-S provides a logic, allowing inferences to be made on the RDF graph. Using the logic defined in RDF-S and derivation rules, new statements can be derived from existing statements. Figure 1 gives the RDF graph for one gene in the Fn genome.

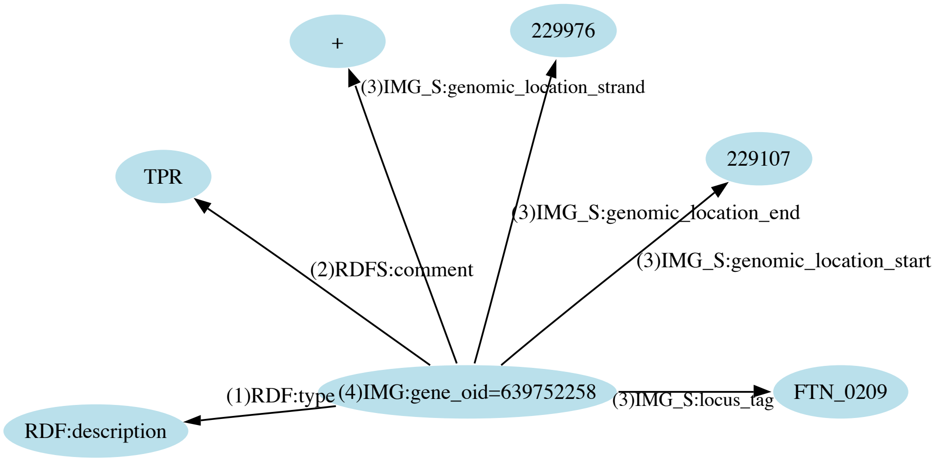
+
229976
(3)IMG_S:genomic_location_strand
TPR
229107
(3)IMG_S:genomic_location_end
(2)RDFS:comment
(3)IMG_S:genomic_location_start
(1)RDF:type
(4)IMG:gene_oid=639752258
(3)IMG_S:locus_tag
FTN_0209
RDF:description

Fig. 1. Example RDF graph for the gene FTN_0209 in the Fn genome data from http://img.jgi.doe.gov/. The following namespaces are used:
(1)RDF=http://www.w3.org/1999/02/22-rdf-syntax-ns#,
(2)RDFS=http://www.w3.org/2000/01/rdf-schema#,
(3)IMG_S=http://img.jgi.doe.gov/schema#,
(4)IMG=http://img.jgi.doe.gov/cgi-bin/pub/main.cgi?section=GeneDetail&page=geneDetail&

5 Implementation

Genome data and annotations (see Table 1) for Fn have been combined in RDF within the Sesame (version 2.1[1]) framework [15] installed on an imac with 2.33 Ghz Intel core 2 Duo processor and 2GB of memory. A single native (disk based B-Tree indexes) repository with default “SPOC,POSC” (**S**ubject **P**redicate **O**bject **C**ontext, **P**redicate, **O**bject **S**ubject, **C**ontext) index configuration was created. RDF triples were loaded using the sesame console. Perl scripts were written to parse each data source into RDF Ntriple format. The repository contains 1,258,677 triples. For the proof of concept we wished to determine how easily annotations of a particular experimental data set could be extended using the combined annotations in an RDF graph. Experimental data from a proteomic experiment studying the transcriptional regulator MglA [16] were also added to the repository in order to test our hypothesis. Additionally, an RDF Schema (RDF-S) was created for the experimental data set and this was also added to the repository. Our preliminary results and observations for Fn are described below.

[1] (www.openrdf.org/)

Table 1. Data sets and Uniform Resource Identifiers (URI's) used in the RDF graph

Resource Identifier	Source file name	No. of triples	load time in seconds
FTN	Ft_novicida_U112_go.n3	135,345	598
Fn genome annotation from Gene Ontology Database http://www.genome.jp/dbget-bin/www_bget?ftn:FTN_0277			
FTN	u112_kegg.n3	3252	415
Fn genome annotation from the KEGG Database http://www.genome.jp/dbget-bin/www_bget?ftn:FTN_0926			
NCBI Protein ID	NC_008601.n3	12,781	71.9
Fn annotation data from Refseq Database http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=118496620			
NCBI Protein ID	francisellaPROTEIN.fasta.n3	5,160	22.7
Fn sequence data from Refseq Database http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=118496625			
NCBI Protein ID	francisellaSUPERFAMILY.n3	16,110	96.4
Fn data from SUPERFAMILY Database http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=118496617			
IMG Gene ID	francisella.rdf2.n3	10,434	26.7
Fn genome data from The integrated microbial genomes (IMG) system http://img.jgi.doe.gov/cgi-bin/pub/main.cgi?section=GeneDetail&gene_oid=639753598			
PSN.V1	Membranes.n3/Soluble.n3/Wholecell.n3	748,157	3,430
University of Washington MglA protein abundance data sets from biological samples Membranes, Soluble and Whole cell https://wwamirce.gs.washington.edu/cgi-bin/fnu112/poson.cgi?poson=PSN081056			
PSN.V1	cogNumberURL.n3	2,548	98.9
University of Washington MglA annotation to COG database https://wwamirce.gs.washington.edu/cgi-bin/fnu112/poson.cgi?poson=PSN035866			
PSN.V3	FnU112Version3.n3	56,754	417.6
Fn genome data from University of Washington https://wwamirce.gs.washington.edu/cgi-bin/fnu112/poson.cgi?poson=PSN0088754.3			
DDB ID	interact-prot-peptides.n3	248,647	-
Fn peptide data from University of Washington http://regis-web.systemsbiology.net/protXML/protein_group/protein/peptide/id/ddb000010839p39			
DDB ID	interact-prot.n3	20,682	147.5
Fn protein identification data from University of Washington http://regis-web.systemsbiology.net/protXML/protein_group/protein/peptide/id/ddb000010839			
DDB ID	mgla_search_db.fasta.blastp4_ypURL.n3	1,719	9.7
DDB/PSN mapping from BLAST comparison http://regis-web.systemsbiology.net/protXML/protein_group/protein/protein_name/ddb000147854			

6 Results

Fn data sources are easily combined into an RDF Graph using Resource Identifiers
The combined RDF graph of Fn data sources can be used as a source for database cross references. The different resources and identifiers used in the RDF sources are shown in Table 1. A graph showing how these identifiers reconcile is shown in Figure 2. The Fn genome and annotation data sources were added into the repository first, and the FTN IDs, IMG (http://img.jgi.doe.gov/) Gene IDs and NCBI (http://ncbi.nlm.nih.gov) Protein IDs were connected through the CONSTRUCT statement shown in Table 2.

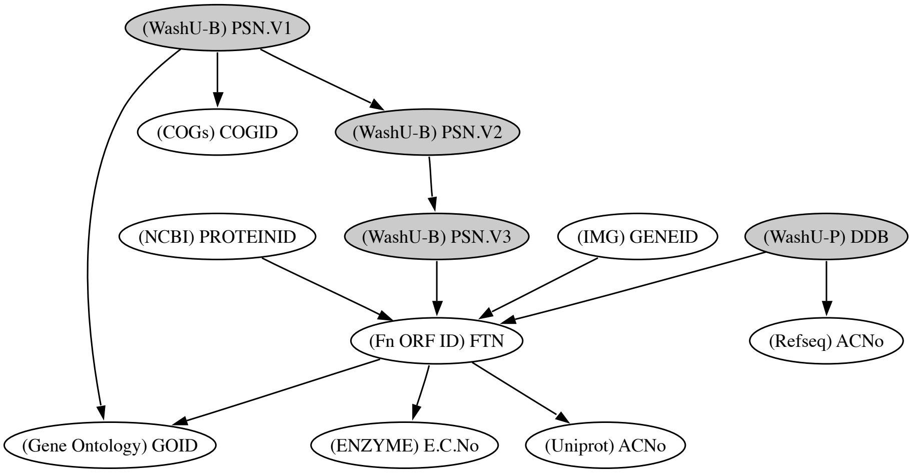

Fig. 2. Reconciled Identifiers in the RDF Graph, with the data source in parenthesis. This graph shows the connections that were made between the different identifiers when the data sources, given in Table 1, were combined in RDF. The shaded nodes are the identifiers used in the experimental data sets while the white nodes are the identifiers used in public domain annotations.

Further data sources were subsequently added and connected to the graph (Gene Ontology data, Fn KEGG data and annotations using COGs derived at the University of Washington). This was done to test our hypothesis that a connected graph of identifiers could increase the depth of annotation available to experimental data set, which included three data sets from the University of Washington (UW). These data sets used a variety of identifiers. The UW genome data (file, FnU112Version3) used internal identifiers called POSON numbers (PSN). There were several versions of these identifiers used internally and the experimental data generated using the MglA mutant strain (files, Membranes.n3/Soluble.n3/Wholecell.n3) used a different version of these identifiers.

Table 2. SeRQL (www.openrdf.org/doc/SeRQLmanual.html) CONSTRUCT statement connecting the identifiers FTN IDs, IMG Gene IDs and NCBI Protein IDs. The query uses two path expressions in the FROM clause. The connection between Protein IDs and Gene IDs is made through the FTN identifier.

```
CONSTRUCT {proteinid} nwrce:hasGeneID {geneid}
FROM {proteinid} G:locus_tag {ftn}, {geneid} G:locus_tag {ftn}
WHERE protein LIKE “http://www.ncbi.nlm.nih.gov*”
AND geneid LIKE “http://img.jgi.doe.gov*”
USING namespace G = <http://img.jgi.doe.gov/schema#>,
nwrce=<https://wwamirce.gs.washington.edu/fnu112/schema#>
```

Data that mapped across the POSON versions were added to the graph which enables the internal genome data and the Fn data graphs to connect. A third data set from a separate lab at the University of Washington used a third identifier, DDBs. These data were mapped to the existing identifiers through the addition of data from a BLAST [37] search against the genome data with sequence identity set to 100%.

Data integration increases the depth of bioinformatics annotation and reduces the effort required to manually annotate data in individual data sources

The depth of annotation available to the experimental data sets was increased through data integration based on database cross references. The RDF graph of the experimental data can be queried via the interposed layer of GO, KEGG and Superfamily descriptions, even though these data were not manually matched to these databases and provided with explicit annotations. These annotations are available by integrating data sets that have been manually annotated previously to at least one data source in the RDF graph. This form of data integration increases the amount of information available to biologists who now do not have to manually create each individual database cross reference. Sample SeRQL queries that show how the MglA experimental data are linked to annotations are shown below for KEGG and SUPERFAMILY data sets.

Querying MglA data through KEGG

The query in Table 3 gives the PSN identifiers and their E.C. numbers from the KEGG database for PSN's whose abundance in the MglA experiment was above 2000. The MglA data was not annotated using KEGG data. These links are available through the identifier cross references established in the RDF graph.

Querying MglA data through Superfamily

The query in Table 4 shows how the MglA data are linked to the SUPERFAMILY database in the RDF graph. PSN identifiers used in the MglA data are connected to FTN identifiers. Superfamily annotations are linked via PID identifiers which are connected to the FTN identifiers.

Table 3. SeRQL select query identifies PSNs and their E.C. numbers, where MglA peptide abundance is greater than 2000. KEGG database annotations are linked to the PSN identifiers in the MglA data through the FTN identifiers. The path expression used is displayed in bold and shown in Figure 3. Peptide abundance data is connected through the PSN identifiers.

```
SELECT psn, ec
FROM
{ftn} rdfs:seeAlso {ec},
{psn} rdfs:seeAlso {ftn},
{analysis} wu:poson {psn},
{analysis} mgla:experiment {exp},
{exp} mgla:abundance {abundance}
WHERE abundance > 2000
USING NAMESPACE
mgla = <https://wwamirce.gs.washington.edu/fnu112/experiments/mgla/schema#>,
wu = <https://wwamirce.gs.washington.edu/fnu112/schema#>
```

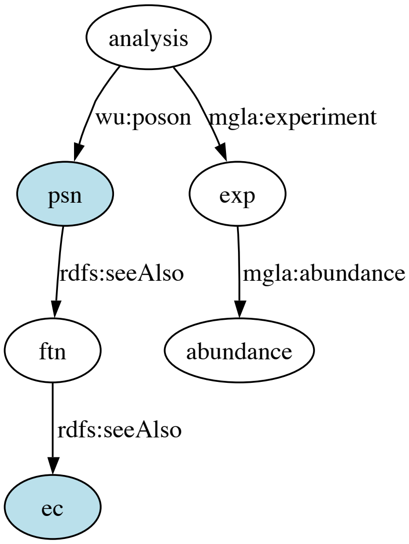

Fig. 3. Path expression used in the query given in Table 3.

7 Discussion and Further Work

Unique Identifiers

URI's, Uniform Resource Identifiers, are the base concept on which the semantic web technologies were developed. All things on the semantic web are resources, and all resources may be identified by URIs. The use of globally unique identification (GUID) can greatly facilitate data integration [38]. For example, when

Table 4. SeRQL select query identifies PSNs, NCBi Protein identifiers and Superfamily annotations from SUPERFAMILY database, where MglA peptide abundance is greater than 2000. SUPERFAMILY database annotations are linked to the PSN identifiers in the MglA data through the FTN identifiers which are linked to the NCBI Protein identifiers. The path expression is displayed in bold and shown graphically in Figure 4.

```
SELECT psn, pid, family
FROM
{psn} rdfs:seeAlso {ftn},
{pid} gen:locus_tag {ftn},
{pid} prot:Protein_Family {family},
{analysis} wu:poson {psn},
{analysis} mgla:experiment {exp},
{exp} mgla:abundance {abundance}
WHERE abundance > 2000
AND family LIKE ”http://supfam.org/SUPERFAMILY/cgi-bin/model.cgi?model=*”
USING NAMESPACE gen = <http://img.jgi.doe.gov/schema#>,
prot = <http://purl.uniprot.org/core/>,
mgla= <https://wwamirce.gs.washington.edu/fnu112/experiments/mgla/schema#>,
wu= <https://wwamirce.gs.washington.edu/fnu112/schema#>
```

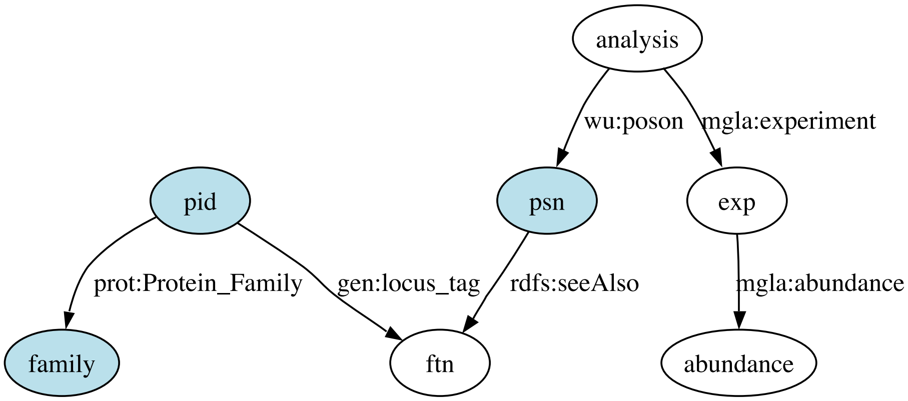

Fig. 4. Path expression used in the query given in Table 4.

two data nodes are the same in two resources, those data can be reconciled very easily if the nodes use GUIDs. In bioinformatics, the databases Genbank and EMBL share a unique identifier called an Accession number. A user can use this identifier to retrieve the same sequence in either database. This also means that this unique identifier can be used to reconcile these sequences if two separate resources make reference to the same sequence. If the unique identifier is used, we know that both resources are referring to the same sequence. Where individual data sources use their own forms of unique identification, a URI can make those identifiers unique, for example, www.protein.org/seq#123456 and www.gene.org/seq#123456. The use of URI's for unique identification can resolve the issue of the same identifier used in different databases referring to different things.

Lack of persistent unique identification in various Fn data sets was a considerable problem that required some manual mediation in order to resolve and combine data sets together. For Fn ORFs alone there were upwards of seven different identifiers used for the same entity (see Figure 2). By combining data in RDF, the different identifiers used in the Fn data can be reconciled and the RDF graph can be used as a source for cross references from the experimental data and the annotation in public domain data sources. However, in the long term for semantic web approaches to be successful in biology, *data producers and users need supported tools that can produce and resolve persistent unique identifiers.*

XML data exchange formats

The bioinformatics community have invested heavily in data exchange formats in XML. There are numerous examples. MIAME [39] is a standard format for microarray experiments. The Proteomics Standards Initiative (www.psi.org) have developed MIAPE for proteomics mass spectrometry data and other standard exchange formats for chromatography and gel electrophoresis. Data interchanged in standard formats like these can be readily transformed into RDF. These formats can also be used as the predicate vocabulary. Wherever possible it was our aim to use a standard term, when a suitable one existed. Currently, standard, easily accessible vocabularies are lacking. This has a lot to do with the fact that the omics XML standards were built as data exchange formats and *using them as vocabularies is out of scope.* However, our experience has highlights that further work is required in this area and some further coordination and extension of vocabularies and checklists is required.

We created simple XSLT scripts to convert data from these standard formats into RDF. Conversion scripts from common data formats such as FASTA and GenBank [2] have been created using Perl. These scripts are far easier to develop and more readily reusable than the traditional data warehouse ETL processes and this mechanism of data interchange is more accessible to biologists.

Standard vocabulary terms can also facilitate data integration. For example, just as two nodes that share the same URI are resolved, nodes in different graphs may be linked together by shared predicates. We required and ultimately created a vocabulary in RDF-S that described the experimental design of the MglA mutant experiment in order to easily integrate the peptide abundance data with the standard protein identification data that was in the ProtXML format. Although this paper focuses on integration at the level of resource identifiers, *further integration can be achieved via combing MglA data and the protein identifications at the level of properties used in both RDF graphs.*

[2] http://www.ncbi.nlm.nih.gov

Annotation of data analysis results

We found that experimental procedures and raw data are easily accessible in standard representations, however, analysed data, such as those found in secondary databases and published in papers are generally only available in ad hoc formats and on journal web pages. While progress has been made in standardisation of experimental data, *the analysis process and the analysed data still require an exchange standard*. This task might be handled partly by workflow descriptions and by standard vocabularies.

Ontologies

An ontology can capture the terms and the rules normally associated with human interpretation into a computationally amenable form. Two domains of data can therefore be described by an ontology and this allows the data within those domains to be queried together to enable data discovery. Currently, there are many tools for developing ontologies and as these mature and become more user friendly, more ontologies will be published and used for biological data integration. A few established ontologies are Gene Ontology [40] , Functional Genomics Ontology [41], Ontology for Biomedical Investigations [42], Influenza Infectious Disease Ontology [43], Mammalian Phenotype Ontology [44]. So far ontologies are being developed within local groups for specific purposes and there are still only a few community based efforts. However, the distributed nature of the development in ontologies is not expected to have any serious effect, since there is an understanding that ontologies can be merged and used together. Also, there are now many ontology repositories which are increasing the accessibility of the growing collection of online ontologies. For example, Ontoselect [45] collects online ontologies, SchemaWeb [46] is a resource to which users can submit ontologies, and the NCBO Bioportal [47] is an ontology repository providing uniform access to online ontologies within the Biology domain. BioPortal provides a valuable resource with very intuitive search and browse functionality and visualisations.

The two ontologies that are relevant to the experimental data sets that we have are PROTON [48] and the MGED Ontology [49]. PROTON models concepts, methods, algorithms, tools and databases relevant to the proteomics domain. The MGED ontology provides terms for concepts used within microarray experiments. These ontologies were loaded into the RDF repository, however, disappointingly, so far very little progress has been made combining proteomics and transcriptomics data with these two ontologies. These ontologies are heavily loaded with concepts specific to their domain (and experimental details) and do not relate the elements of integration, which is the abundance of mRNA and the abundance of peptides extracted from the organism. They do not describe the relationship between transcripts and peptides. Further work is required to make best use of the available ontologies to integrate data that share no common data elements but are related.

8 Conclusion

This paper demonstrates the progress made while testing semantic web technologies for data integration and highlights gaps and further requirements in data integration support. We demonstrated that data integration using RDF is easy to carry out and that simple integration at the level of resource identifiers can be achieved cheaply and efficiently. The combined data in the RDF graph provides a resource for database cross references for Fn data. An RDF dump of the Sesame repository (in N-triple format) can be downloaded from: http://spira.bio.gla.ac.uk/Francisella/swat4ls.nt.

This resource increases the depth of annotation available to biologists and this form of integration reduces the manual effort that would normally be required to gain this depth of annotation. Further work will include extending the integration semantically using RDF-S that maps between predicates used in different graphs. This will be tested in the first instance on peptide and transcript abundance data.

Acknowledgements

This work is funded by the BBSRC RASOR grant (BBC5115721). Data from the University of Washington was provided by Professor Dave Goodlett and Dr Mitch Brittnacher.

References

1. Hack C. J.: Integrated transcriptome and proteome data: the challenges ahead. Briefings in Functional Genomics and Proteomics **3:3**(2004) 212–219.
2. Lassila, O. and Swick R.R.: Resource Description Framework (RDF) Model and Syntax Specification. World Wide Web Consortium W3C (1999) http://citeseer.ist.psu.edu/212974.html
3. Joyce AR, Palsson B: The model organism as a system: integrating 'omics' data sets. Nature Reviews Molecular Cell Biology **7:3** (2006)
4. Ge, H. and Walhout, A.J.M. and Vidal, M.: Integrating omicinformation: a bridge between genomics and systems biology. Trends in Genetics **19:10** (2003) 551–560
5. Conway, T. et. al.: Microarray expression profiling: capturing a genome-wide portrait of the transcriptome. Molecular Microbiology **47:4** (2003) 879–889
6. Tyers, M. and Mann, M.: From genomics to proteomics. Nature **422:6928** (2003)193–197
7. Patterson, S.D.: Data analysis-the Achilles heel of proteomics. Nature Biotechnology **21:3** (2003) 221–222
8. Yates, J.R.: Mass spectrometry from genomics to proteomics. Trends in Genetics **16:1** (2000) 5–8
9. Schadt, E.E et. al.: Genetics of gene expression surveyed in maize, mouse and man. Nature **422: 6929** (2003) 297–302
10. Schadt, E.E. et. al.: An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics **37** (2005) 710–717
11. Karp, C.L. et. al.: Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. Nature Immunology **1** (2000) 221–226

12. Barker, J.R. and Klose, K.E: Molecular and Genetic Basis of Pathogenesis in Francisella Tularensis. Annals of the New York Academy of Sciences **1105** (2007) 138–159
13. Rohmer, L. et. al.: Comparison of Francisella tularensis genomes reveals evolutionary events associated with the emergence of human pathogenic strains. Genome Biology **8:6** (2007) R102
14. Nano, F.E. et. al.: A Francisella tularensis Pathogenicity Island Required for Intramacrophage Growth. Journal of Bacteriology **186:19** (2004) 6430–6436
15. Broekstra, J. and Kampman, A. and van Harmelen, F.: Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. Proceedings of the First International Semantic Web Conference (ISWC 2002) **2342** 54–68
16. Guina, T et. al.: MglA Regulates Francisella tularensis subsp. novicida Response to Starvation and Oxidative Stress. Journal of Bacteriology **189:18** (2007) 6580–6586
17. Brotcke, A. et. al.: Identification of MglA-Regulated Genes Reveals Novel Virulence Factors in Francisella tularensis. Infection and Immunity **74:12** (2006) 6642–6655
18. Lacroix Z. and Crichlow T.: Bioinformatics, Managing Scientific Data. Morgan Kaufman (2003)
19. Gorton, I. and Liu, A.: Architectures and technologies for enterprise application integration. Proceedings of the 26th International Conference on Software Engineering (ICSE 2004) 726–727
20. Lord, P. et. al.: Applying Semantic Web Services to Bioinformatics: Experiences Gained, Lessons Learnt. Lecture Notes in Computer Science (2004) 350–364
21. Curbera, F. and Duftler, M. and Khalaf, R. and Nagy, W. and Mukhi, N. and Weerawarana, S. Unraveling the Web Services Web: An Introduction to SOAP, WSDL, and UDDI. IEEE Internet computing **6:22** (2002) 86–93
22. Karp, P.D.: Database links are a foundation for interoperability. Trends in Biotechnology **14:8** (1996) 273–279
23. Etzold, T. and Ulyanov, A. and Argos, P.: SRS: information retrieval system for molecular biology data banks. Methods Enzymol **266** (1996) 114–28
24. Schuler, GD and Epstein, JA and Ohkawa, H. and Kans, JA: Entrez: molecular biology database and retrieval system. Methods Enzymol **266** (1996) 141–62
25. Stevens, R. and Miller, C.: Wrapping and interoperating bioinformatics resources using CORBA. Briefings in Bioinformatics **1** (2000) 9–21
26. Davidson, SB and Overton, C. and Tannen, V. and Wong, L.: BioKleisli: a digital library for biomedical researchers. International Journal on Digital Libraries **1** (1997) 36–53
27. Haas, L.M. and Schwarz, P.M. and Kodali, P. and Kotlar, E. and Rice, J.E. and Swope, W.C.: DiscoveryLink: A system for integrated access to life sciences data sources. IBM Systems Journal **40:2** (2001) 489–511
28. Sohrab, S. and Yong, H. and Tao, X. and Macaire, Y. and John, L. and Francis, O.B.F.: Atlas–a data warehouse for integrative bioinformatics. http://www.biomedcentral.com/1471-2105/6/34, BMC Bioinformatics **6:34**
29. Birkland, A. and Yona, G.: BIOZON: a system for unification, management and analysis of heterogeneous biological data. BMC Bioinformatics **7:1** (2006)
30. Kasprzyk, A. et. al.: EnsMart: A Generic System for Fast and Flexible Access to Biological Data. Genome Research **14:1** (2004) 160–169
31. Pasquier, C.: Biological data integration using Semantic Web technologies. Biochimie **90:4** (2008) 584–594
32. Smith, A.K. and Cheung, K.H. and Yip, K.Y. and Schultz, M. and Gerstein, M.B.: LinkHub: a Semantic Web system that facilitates cross-database queries and information retrieval in proteomics. BMC Bioinformatics **8: Suppl 3** (2007)

33. Villanueva-Rosales, N. and Dumontier, M.: yOWL: An ontology-driven knowledge base for yeast biologists. Journal of Biomedical Informatics **41:5** (2008) 779–789
34. Lam, H.Y.K. et. al: AlzPharm: integration of neurodegeneration data using RDF. BMC Bioinformatics **8:3** (2007) S4
35. Cheung, K.H. and Yip, K.Y. and Smith, A. and Deknikker, R. and Masiar, A. and Gerstein, M.: YeastHub: a semantic web use case for integrating data in the life sciences domain. Bioinformatics **21:1** (2005) i85–i96
36. Decker, S. and Mitra, P. and Melnik, S.: Framework for the Semantic Web: An RDF Tutorial. IEEE Internet Computing (2000) 68–73
37. Altschul, S.F. and Madden, T.L. and Schäffer, A.A. and Zhang, J. and Zhang, Z. and Miller, W. and Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25:17** (1997) 3390–3402
38. Clark, T. and Martin, S. and Liefeld, T.: Globally distributed object identification for biological knowledgebases. Briefings in Bioinformatics **5:1** (2004) 59
39. Brazma, A. et. al.: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nature Genetics **29** (2001) 365–372
40. Harris, M.A. et. al: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res **32:1** (2004) D258–61
41. Whetzel, P.L. et. al.: Development of FuGO: An Ontology for Functional Genomics Investigations. OMICS: A Journal of Integrative Biology **10:2** (2006) 199–204
42. Smith, B. et. al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology **25:11** (2007) 1251–1255
43. Lindsay Cowell and Barry Smith: Infectious Disease Ontology (IDO). www.infectiousdiseaseontology.org/
44. Smith, C.L. and Goldsmith, C.A.W. and Eppig, J.T.: The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biology **6:1** (2005)
45. Ontoselect http://olp.dfki.de/ontoselect/
46. SchemaWeb http://www.schemaweb.info/
47. NCBO BioPortal http://bioportal.bioontology.org/
48. PROTON (PROTo ONtology) Home Page http://proton.semanticweb.org
49. MGED - Microarray and Gene Expression Data Home http://mged.sourceforge.net