Semantic Data and Models Sharing in Systems Biology: The Just Enough Results Model and the SEEK Platform

Katherine Wolstencroft[1&3], Stuart Owen[1], Olga Krebs[2], Wolfgang Mueller[2], Quyen Nguyen[2], Jacky L. Snoep[1], Carole Goble[1]

1. School of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL
2. HITS gGmbH, Schloss-Wolfsbrunnenweg 35, Heidelberg, Germany, 69118
3. Leiden Institute of Advanced Computer Science, 2300 RA Leiden, The Netherlands

kwolstencroft@cs.man.ac.uk, sowen@cs.manchester.ac.uk, olga.krebs@h-its.org, Wolfgang.Mueller@h-its.org, Quyen.Nguyen@h-its.org, jls@SUN.AC.ZA, carole.goble@manchester.ac.uk

Abstract. Research in Systems Biology involves integrating data and knowledge about the dynamic processes in biological systems in order to understand and model them. Semantic web technologies should be ideal for exploring the complex networks of genes, proteins and metabolites that interact, but much of this data is not natively available to the semantic web. Data is typically collected and stored with free-text annotations in spreadsheets, many of which do not conform to existing metadata standards and are often not publically released.

Along with initiatives to promote more data sharing, one of the main challenges is therefore to semantically annotate and extract this data so that it is available to the research community. Data annotation and curation are expensive and undervalued tasks that have enormous benefits to the discipline as a whole, but fewer benefits to the individual data producers.

By embedding semantic annotation into spreadsheets, however, and automatically extracting this data into RDF at the time of repository submission, the process of producing standards-compliant data, that is available for semantic web querying, can be achieved without adding additional overheads to laboratory data management. This paper describes these strategies in the context of semantic data management in the SEEK. The SEEK is a web-based resource for sharing and exchanging Systems Biology data and models that is underpinned by the JERM ontology (Just Enough Results Model), which describes the relationships between data, models, protocols and experiments. The SEEK was originally developed for SysMO, a large European Systems Biology consortium studying micro-organisms, but it has since had widespread adoption across European Systems Biology.

Keywords: Semantic Systems Biology, Semantic Data Management, OWL Ontology, RDF Extraction from spreadsheets, Standard Metadata

1 Introduction

Systems Biology is a field of study that aims to understand biological and biomedical *systems* by analyzing and modeling their dynamic behavior. Mathematical models describing, for example, metabolic processes or genetic networks, can be used to predict the behavior of the system under different biological conditions or stresses. Linking together experimental data, models and model simulation results is therefore central to Systems Biology. Naturally, this involves a large amount of data integration.. Scientists need to combine different sources of heterogeneous information in order to model biological systems, and relate those models to available experimental data for validation.

The semantic web should be an ideal technology to assist with the process of identifying relevant data and their relationships; and there are a growing collection of semantic web resources for Systems Biology. For example, the Semantic Systems Biology portal [1] and the Chem2Bio2RDF [2] resources allow SPARQL (http://www.w3.org/TR/rdf-sparql-query/) queries across Life Science data, and the Systems Biology Ontology [3] allows scientists to semantically describe their models.

The bottleneck, however, is not with available vocabularies and frameworks for querying semantic Systems Biology data, it is with collecting and sharing Systems Biology data in a format that is amenable to semantic web querying. Currently, only a small fraction of the data and models produced during Systems Biology investigations are deposited for reuse by the community, and only a smaller fraction of that data is standards compliant, semantic content.

Funding agencies in Europe, such as the BBSRC in the UK and the BMBF in Germany, have developed new data policies to encourage and increase sharing. These policies stipulate that all data produced using public funding should be shared with the scientific community, and should be made available for a period of 10 years (http://www.bbsrc.ac.uk/datasharing). However, for this to happen on a large-scale, the community needs tools, repositories and standards to allow the systematic collection of adequately described data and to ensure the data is computationally amenable. For many types of experimental data, this situation has already improved. For example, there are repositories like GEO [4] and ArrayExpress [5] for microarray data, with corresponding standard formats and ontologies (MAGEML and MGED ontology respectively [6]).

The complication for Systems Biology is that there isn't one type of data, and there is added value from understanding and preserving the relationships between multiple different data sets. One Systems Biology model, for instance, could be constructed from the interactions between transcriptomics, proteomics and metabolomics data. If these data sets are publically shared, but submitted to their respective omics data silos, these relationships could be lost.

The SEEK platform [7] is a web-based resource for sharing heterogeneous Systems Biology data and models and preserving associations between datasets and

models. It is based on the ISA infrastructure (Investigations, Studies and Assays), a standard format for describing how individual experiments (assays) are aggregated into wider studies and investigations [8]. For the SEEK, ISA has been extended in order to encompass the description of mathematical models, and the relationships between them and the data.

The SEEK is a semantic integration resource. All metadata from experiments is extracted and stored in RDF (Resource Description Framework, http://www.w3.org/RDF/) and the relationships are defined and described by the underlying JERM ontology (the Just Enough Results Model Ontology). The majority of data is uploaded to the SEEK as Excel spreadsheets, so the RightField semantic spreadsheet application [9] (also developed during this work) is used to embed semantic annotation into the data.

This paper describes the semantic data integration in the SEEK, and how it supports the whole life cycle of data collection, annotation, sharing and reuse in Systems Biology. It draws on experiences in deploying this system in the SysMO Consortium, consisting of over 350 scientists, in over 100 laboratories.

The SysMO Consortium (Systems Biology of Micro-Organisms) is investigating systems approaches to studying wide variety of micro-organisms, including model organisms (like *E. Coli* and yeast) and microbes that are industrially important, such as those used in bio fuel or food production (e.g. *Clostridium acetobutylicum* or Lactic Acid bacteria, respectively). One of the main aims of the SysMO initiative is to: "record and describe the dynamic molecular processes going on in unicellular microorganisms in a comprehensive way and to present these processes in the form of computerized mathematical models." (http://www.sysmo.net)

Work with SysMO demonstrates that semantic data sharing and integration can be achieved by lowering the barriers to semantic annotation and extraction to RDF, whilst providing greater incentives to encourage the initial data sharing.

The paper is structured in the following way. Section 2 describes the JERM ontology and its function as both a central organizational framework and a vocabulary for describing Systems Biology data. Section 3 describes the capturing and extraction of RDF and section 4 evaluates the richness of queries possible compared to more conventional methods. Section 5 discusses related work, and section 6 describes experiences and conclusions.

2 The Just Enough Results Model Ontology

The JERM Ontology is an application ontology designed to describe the items in SEEK and the relationships between them (for example, data, models, experiment descriptions, results, samples, protocols, standard operating procedures and publications - subsequently referred to as SEEK assets); and to enable these relationships to be expressed with formal semantics. It is based on the idea of the Minimal Information Model [10]. Minimum Information Models have recently gained popularity in the Life Sciences because they offer a pragmatic solution to the provision of *sufficient* metadata. Metadata annotation is both time-consuming and costly, and the

biggest benefits are not for the data producers, but for scientists wishing to reuse data. A Minimum Information Model is the smallest amount of metadata required in order to make experimental data discoverable and interpretable by other scientists.

Most Minimum Information Models have been developed by communities of scientists working with a particular technology or experimental method. They are expressed as checklists, or XML specifications (schemas). There are already over 50, which have been collected under the umbrella of MIBBI (Minimum Information for Biological and Biomedical Investigations). The JERM takes the specification one step further, expressing the minimum information model as an OWL ontology.

The JERM provides a unifying framework for metadata elements common across MIBBI, and complies with existing MIBBI guidelines where they are available. The JERM describes SEEK assets and the relationships between assets and the experiments that created them (available from the BioPortal http://bioportal.bioontology.org/ontologies/1488). Crucially, all assets are related to the scientists that created them and the projects they originate from, so the ontology captures provenance information as well as physical links between assets. JERM definitions for each type of data in SEEK is different, but highly overlapping.

The JERM describes what type of experiment was performed, who performed it, and what was measured. These elements are common to all data types in Systems Biology, but each data type (e.g. microarray, mass spectrometry, enzymatic reactions), requires a different set of additional metadata. For enzyme experiments, the reactions being catalyzed need to be recorded, with substrates, products, and details of inhibition. For microarray experiments, the methods for quality control and normalization of the data should be recorded. For proteomics or metabolomics with Mass spectrometry, detailed descriptions of the instruments are required. For all cases, however, a description of the biological samples and any treatments applied is essential in order to understand the results of the experiment.

Using the JERM ontology to describe SEEK assets, and representing them in RDF, is essential to the sustainability of the resource. Data produced using continuously developed new experimental techniques must be incorporated into the SEEK. The flexibility and extensibility of RDF is ideal for such conditions. Using the JERM framework means that adding a new data type is possible without requiring the redesign of the underlying data model. Any new data type would have the same set of minimal metadata elements, plus some elements specific to that data. This would allow aggregation at any point of commonality (for example, the biological samples, the same factors studied, or membership of the same study), but the fact that there are differences between datasets is an advantage and not a complication.

At the time of writing, the JERM ontology contained 262 classes and 43 properties. It is represented in OWL in order to capture rich associations between SEEK assets and to allow complex queries and reasoning. The terms and properties of the JERM ontology therefore provide a schema for managing and extracting SEEK metadata. An OWL ontology, however, is not a suitable representation for laboratory biologists. The JERM schemas must be presented in a way that enables their use, without requiring the Systems Biologists to invest time and resources in learning new tools and ontology/RDF skills.

In the SysMO consortium, the majority of laboratory scientists (microbiologists, biochemists, molecular biologists, and geneticists) use spreadsheets for the daily management and manipulation of data. By embedding the JERM metadata model in a spreadsheet format, and enabling the use of JERM (and other) vocabulary terms for annotation, the process of standardized semantic data collection can become part of the existing data management activities in the laboratory. JERM-compliance in SEEK is therefore achieved by the sharing of JERM-compliant spreadsheet templates.

2.1 JERM Templates

JERM spreadsheet templates have been developed for a wide range of experimental data types. In collaboration with members of the SysMO consortium, templates have been designed for numerous different types of microarray and RNA-Seq data, proteomics, interactomics, metabolomics, and enzyme kinetics. Figure 2 shows how JERMs for different experiment types, with different metadata content and structures can overlap and interrelate in the SEEK.

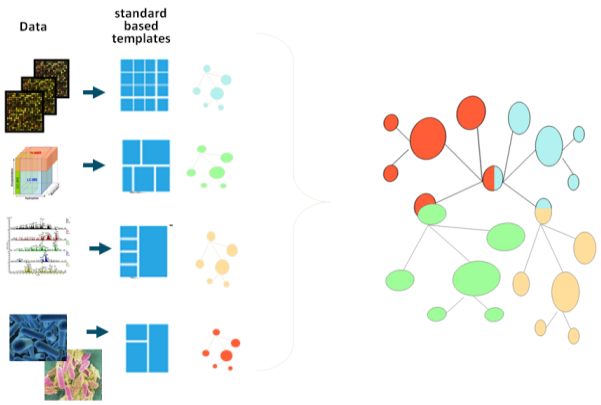

Fig. 1. A depiction of different JERM representations for different data types, showing how they can be aggregated at points of commonality and how they can retain their different structures where they differ.

The JERM templates each follow the same basic format. The first worksheet contains metadata elements describing the data set and its provenance. Worksheet 2 describes the properties and conditions of samples, which includes the organism, strains and any genetic modifications (compliant with the BioSamples metadata specification from the EBI). The final mandatory worksheet is a matrix that contains the actual data values obtained from measurements on each sample and their specific conditions (e.g. specific time-points, concentrations of metabolites or carbon sources, etc). Sample labels should correspond to the labels in the data matrix, so that the BioSamples sheet

acts as a key to the details of the data matrix. Optional worksheets describing derived results or details of the instrument specifications can also be added, but these are not currently converted to RDF by default. Where metadata standards require the use of ontology terms for annotation, the JERM templates contain drop-down lists of only those terms that are permissible for a certain metadata element. These terms, and other semantic content, are embedded in the spreadsheet using the RightField application (see section 3 for a full description of RightField).

The resulting collection of data sets have a uniform structure and uniform semantics, and can be interpreted by both scientists and computational systems. This is achieved without exposing scientists to any semantic web infrastructure and without requiring scientists to adopt new technologies and data management techniques. Figure 3 shows an excerpt from a typical metadata sheet in a JERM template.

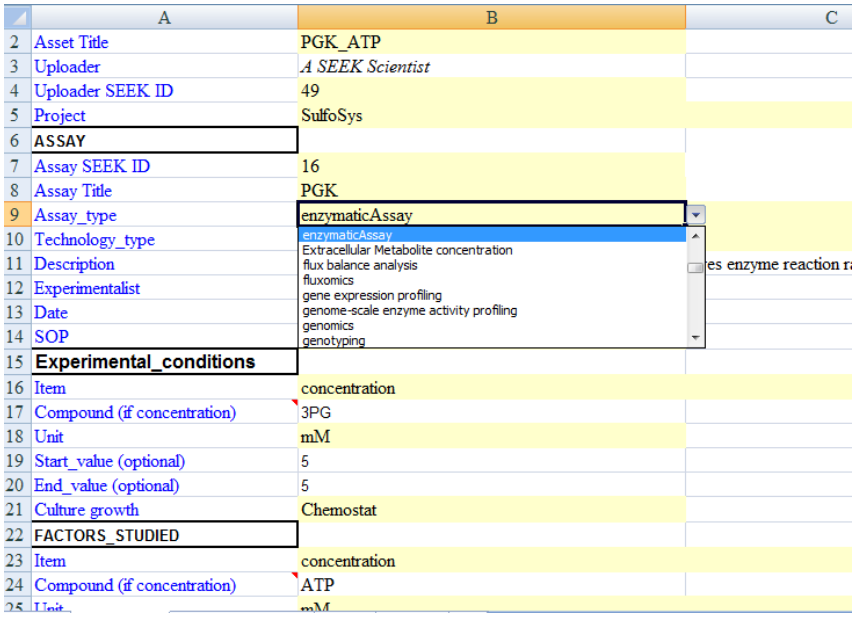

Fig. 2. An extract from a JERM template describing an enzymatic activity assay. The drop-down list shows the terms permitted for the JERM "Assay Type" classification. Each yellow shaded box represents a cell with embedded semantic content, which can be automatically extracted into RDF using the RightField application

3 Capturing and Extracting Semantic Data

JERM templates contain semantic content. An RDF triple can be defined for each spreadsheet cell that contains JERM-compliant metadata. For example, in the cell containing the title, the underlying triple states that "Asset *has_title*, title", where the scientists supplies the title as free text. In the cell describing the environmental conditions, "Asset *has_part* environmentalCondition", where the permissible environmen-

tal conditions are presented as a simple, drop-down list in the cell. The cell therefore contains the vocabulary of terms to be used for annotation (taken from the JERM ontology in this case), and the property that describes the relationship between the dataset and the annotation.

These semantic augmentations are provided using RightField. RightField is an open-source cross-platform java application that provides a mechanism for embedding semantic content into Excel or Open Office Spreadsheets. It was developed as part of SEEK, but it is a stand-alone application that has been used in applications in the Life Sciences and others, including standardizing medical record collection, and the cataloguing of Egyptian mummy samples.

RightField allows the marking-up of individual cells, ranges of cells, or whole rows and columns, with particular selections of ontology terms. For example, all sub-classes (or all direct sub-classes), or all instances (or only direct instances), can be included from a particular class in an ontology. Multiple ontologies can be used to embed content into any one spreadsheet, but each cell can only display a selection of terms from a single ontology.

The RightField client interface is not designed to be used by all SEEK end-user scientists. Preparing a RightField-enabled spreadsheet is an administrative task for the bioinformatics specialists on the project. Once a template has been prepared, the spreadsheet can be shared and used by the scientists, without exposing them to the underlying semantic content. In SysMO SEEK, a collection of templates have been prepared and shared for a number of different experimental data types (https://seek.sysmo-db.org/help/templates).

RightField also enables the automated extraction of data into RDF. Due to the fact that the subject, predicate and object have already been defined by filling in the template with experimental data, RDF statements can be automatically generated for each cell. Therefore an RDF graph, or collection of graphs, can be generated for each data set. In SysMO, the resulting RDF graphs comply with the JERM ontology model, so more complex queries and reasoning can be performed using the OWL representation.

RightField transforms the experimental metadata into RDF, but does not currently transform data value literals. However, future versions of RightField will allow ranges of cells to be referenced from other cells as part of an RDF triple. This will enable the actual data values to be defined and treated as data sets, associating data more easily with error and standard deviation values, for example.

As RightField-enabled data is uploaded to SEEK, an RDF representation of its metadata is extracted and stored in a Virtuoso triplestore (http://virtuoso.openlinksw.com/). A SPARQL end-point to public data in SEEK is available at: http://iswc.sysmo-db.org/sparql.

RightField-enabled spreadsheets allow the collection of semantic information by stealth. The semantics are embedded in the spreadsheet, a tool that is already in common use for data management. This provides a low barrier for uptake and ensures that those using the JERM templates do not require prior knowledge of the Semantic Web and ontologies. Term labels are displayed in the spreadsheets, but the IRI of each label is also stored in hidden sheets, along with the ontology URI and ontology version information. By embedding the terms and their provenance, the spreadsheet re-

mains self-contained and can therefore be used and shared in the same way as a regular spreadsheet. If ontology versions change after the creation of a JERM spreadsheet, they will not be updated until, or unless, the template is re-opened in the RightField client. This is essential for the consistent collection and annotation of data. If an experiment takes months to complete, *all* data from that experiment should be annotated with the same versions of the same ontologies. Updating templates to use newer versions of ontologies should be a conscious decision and not an automated process.

One potential limitation of the system is that if SEEK scientists choose not to use JERM templates, the amount of metadata that can be collected, and therefore the RDF graph of that data, is reduced. However, anything uploaded to SEEK is linked to the person and the project it belongs to, the organism under investigation, and the assay type and technology type of the asset is also recorded. This provides enough information for non-standard datasets to be discovered and reused through SEEK, although the contents cannot be fully explored.

4 Evaluation: 20 Questions

The user interface to the SEEK platform (and its search capabilities) were designed by following Jim Gray's 20 questions model [11] A focus group of scientists from the SysMO consortium were asked to list all the questions they envisioned asking of SEEK content. These questions were distilled into the top 20 queries that SEEK must be able to answer in order to serve the whole SysMO community. The performance of the SEEK was evaluated against these questions, initially using a traditional *classic* search through the SEEK web interface (driven by the Lucene search engine), and then using the RDF/SPARQL end-point. It was determined that whilst the classic search was sufficient for most questions, querying the RDF enabled a greater number of queries to be answered.

The SEEK focus group (also known as the SysMO PALs network) were a collection of post-docs and PhD students from each of the SysMO projects, covering a broad range of research areas in experimental biology, mathematical modeling and bioinformatics. The questions were obtained before the development of any prototype interface, to prevent focus group members being constrained by what they considered technically possible. Table 1 shows the set of 20 questions and the ability of the SEEK to return those queries using the classic search and using semantic search over the SEEK RDF with SPARQL.

Question	Classic	RDF
1. Which experiments were carried out on *E.coli* (organism X)?	+	+
2. Which strains of *E.coli* (organism X) are being used in SysMO?	+	+
3. What proteomic (experiment x) data is available? What types of transcriptomics (assay type x) experiments were performed?	+/+	+/+

4. Who has experimental data on gene/protein/metabolite X	+/+/+	+/+/+
5. Which microarray data files show up-regulation in genes with Gene ontology molecular function X	-	+
6. What data is available from SysMO-LAB (project X)?	+	+
7.What data was used to construct the model and what data was used to validate it?	+	+
8. Who is in the COSMIC (project X) project?	+	+
9. What Standard Operating Procedures were used in experiment X? Are there any protocols for Mass Spectroscopy (technology type X) experiments?	+/-	+/+
10. Who is working on growth rate (assay type X) experiments?	+	+
11. What publications are available for models in *Pseudomonas* (organism X)?	+	+
12. Are there any models on yeast (models on X)?	+	+
13. Who is in more than one SysMO project?	-	+
14. What are the factors studied in the MOSES project (project X)?	-	+
15. What are the data on steady state fluxes in organism XXX in condition XXX?	-	+
16. What type of experimental data should I collect to apply the Teusink model (model X)?	+	+
17. What model simulation results are available?	+	+
18. Is the original data from my archive sufficient for model X	+	+
19. How good is the correlation between transcriptome levels, proteome levels and enzyme activities in organism X in study Y? Is a time delay observed?	-	-
20. What range of concentrations of metabolites (extra- and intracellular) are detectable from organism XXX	-	+

Table 1. The "20 Questions" identified by the SEEK focus group as being the most important queries to perform over the SEEK

The results of the comparison show that the classic search and RDF/SPARQL query perform equally well on the majority of questions. In six cases, however, the question can only be answered using RDF/SPARQL, and for question 19, neither search method is successful. Question 19 cannot be answered directly because it requires both the extraction of relevant data and the analysis of that data to determine correlations. The other differences in results were caused by either limitations on aggregating information, or a reliance on data held externally to SEEK. SEEK without RDF capabilities is (of course) limited in providing all conceivable aggregates and projections of information. Here, semantic web techniques *complement* SEEK's capabilities.

Questions depending on outside data, or information from outside ontologies are not feasible with the classic SEEK. Here the use of Semantic Web techniques gives the opportunity to pull in several data sources and then combine them for answering the query. The semantic web therefore helps to *extend* SEEK's capabilities.

Section 4.2 shows concrete examples of SPARQL queries used to answer a selection of the original SEEK 20 questions.

As shown in table 1, the majority of queries could be answered using either the classic or RDF/SPARQL approach. However, data and models in individual SEEK instances exist in a wider ecosystem. There are over 1500 other database resources in the Life Sciences, and many of these are available as Linked Data. There are also other SEEK instances, containing data and models from other consortia. This is the primary advantage of the RDF/SPARQL approach. The Linked Open Data initiative already provides conventions for querying multiple SPARQL end-points.

The ability to query the contents in one data source is useful, but the ability to query across multiple, related data sources is therefore the ultimate goal. By extracting and serving SEEK data as RDF, SEEK data can also be served as Linked Data.

4.1 20 Questions Revised

A review of the 20 questions, to coincide with the release of the first RDF-enabled SEEK version, presented an opportunity to revise the questions and adapt to the changing requirements of the SysMO consortium. Over half of the new questions were variations on those already posed, but the rest were much richer and explored the detailed content of the data and models. For example, questions such as "Which data files contain compound X and/or receptor Y, in organism Z?", or "what is the concentration of fructose bisphosphate in *Lactococcus lactis?*" make use of the RDF aggregation capabilities. Other questions further exploit external data links, for example, "What experimental data exists in SEEK for the Gene Ontology biological process X". Some SEEK data may not be annotated with Gene Ontology terms, so the query must identify all gene products annotated with that term, and then match these to the gene products in SEEK. This would require interrogating external primary sequence databases, such as UniProt [12], as well as the Gene Ontology [13]. More extensive searching over a greater number of external resources is required to answer questions such as, "what additional information is known about compound X?".

The revised 20 questions are being used to inform the design of the next version of the SEEK, which will be served as Linked Open Data.

4.2 Example Queries

The following SPARQL queries can be performed to answer a selection of the questions from table 1. All data returned in SEEK is restricted to data that has been publically shared. Every asset uploaded to SEEK can be shared with named individuals, groups, the whole consortium, or released for public view. However, logging-in is not required to use the SPARQL endpoint, so SPARQL users are therefore considered anonymous users.

1. Which experiments were carried out on *E.coli*?

```
SELECT ?title ?assay ?organism  WHERE
   {
      GRAPH <iswc13:rdf>
      {
        ?organism jerm:NCBI_ID
<http://purl.obolibrary.org/obo/NCBITaxon_562> .
        ?organism a jerm:organism .
        ?assay jerm:investigates ?organism .
        ?assay a jerm:Assay .
        ?assay dcterms:title ?title .
      }
}
```

E.coli is identified by its NCBI identifier. All assays associated with this organism are returned.

2. What metabolomics data is available?

```
SELECT  ?data ?title   WHERE
{
{
?types rdfs:subClassOf jerm:metabolomics .
}
GRAPH<iswc13:rdf>
{
{
?assay jerm:hasType jerm:metabolomics.
}
UNION {
?assay jerm:hasType ?types.
}
?data jerm:isPartOf ?assay;
 a jerm:Data.
?data dcterms:title ?title
}
}
       GROUP BY ?data
```

3. Are there any models on yeast, and what data is associated with those models?

```
SELECT ?model ?model_title ?assay ?assay_title ?data ?data_title WHERE
{
 GRAPH<iswc13:rdf>
```

```
{
 ?organism jerm:NCBI_ID <http://purl.obolibrary.org/obo/NCBITaxon_4932>;
 a jerm:organism .
 ?model jerm:investigates ?organism;
   a jerm:Model .
 ?assay jerm:hasPart ?model;
  jerm:hasPart ?data.

 ?data dcterms:title ?data_title .
  ?assay dcterms:title ?assay_title .
?model dcterms:title ?model_title .

}
}
```

4. Find all strains that have had samples derived from them during the second quarter of 2012

```
SELECT ?specimen ?strain ?strain_title ?ncbi ?sample ?sampling_date
WHERE {
  ?specimen a jerm:specimen.
  ?specimen jerm:isDerivedFrom ?strain.
  ?strain dcterms:title ?strain_title.
  ?strain jerm:NCBI_ID ?ncbi.
  ?sample ?isDerivedFrom ?specimen.
  ?sample jerm:sampling_date ?sampling_date
  FILTER (
   ?sampling_date > "2012-04-01"^^xsd:date &&
   ?sampling_date < "2012-08-31"^^xsd:date
 )
}
```

Query 4 represents one of the more complex queries collected in the second requirements session.

4.3 Implications for SEEK Querying

The SEEK triple store provides a powerful mechanism for querying the SEEK contents and items related to that contents in other resources. It provides a flexible, sustainable platform that can be extended and expanded to incorporate new experimental data or model types, without requiring the redesign of the underlying data model (the JERM). In order to query the SEEK triple store in its current form, users are required to construct SPARQL queries. This presents a barrier to most SEEK users. For data collection, SEEK users are shielded from the details of the semantic technologies by embedding them in familiar tools, such as spreadsheets. The same approach will be adopted for querying the contents. Initially, canned queries will be offered through a

web interface, allowing users to formulate queries to answer the 20 questions. This approach is restrictive and loses the flexibility offered by SPARQL, but it enables the queries that have been identified as essential in a universally accessible way. The SPARQL end-point will still be available to those that require it, and in the long-term, users with SPARQL expertise will be able to construct new queries and serve then through the same web front-end.

5 Related Work

The SEEK is a platform that addresses data collection, annotation, submission and reuse in Systems Biology. It is unique in its approach to embed semantics into existing and familiar tools and it is also unique in the way it collects and stores information on both models and data. The integration and interaction between data and models is the definitive characteristic of the Systems Biology community.

There are a number of related resources that address some of the same problems as those in the SEEK, but most do not support the whole workflow. For example, initiatives from the BioSharing community tackle data standardization and annotation, other initiatives from Systems Biology provide repositories for data and models, initiatives from bioinformatics provide RDF representations and Linked Data versions of commonly used resources, and initiatives from computer science provide automated extraction of RDF from spreadsheets. Summaries of these are described below.

Bio Sharing

The BioSharing portal (http://biosharing.org) is a catalogue of standards, formats and ontologies that are in use in the Life Sciences. For SEEK, it is a valuable resource for identifying community standards that should be used in SEEK.

The ISA tools suite [14] enables the creation and management of ISA-TAB files. ISA creator has similar functionality to RightField. It enables the creation of ISA-TAB compliant metadata templates to allow groups of scientists to collect standards-compliant semantic metadata. ISA creator also has a spreadsheet-like interface, but it is operated from bespoke client software and designed for expert users rather than laboratory scientists.

ISA tools focuses only on experimental data. In SEEK, the ISA structure is used to organize and link related experiments, but it has been extended to incorporate the relationships between the omics data and models. ISA tools and SEEK have similar and complimentary approaches to multi omics data exchange. ISA tools are also developing an RDF representation, which should enable queries between SEEK and other ISA resources in the future.

Systems Biology Semantic Data Resources

The semantic Systems Biology portal [1] provides access to a data warehouse of Systems Biology data in RDF. Unlike the resources described above, this portal compiles available public data and serves it from the same end-point. For SEEK and related resources, it is another source of external data, although the frequency of updates to underlying resources may not reflect the frequency of updates in those underlying resources. The Linked Life Data resource (http://linkedlifedata.com/) provides RDF

and SPARQL interfaces to a broader range of biological data collections, and both Bio2RDF [15] and Chem2bio2RDF [2] provide RDF formatted collections of biological and chemical data respectively.

Systems Biology Data Management

The Data Integration Platform for Systems Biology Collaborations (DIPSBC) [16] was designed specifically for managing Systems Biology data. Like SEEK, it is compliant with existing community metadata standards, but it accepts and parses XML representations of the data, rather than spreadsheets. Data uploaded to DIPSBC is indexed and searched using Lucene, and a Foswiki interface allows users to create, share and manage versions of pages and resources as required. It currently does serve data as RDF and there is no specific ontology support.

The Bioinformatics Resource Manager (BRM) is a java based client/server database system with a PostgreSQL back end. It is a data warehouse system that has been designed for managing Systems Biology data. It imports data from public sources, such as KEGG, NCBI and the Gene ontology, and allows users to combine this public data with local data files. Data is incorporated or exported using wizards in the client. Like SEEK, local data can be uploaded centrally or stored locally, but unlike SEEK, it has no support for Systems Biology model management, and it is not a semantically aware resource.

RDF Extraction from Spreadsheets

There are several tools that perform extractions of spreadsheet data to RDF. For example, Excel2RDF (http://www.mindswap.org/~rreck/excel2rdf.shtml), and RDF123. A key difference between these resources and the RightField RDF generation is that they focus on the transformation of spreadsheet content, rather than the structure and consistency of that content. Therefore, RDF relationships between spreadsheets cells are produced, rather than relationships between the concepts in the content. RightField templates allow the extraction of data to a particular metadata model, allowing the expression of complex relationships between cell content across datasheets. In addition, RightField does not require a separate mapping file because this information is self-contained. Therefore, cells can be moved around or copied without affected the expected RDF produced.

6 Discussion

Standards-compliant data collection and annotation are becoming increasingly important in the Life Sciences. The effective management and reuse of data is essential in large, collaborative projects and is increasingly becoming a condition of public funding. Semantic web technologies can have an important role in this process and semantic annotation makes data more valuable for reuse.

In order to describe Systems Biology studies, multiple experiments, producing diverse data-types must be described, interlinked and associated with corresponding mathematical models. Systems Biology is therefore a discipline with complex data management requirements. These must be balanced against the time-consuming pro-

cess of data curation and annotation, to enable enough information to be collected for discovery and reuse.

There are a large number of ontologies and standards available in the Life Sciences. Many of these are directly relevant to Systems Biology and the SEEK draws upon them as annotation vocabularies and metadata schemas in the JERM templates. Above these resources, the JERM Ontology provides a formal representation of the relationships between SEEK assets and a framework for aggregating and integrating metadata. This is a common infrastructure for semantic resources, but SEEK differs in the way that the semantics are embedded and hidden behind other applications. Users do not require prior knowledge of the semantic web or ontologies in order to annotate data to a standards-compliant format, or to generate RDF graphs of that data. This is one of the largest advantages of the SEEK approach. SEEK users do not have to change their general data management working practices, so the barrier to adoption is low and attainable by all.

The SEEK is a system that is deployed and used by multiple consortia of Systems Biologists across Europe. At the time of writing, the SysMO SEEK contained over 2000 assets that had been uploaded by SysMO consortium members. This demonstrates a high level of uptake and success in the approach.

The SEEK offers an off-the-shelf solution to data management and promotes the use of existing metadata standards and ontologies wherever they are available. It does not enforce standards-compliance, but it streamlines the process and provides incentives for compliance. The SEEK is a semantic web resources with an interface that meets the requirements and capabilities of its end-user scientists.

The next step for the SEEK will be to serve the generated RDF as Linked Data, which will enable easier federated searching between different SEEK instances and other Life Science Linked Data resources. New interfaces to the SPARQL end-point will also be developed, to make the writing and use of SPARQL queries more accessible.

7 Acknowledgements

This work was funded as part of the SysMO-DB2 grant awarded by the BBSRC (BB/I004637/1) and the BNBF grant FKZ:0315781

8 References

1. Antezana, E., Blonde, W., Egana, M., Rutherford, A., Stevens, R., De Baets, B., Mironov, V. and Kuiper, M. (2009) BioGateway: a semantic Systems Biology tool for the life sciences. *BMC Bioinformatics*, **10 Suppl 10**, S11.
2. Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y. and Wild, D.J. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics*, **11**, 255.

3. Courtot, M., Juty, N., Knupfer, C., Waltemath, D., Zhukova, A., Drager, A., Dumontier, M., Finney, A., Golebiewski, M., Hastings, J. *et al.* Controlled vocabularies and semantics in Systems Biology. *Mol Syst Biol*, **7**, 543.
4. Barrett, T. and Edgar, R. (2006) Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*, **411**, 352-369.
5. Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A. *et al.* (2009) ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, **37**, D868-872.
6. Ball, C.A. and Brazma, A. (2006) MGED standards: work in progress. *OMICS*, **10**, 138-144.
7. Wolstencroft, K., Owen, S., du Preez, F., Krebs, O., Mueller, W., Goble, C. and Snoep, J.L. The SEEK: a platform for sharing data and models in Systems Biology. *Methods Enzymol*, **500**, 629-655.
8. Sansone, S.A., Rocca-Serra, P., Brandizi, M., Brazma, A., Field, D., Fostel, J., Garrow, A.G., Gilbert, J., Goodsaid, F., Hardy, N. *et al.* (2008) The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?" *OMICS*, **12**, 143-149.
9. Wolstencroft, K., Owen, S., Horridge, M., Krebs, O., Mueller, W., Snoep, J.L., du Preez, F. and Goble, C. RightField: embedding ontology annotation in spreadsheets. *Bioinformatics*, **27**, 2021-2022.
10. Taylor, C.F., Field, D., Sansone, S.A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C.A., Binz, P.A., Bogue, M., Booth, T. *et al.* (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol*, **26**, 889-896.
11. Gray, J. and Szalay, A. (2004). Microsoft Research, Microsoft Corporation.
12. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*, **40**, D71-75.
13. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, **32**, D258-261.
14. Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O. *et al.* ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, **26**, 2354-2356.
15. Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P. and Morissette, J. (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform*, **41**, 706-716.
16. Dreher, F., Kreitler, T., Hardt, C., Kamburov, A., Yildirimman, R., Schellander, K., Lehrach, H., Lange, B.M. and Herwig, R. DIPSBC--data integration platform for Systems Biology collaborations. *BMC Bioinformatics*, **13**, 85