

# RightField: Semantic enrichment of systems biology data using spreadsheets

**DOI:**
10.1109/eScience.2012.6404412

**Document Version**
Accepted author manuscript



**Published in:**
2012 IEEE 8th International Conference on E-Science, e-Science 2012|IEEE Int. Conf. E-Sci., e-Science

# RightField: Semantic Enrichment of Systems Biology Data using Spreadsheets

Katherine Wolstencroft, Stuart Owen, Carole Goble
School of Computer Science, University of Manchester
Kilburn Building, Oxford Road
Manchester, UK
*firstname.lastname*@manchester.ac.uk

Quyen Nguyen, Olga Krebs, Wolfgang Müller
Heidelberg Institute of Theoretical Studies gGmbH, Schloß-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
*firstname.lastname*@h-its.org

*Abstract*— **The interpretation and integration of experimental data depends on consistent metadata and uniform annotation. However, there are many barriers to the acquisition of this rich semantic metadata, not least the overhead and complexity of its collection by scientists. We present RightField, a lightweight spreadsheet-based annotation tool for lowering the barrier of manual metadata acquisition; and a data integration application for extracting and querying RDF data from these enriched spreadsheets. By hiding the complexities of semantic annotation, we can improve the collection of rich metadata, at source, by scientists. We illustrate the approach with results from the SysMO program, showing that RightField supports the whole workflow of semantic data collection, submission and RDF querying in Systems Biology. The RightField tool is freely available from *http://www.rightfield.org.uk*, and the code is open source under the BSD License.**

***Keywords RightField, RDF, Semantic annotation, spreadsheet science, biological data management (key words)***

## I. INTRODUCTION

The complexity and heterogeneity of Life Science data has been increasing rapidly with the development of new high throughput experimental techniques for studying 'omics (e.g. genomics, proteomics, metabolomics) [1]. In Systems Biology, integrating this diverse set of data is necessary in order to study the processes across whole biological systems, and this is only possible with the use of standard metadata and ontologies for data annotation [2].

There are a growing number of Semantic Web resources for Systems Biology, all designed to assist in the integration and comparisons of data and knowledge. The Semantic Systems Biology framework [3] and the Linked Life Data resource (http://linkedlifedata.com/), for example, provide RDF and SPARQL interfaces to large biological data collections, and both Bio2RDF [4] and chem2bio2rdf [5] provide RDF (Resource Description Framework, http://www.w3.org/RDF/*)* formatted collections of biological and chemical data respectively. Unlike conventional databases, combining knowledge with RDF allows both integration *and* aggregation. Different types of experimental data have different, but overlapping sets of metadata. The flexibility of RDF enables the schema to be aggregated at points of commonality, without attempting to homogenize the content of each. New types of datasets can, therefore, be assimilated and compared as required. However, the bottleneck is not in displaying available data as RDF , but in encouraging scientists to publish their data to conform to community standards and guidelines so that it can be easily served as RDF and Linked Data.

There are a myriad of metadata standards and ontologies available for Systems Biology data [2, 6-11], but, standardizing and annotating data is often regarded as an optional extra step in laboratory data management. Tools to assist this process have only recently begun to emerge [12]. Consequently, only a small fraction of experimental data is actually published and shared with the community, and a smaller fraction is shared in a standards-compliant, semantically annotated format [13]. To encourage greater semantic data sharing: (a) the annotation and submission of data must be more accessible to the scientists producing the data; and (b) the direct benefits of querying their data in the context of related work on the Semantic Web must be made more apparent.

The pan-European Systems Biology initiative, SysMO (Systems Biology for MicroOrganisms, http://www.sysmo.net), is an example of a large, dispersed and data-rich community that could significantly benefit from a Semantic Web approach to the integration, aggregation and comparison of their data. The initiative supports 13 multi-site consortia, ranging from 4 to 17 partners, spread across 7 European countries. Over 100 institutions and over 340 scientists participate in experiments on the dynamic processes of micro-organisms. Experimental data from all consortia are stored, exchanged and linked with mathematical prediction models and standard operating procedures using a custom collaboration and sharing platform, the SEEK (https://seek.sysmo-db.org). Currently, over 900 datasets have been deposited by consortia members.

One of the most popular and familiar tools for scientific data capture is the spreadsheet. Microsoft Excel and Open Office are key tools for experimental scientists, and Google Refine is gaining popularity. If we could embed semantic annotation support into these tools, we could gather semantically enriched experimental data “by stealth”. RDF Linked Data generated from the spreadsheets would provide

a platform for cross spreadsheet integration and querying and therefore, semantically enriched content for e-science infrastructures.

In this paper we present RightField which is both (a) a lightweight spreadsheet-based annotation tool for lowering the barrier of manual metadata acquisition; and (b) a data integration application for extracting and querying RDF data from these enriched spreadsheets. We show that by hiding the complexities of semantic annotation from the end user scientist we can collect more accurate and rich metadata at source, demonstrating the advantages of adopting Semantic Web approaches to biological data integration. We illustrate the approach with results from the SysMO program, showing that RightField supports the whole workflow of semantic data collection, submission and RDF querying in Systems Biology. The RightField tool is freely available from http://www.rightfield.org.uk, and the code is open source under the BSD License.

## II. SEMANTIC ANNOTATION AND QUERYING

### A. *Requirements to Support Semantic Data Annotation*

Data integration relies on accurate and uniform annotation. In the Life Sciences, large public initiatives have generated controlled vocabularies and ontologies for annotating biological entities, such as the Gene Ontology [8] (for gene products) and ChEBI [14] (Chemical Entities of Biological interest, for metabolites and small molecules). Using such resources to semantically annotate biological entities with their functional properties enables integration and inferences across data sets.

The metadata standards and ontologies in the Life Sciences are essential for data sharing and reuse, but achieving compliance can add a considerable overhead to data management. For scientists working in multidisciplinary areas, like Systems Biology, data annotation involves understanding and tracking developments in many community standards and ontology development initiatives. This only occurs on a large scale when it is a mandatory step for publication. For microarray data, for example, scientists cannot submit a journal publication before data is compliant with the community metadata standard (MIAME [15] - Minimum Information About a Microarray Experiment), which includes annotation with terms from recommended ontologies. For other data types, there are similar minimum information specifications, (under the umbrella of MIBBI [16], Minimum Information for Biological and Biomedical Investigations), but most are optional for publication so the uptake of these is lower. The problem is further exacerbated by the fact that some MIBBI models are provided as XML schemas, which many scientists have limited experience of, or they are simple guidelines that do not assist scientists with data formatting.

If we instantiate these standards and vocabularies into tools that scientists already use to capture and store data (namely, spreadsheets), we lower the barriers to standards compliance for the researcher. In effect, standards-compliant, semantic annotation becomes part of the laboratory data management process, and can be carried out by the researchers generating the data. The requirements for semantic annotation are therefore as follows

1. support acquisition of annotation at source;
2. use familiar tools and fit into the laboratory data management process without change to those tools – i.e. no additional plug-ins or libraries;
3. collect data offline;
4. collect consistent data across sets of experiments;
5. cope with heterogeneous data; and
6. conform to community standards and vocabularies.

### B. *Requirements to Support Data Extraction and Querying*

In large research consortia typical in Systems Biology, like SysMO, data is continually produced and submitted, (and potentially updated) from multiple physical locations. As new experiments are devised, new types of data are deposited and compared to existing data locally or in public repositories. In order to support this process, new data must be assimilated quickly and automatically from distributed locations, and tools must assist users in forming queries across both their own repository and public data sources. Figure 1 shows the flow of work from data acquisition and annotation, to data extraction and querying. In the next section, we show how the application of RightField supports and enables each phase.

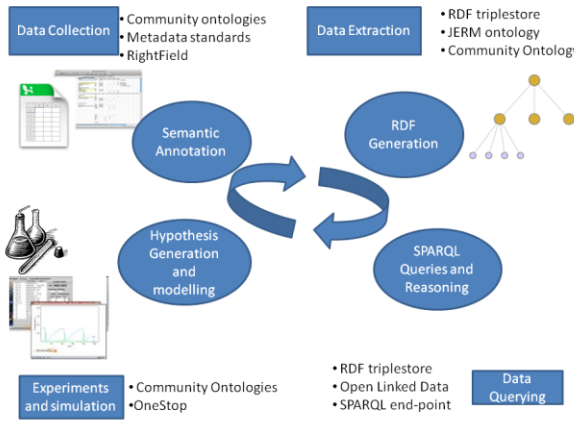


**Figure 1.** A flow diagram of data acquisition and querying in Systems Biology, using RightField.

## III. INTRODUCING RIGHTFIELD

RightField is an open-source, cross-platform Java application that provides a mechanism for embedding ontology annotation support for data in Excel or Open Office spreadsheets. We present its capabilities in three steps:

**Step 1:** The RightField client is used to define MS Excel or Open Office templates for an experiment. Individual cells,

columns, or rows can be restricted to display particular ranges of allowed classes or instances from multiple chosen ontologies (Figure 2a). Ontology *properties* (data or object type) can also be defined for spreadsheet cells.

**Step 2:** The RightField-enabled spreadsheets are distributed to the experimental scientists who use a regular Excel or Open Office Application to open them. The selected ontology terms are presented as a simple drop-down list, enabling scientists to consistently annotate their data without

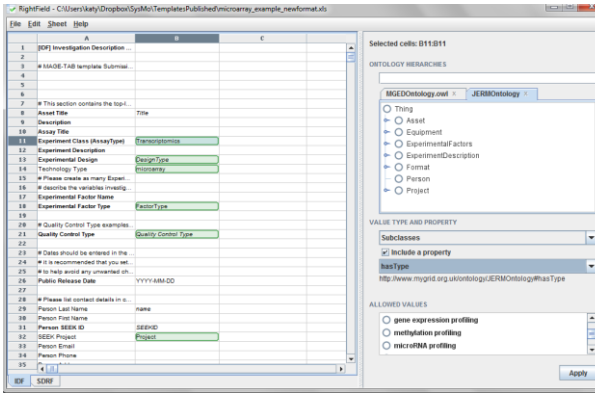

**Figure 2a:**The RightField application showing a spreadsheet template (left) being marked-up with terms from an ontology (right).

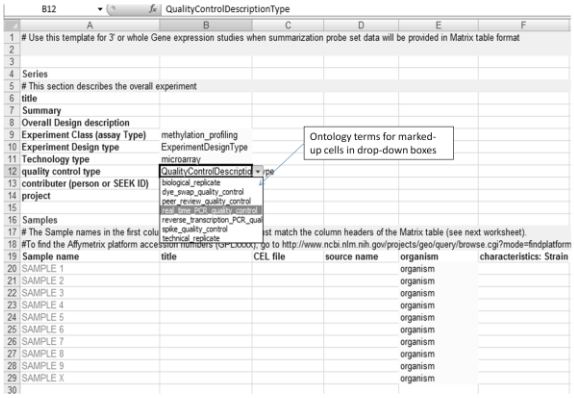


**Figure 2b:** A RightField-enabled spreadsheet showing the dropdown lists of ontology terms presented to the user for annotation.

requiring detailed knowledge of the ontology structures and content (Figure 2b). The architecture is such that no additional plug-ins, special macros or scripts are required for the Excel/Open Office applications.

**Step 3:** By defining the classes, instances and properties of each cell, RDF statements can be automatically generated for each cell and therefore an RDF graph, or collection of graphs can be generated for each data set. RightField produces RDF, but these statements can be extracted to comply with richer models in OWL or RDFS to allow more complex reasoning. Depositing this RDF to a triple-store provides a rich, querying environment that allows the scientists to search their data and other Linked Data resources interchangeably.

The architecture has been designed with several features to fulfill the requirements (Figure 3). The Apache POI library is used to read and manipulate spreadsheets, and terms are applied to cells using data validation provided by Apache POI.

*Access to standard ontologies*

The tool supports annotation with OWL, OBO and RDFS ontologies and RDF vocabularies, available from local file systems, a URL, or from the BioPortal [1] repository of biological ontologies available at http://bioportal.bioontology.org/. The Protégé OWL API is used to read and process ontology files. This satisfies requirement 5. The heterogeneity of the data means that each experiment may require annotation with terms from multiple ontologies in multiple formats.

*Ontology embedding and encapsulation*

Once marked-up and saved the full IRIs (Internationalized Resource Identifiers), label and version for selected elements of the ontology are stored within hidden sheets. These IRIs are used for both data provenance, to link back to the community ontologies; and for extracting and storing information in RDF. The use of hidden sheets enables the spreadsheet to be self-contained and avoids the use of customized extensions (requirement 2). This method of ontology term encapsulation is essential for the efficient performance of RightField and for its mode of use. The spreadsheets must be available for scientists to work offline and to continue to function in the event that the ontology server is unavailable (in particular for ontologies sourced through BioPortal, as this would constitute a single point-of-failure) (requirement 3). By encapsulating only the terms required in the spreadsheet, their IRIs, and the version of the ontology they originate from, no live link is required between the ontologies and the spreadsheets. This also eliminates any requirement to encapsulate whole ontologies. Life Science ontologies can be large (e.g. the Gene Ontology contains over 37500 terms), so the size could potentially affect performance.

The fixing of ontology terms is also a desirable property (requirement 4). Since there is no live link between the ontology and the RightField-enabled spreadsheet, there is no mechanism for updating ontology terms when ontologies are updated, until spreadsheets are re-opened in RightField. This is a deliberate design decision which ensures that a series of experiments can be annotated with the same version of the same ontologies. If ontology versions changed during experiments, annotating the data with different versions could make comparisons and integration difficult. Ontology updates should be performed as part of a data curation process, where all data from a series could be updated, or all data could preserve previous versions. When RightField-enabled spreadsheets are re-opened in RightField, the tool compares embedded ontology versions to live ontologies. If they have been updated, RightField warns the user and offers the possibility of updating them.

*RDF export*

Once annotated, RightField spreadsheet data can be automatically exported as RDF. This is achieved by reading the embedded information hidden within the spreadsheet to determine the term, property, and ontology related to each annotation. An identifier for the spreadsheet itself is required and is provided by the user of the graphical user interface (GUI), or software client, using the API. The RDF is then composed with the aid of the Jena RDF libraries. Within the RightField GUI the RDF can be stored to a file, but when using the RightField API it may be exported to an arbitrary stream.

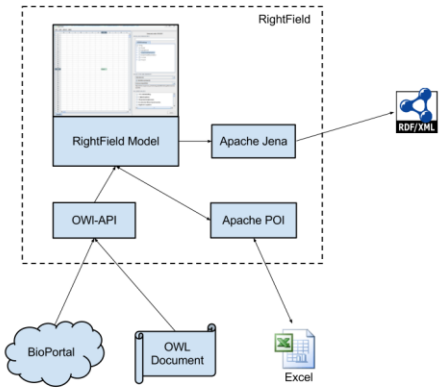


**Figure 3:** The architecture of RightField

RightField is in routine use in SysMO. Examples of spreadsheet templates can be found at http://www.rightfield.org.uk. Templates are developed centrally and distributed to the consortia, to promote data consistency. Figure 4 shows an excerpt from a template, showing how the classes and properties relate to particular cells, using the JERM Ontology (as described in section IV). The RightField templates encapsulate community standards (requirement 1) without requiring users to adopt additional tools, acquire detailed knowledge of community metadata standards or ontologies, or be exposed to the underlying Semantic Web technologies.

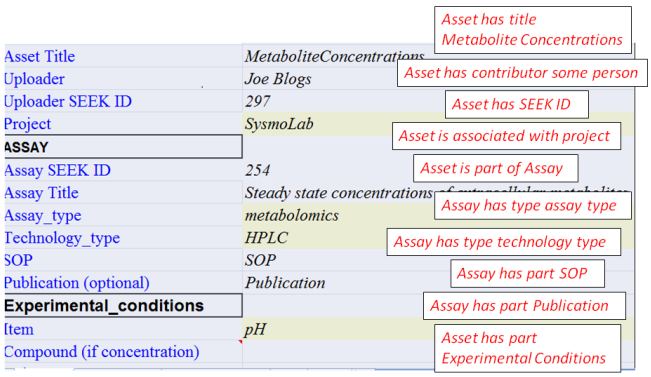


**Figure 4**: A RightField template showing the underlying ontology properties that describe the relationships between the dataset and the metadata elements it is annotated with.

## IV. Evaluation: Using RightField for Semantic Annotation

The main hypothesis of this work is that the use of RightField results in datasets that are standards-compliant and semantically annotated, and that the annotation is consistent and more complete. To assess completeness and consistency, we can examine RightField-enabled data collected during the SysMO project. To date, the consortium has produced and shared over 900 datasets.

RightField was only introduced mid-way through the project. Therefore, we have a heterogeneous collection of data submitted pre-RightField, data submitted pre-RightField which has since been transformed to RightField templates, and data submitted directly using RightField templates. For this evaluation, we examined a collection of datasets that existed as different versions pre and post RightField, so that individual datasets were directly comparable.

### *A. Metadata Completeness*

To assess completeness of metadata, we compared each dataset to the SysMO-JERM metadata checklist. The JERM is the "Just Enough Results Model" a minimum information model (and ontology) to describe the relationship between different data sets and mathematical models in SysMO. It is described in detail in [17], but the checklist is available at https://seek.sysmo-db.org/help/metadata-guidelines. Most metadata elements are mandatory, but some are optional and some are mandatory only for certain types of data. For example, all data sets have a *title* and a *creator*, but when describing a biological sample, recording the phenotype is optional, and for some high throughput experiments involving microarrays or mass spectrometry, for instance, recording the normalization method is mandatory, but other types of data do not require normalization.

For each dataset in the study, we gave a score of 2 for each general mandatory metadata element recorded and a score of 1 for each data-specific mandatory metadata element. Optional elements were also given a score of 1. To reduce the problems associated with different numbers of data-specific mandatory metadata elements, the study only included data relating to metabolomics. The study contained 5 different datasets, describing 11 experiments and 62 samples. Table 1 shows the results of analyzing pre- and post-RightField metabolomics datasets.

| JERM Metadata Element Scores | | |
|---|---|---|
| ***Dataset ID*** | ***RightField Template*** | ***Pre-RightField Template*** |
| 598 | 616 | 244 |
| 599 | 319 | 402 |
| 72 | 119 | 85 |
| 868 | 203 | 62 |
| 69 | 127 | 88 |

**Table 1:** A comparison of metadata elements recorded for SysMO metabolomics datasets before and after RightField

For all comparisons, the RightField-enabled spreadsheets scored significantly higher than pre-RightField spreadsheets, resulting in more complete metadata descriptions of the datasets. However, there is a large variation of scores within both the RightField set and the pre-RightField set. This is due to the fact that the JERM checklist covers experiment metadata and samples metadata. Therefore, the greater the number of samples in a dataset, the higher the potential JERM compliance score. Dataset 598, for example, describes 28 different samples, whereas dataset 72 only describes 3.

### B. *Consistency of Annotation*

Consistent metadata elements across datasets assist comparisons between them, but free text annotations in these elements can add ambiguity. Simple mapping between synonyms (e.g. yeast, *Saccharomyces cerevisiae* and S. cerevisiae) can be achieved without a semantic framework, but classifications of experimental types, for example, and the experimental conditions and factors studied in each require more precision. The JERM ontology (available from http://bioportal.bioontology.org/ontologies/1488) provides a vocabulary for these concepts in RightField-enabled spreadsheets, along with other community ontologies.

Out of the five pre-RightField datasets, only two defined the type of experiment they described, and both used the same terms for that description. However, for the other three, it was only possible to determine the type of experiment being described by human inference over the data and results sheets. For experimental conditions and factors studied, only pH and temperature were described consistently. Other factors, like growth media, buffers, or metabolite concentrations, were inconsistent. The JERM checklist recommends the annotation of all biological objects with terms from community ontologies or databases, but in the pre-RightField datasets, they were simply recorded with free text (e.g. the concentration of glucose, or MES buffer).

In the post-RightField spreadsheets, every dataset defined the type of experiment being performed and the experimental conditions and factors studied using terms from the JERM ontology. Consequently, there is much less ambiguity involved in comparing the experiments. For pre- and post-datasets, the JERM metadata requirements were the same, but without the drop-down lists assisting experimentalists, much of this information was omitted. These results show that the introduction of RightField improves both the completeness and consistency of annotation.

## V. USING RIGHTFIELD TO GENERATE AND EXTRACT RDF

One of the major purposes of improving data annotation is to enable greater discovery and reuse by others. By extracting and storing RightField data in RDF, we can make it available for querying by a much larger community, in conjunction with other available resources.

RightField allows the encapsulation of a metadata model and its associated ontologies. Due to the innate flexibility of RDF, there are many ways this information can be extracted and expressed. There is no 'standard' RDF format for MIBBI metadata models, so a direct comparison with community resources is not currently possible (although the ISA standard is working towards an RDF export specification. See related work section for further details). Therefore, in order to evaluate the effectiveness of the approach, we must consider how the RDF performs against more conventional relational database solutions, and how RDF supports the integration of data as well as allowing the aggregation of data from other sources.

### A. *Extracting and Storing RightField Data*

A test data set was created for this evaluation by extracting public data from ArrayExpress and marking it up with RightField. Data submitted to ArrayExpress should already conform to the MIAME specification (Minimum Information About a Microarray Experiment), so we can assume a high level of metadata completeness. The purpose of RightField-enabling this data was simply to ensure a connection between metadata elements and community ontologies, and to specify the properties of associations between the datasets and the metadata annotation (as specified in Figure 4).

The dataset was selected by searching for ArrayExpress [18] submissions related to the organism *Lactococcus lactis*. The IDF (Investigation Description Format) and SDRF (Sample and Data Relationship Format) metadata files were extracted, which describe the overall experimental aims, the origins of the experiment; and the conditions of the samples used. Once marked-up, the data was automatically extracted to RDF and stored in a Virtuoso triple store. The contents of this test triple store currently contains 9907 triples.

Although the test dataset was MIAME compliant, there are many optional metadata elements for describing the experiment samples in MIAME. Therefore, across the datasets, there is large variability. For example, 2 datasets describe a particular strain of the organism, whereas others simply record the organism name. For environmental conditions, some describe the components of the growth media and the conditions of the culture, whereas others reference a protocol.

To construct a relational database to house MIAME-compliant data, we would require a complete relational model that describes the whole MIAME specification of mandatory and optional elements. For many records, database fields would remain empty for some optional metadata elements. As the schema changes over time and is potentially extended, the underlying model would have to be altered to reflect this.

For RDF, this variability is not an issue. The RDF graph for each dataset exports only those metadata elements present, and there is no requirement for an over-arching model of MIAME. If a new dataset was included with extensions to the MIAME specification, no re-modeling would be required. Similarly, with the RDF model, scientists could immediately use RightField RDF extraction

to capture data from other types of experiments (using other metadata models), allowing aggregation across different data sets. The JERM models developed in SysMO, for example, could be aggregated with the MIAME test dataset. The JERM microarray model is also MIAME compliant and can be directly compared, but other data, for example, for metabolomics or proteomics, could be aggregated at the level of biological samples and experimental conditions. Comparisons across omics data are essential for an understanding of the dynamic processes of whole biological systems, which is a central concern in Systems Biology.

### *B. Querying the RDF Extracted from RightField*

The data contained in the example triple store has rich descriptions of experimental designs and samples. This dataset can be queried using the public SPARQL interface at http://escience.rightfield.org.uk/sparql. The following SPARQL query retrieves all data associated with the MGED:ExperimentalFactorCategory, for example:

```
SELECT ?df ?p ?instance
WHERE {
 {
   ?subject rdfs:subClassOf
mged:ExperimentalFactorCategory option (transitive) .
   ?instance a ?subject .
}
GRAPH <rf:test> {
   ?df ?p ?instance
}
}
```

The MGED:ExperimentalFactorCategory terms describe the biological, environmental and methodological factors influencing the experiment. For microarrays, this includes the BioMaterial (e.g. whole organism, total RNA, cytoplasmic RNA), age, cell line, disease state, etc; the environmental factors like temperature and growth media; and methodological factors like experimental protocols and equipment used.

In this example, a total of 81 triples are retrieved, which describe the hybridization methods, the labeling molecules, and the types of extraction and image acquisition for each experiment.

For the SysMO-SEEK, a SPARQL end-point will be provided for advanced users, but regular access will be through common, canned queries and query templates in the SEEK web interface. The underlying Semantic Web technologies will again be hidden from scientists.

Queries can also include access to external resources. This allows scientists to place their data into context with public data. For example, the local store can be queried for all data relating to the organism *Lactococcus Lactis* in conjunction with the same query through the ArrayExpress SPARQL endpoint (http://www.ebi.ac.uk/efo/semanticweb/atlas). The retrieved RDF graphs can be aggregated due to the common ontology IRIs used for annotation.

## VI. Related Work

RightField is an application that addresses the whole workflow of data collection, annotation, RDF representation and querying in Systems Biology. There are therefore a number of other resources that address one or more of the same issues, but RightField is unique in its approach to support the process end to end. Related work therefore encompasses BioSharing initiatives, using spreadsheets for knowledge acquisition, and extracting spreadsheet data to RDF.

### *A. BioSharing*

BioSharing.org is a global initiative to co-ordinate the standardization of biological data in order to promote data sharing and interoperation [19, 20]. This organization catalogues reporting standards (formats, terminologies and checklists), as well as developing data sharing policies. BioSharing stops short of producing tools, but RightField and related applications build on their standards.

The ISA tool-suite also builds on BioSharing resources. ISA tools [12] are a set of applications designed to create and manage ISA-TAB files, a tabular format for describing the relationships between different experiments (in fact, the SysMO-JERM implements an ISA-TAB model).

The ISA tools are implemented in bespoke client software. They have the same *'look and feel'* of spreadsheets, but they are not designed for laboratory scientists, but more for informaticians and data management experts. The 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. However, it is a pre-configured application that is designed to work only with the ISA-TAB specification and associated vocabularies, whereas RightField can be configured to use any ontology and metadata schema. The ISA suite will soon offer the ability to automatically convert ISA-TAB to RDF. Since the SysMO-JERM follows the ISA specification, queries between ISA RDF and SysMO RDF should be straightforward.

Other ontology annotation tools in the Life Sciences include Phenote, which assists with the annotation of biological phenotypes (http://phenote.org), and the PRIDE Proteome Harvest Spreadsheet submission tool (http://www.ebi.ac.uk/pride/proteomeharvest/), which assists with the annotation and submission of proteomics data to the PRIDE public repository. These are powerful annotation tools for specific biological disciplines, and are not generically applicable.

### B. *Spreadsheets for knowledge acquisition and manipulation*

Google Refine (http://code.google.com/p/google-refine/) is a tool designed to help users deal with inconsistent data in spreadsheets. It allows spreadsheet manipulation, format conversion and the incorporation of extra data via web services or databases. Google Refine is therefore a useful tool for data curation and is designed to manage legacy data, rather than in the creation of new data. RightField is designed to improve the accuracy of data as it is collected, so the aims are quite different, but they both address the same fundamental problem of inconsistencies in data.

DataScopes, from MicroSoft, is an application that provides facilities for running analytics over spreadsheets and for linking to analyses and other data in the cloud. It is therefore much more related to the querying of the extracted RightField data

### C. *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), XLWrap, and RDF123, all perform this function. However, 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. Also, RightField does not require a separate mapping file since this information is self-contained. Therefore, cells can be moved around or copied without affected the expected RDF produced. The Anzo platform (http://www.cambridgesemantics.com/) is a commercial product with similar goals, focusing on spreadsheet coordination and coherency through a common RDF bus.

## VII. DISCUSSION AND FUTURE WORK

RightField-enabled spreadsheets show a marked increase in the consistency of annotation when compared with free text annotation or other template approaches. The success of RightField comes from embedding functionality in tools that are already familiar to the people collecting data. The result is that semantic annotation and metadata management becomes part of the day-to-day data management process.

Most experimental biologists have little experience in the use of ontologies, and moreover see no immediate personal benefits in semantically annotating data for the community. However, they understand the value of such annotations when attempting to reuse data from others, and also when required to share data within large, distributed consortia, like SysMO. RightField helps address the pre-publication data sharing bottleneck as well as public data sharing. It provides a framework for sharing data without necessarily publishing it since the extracted RDF can be shared, *or* the semantic spreadsheets can be exchanged directly. For the individual scientist, the benefit of using RightField is the ability to immediately share with their consortium. The added benefits of making the data amenable to Semantic Web querying is a side-effect of addressing this original problem.

RightField is a key part of the SysMO-DB project and subject to active development. We are currently working on several improvements:

**Large ontologies**: RightField displays *all* classes and/or individuals from a chosen section of an ontology. If it has a shallow structure, with hundreds of sibling classes at any given level, the numbers of terms in the drop-down box becomes unmanageable. Current work will allow auto-complete style searches in the drop-down boxes, and an advanced option could show hierarchical relationships between terms. The requirement for these extra features pinpoints areas where ontologies designed as a vocabulary for data annotation need further development (i.e. where classes should be further defined to assist users with their classification). For ontologies accessed through the BioPortal, submitting feedback through RightField would be valuable.

**Ontology label ambiguities.** RightField hides most of the complexity of the underlying ontologies away from end users, but in cases where term labels describe similar concepts, the ability to access and compare the ontology definitions would be useful.

**Linked Data output**. We will add VOID support to RightField to enable dataset discovery and tracing, and we will provide a mechanism to publish RightField RDF with persistent URLs. Releasing the data to the community for long term access and reuse is a requirement of the SysMO project, and releasing it as RDF would allow interoperability with related resources.

RightField is already in use in Systems Biology, but managing heterogeneous data with complex metadata is a common problem. Work has also started in projects in other disciplines., For example, to build knowledge bases for Kidney and Urinary Pathways, inflammatory bowel disease and Chagas disease. Further afield, RightField is being used in archaeology, for the annotation of historical samples. In particular, for developing 'patient records' for Egyptian mummies, in collaboration with Manchester Museum.

Organizations, like BioSharing.org, international data repositories, and national funding councils are encouraging individuals and academic institutions to release more of their data. However, making data available is only the first step. Data must also be computationally accessible so that scientists can discover and evaluate it. Semantic Web technologies should be ideal for exploring the complex networks of genes, proteins and metabolites that interact in biological systems. They provide cutting-edge methods from computer science to address problems that are not only

confined to the Life Sciences, but span most science disciplines. A challenge is making these (often new) technologies accessible to scientists, without significantly increasing workloads. We have shown that the instrumentation of widely used, commodity applications is an effective approach.

## REFERENCES


[1] Kandpal, R., Saviola, B., Felton, J.: The era of 'omics unlimited. Biotechniques **46** (2009) 351-352, 354-355

[2] Juty, N., Le Novere, N., Laibe, C.: Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res **40** D580-586

[3] Antezana, E., Blonde, W., Egana, M., Rutherford, A., Stevens, R., De Baets, B., Mironov, V., Kuiper, M.: BioGateway: a semantic systems biology tool for the life sciences. BMC Bioinformatics **10 Suppl 10** (2009) S11

[4] Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform **41** (2008) 706-716

[5] Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., Wild, D.J.: Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics **11** 255

[6] Ball, C.A., Brazma, A.: MGED standards: work in progress. OMICS **10** (2006) 138-144

[7] Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M., Ashburner, M.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res **36** (2008) D344-350

[8] Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R.: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res **32** (2004) D258-261

[9] Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, II, Hedley, W.J., Hodgman, T.C., Hofmeyr, J.H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novere, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J.: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics **19** (2003) 524-531

[10] Rubin, D.L., Shah, N.H., Noy, N.F.: Biomedical ontologies: a functional perspective. Brief Bioinform **9** (2008) 75-90

[11] Sansone, S.A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L.A., Copeland, J., Das, S., de Daruvar, A., de Matos, P., Dix, I., Edmunds, S., Evelo, C.T., Forster, M.J., Gaudet, P., Gilbert, J., Goble, C., Griffin, J.L., Jacob, D., Kleinjans, J., Harland, L., Haug, K., Hermjakob, H., Ho Sui, S.J., Laederach, A., Liang, S., Marshall, S., McGrath, A., Merrill, E., Reilly, D., Roux, M., Shamu, C.E., Shang, C.A., Steinbeck, C., Trefethen, A., Williams-Jones, B., Wolstencroft, K., Xenarios, I., Hide, W.: Toward interoperable bioscience data. Nat Genet **44** 121-126

[12] Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., Neumann, S., Sterk, P., Tong, W., Sansone, S.A.: ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics **26** 2354-2356

[13] Nelson, B.: Data sharing: Empty archives. Nature **461** (2009) 160-163

[14] Degtyarenko, K., Hastings, J., de Matos, P., Ennis, M.: ChEBI: an open bioinformatics and cheminformatics resource. Curr Protoc Bioinformatics **Chapter 14** (2009) Unit 14 19

[15] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M.: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet **29** (2001) 365-371

[16] Taylor, C.F., Field, D., Sansone, S.A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C.A., Binz, P.A., Bogue, M., Booth, T., Brazma, A., Brinkman, R.R., Michael Clark, A., Deutsch, E.W., Fiehn, O., Fostel, J., Ghazal, P., Gibson, F., Gray, T., Grimes, G., Hancock, J.M., Hardy, N.W., Hermjakob, H., Julian, R.K., Jr., Kane, M., Kettner, C., Kinsinger, C., Kolker, E., Kuiper, M., Le Novere, N., Leebens-Mack, J., Lewis, S.E., Lord, P., Mallon, A.M., Marthandan, N., Masuya, H., McNally, R., Mehrle, A., Morrison, N., Orchard, S., Quackenbush, J., Reecy, J.M., Robertson, D.G., Rocca-Serra, P., Rodriguez, H., Rosenfelder, H., Santoyo-Lopez, J., Scheuermann, R.H., Schober, D., Smith, B., Snape, J., Stoeckert, C.J., Jr., Tipton, K., Sterk, P., Untergasser, A., Vandesompele, J., Wiemann, S.: Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol **26** (2008) 889-896

[17] Wolstencroft, K., Owen, S., du Preez, F., Krebs, O., Mueller, W., Goble, C., Snoep, J.L.: The SEEK: a platform for sharing data and models in systems biology. Methods Enzymol **500** 629-655

[18] Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., Brazma, A.: ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res **35** (2007) D747-750

[19] Field, D., Sansone, S., Delong, E.F., Sterk, P., Friedberg, I., Gaudet, P., Lewis, S., Kottmann, R., Hirschman, L., Garrity, G., Cochrane, G., Wooley, J., Meyer, F., Hunter, S., White, O., Bramlett, B., Gregurick, S., Lapp, H., Orchard, S., Rocca-Serra, P., Ruttenberg, A., Shah, N., Taylor, C., Thessen , A.: Meeting Report: BioSharing at ISMB 2010. Stand Genomic Sci **3** 254-258

[20] Orchard, S., Albar, J.P., Deutsch, E.W., Eisenacher, M., Vizcaino, J.A., Hermjakob, H.: Enabling BioSharing - a report on the Annual Spring Workshop of the HUPO-PSI April 11-13, 2011, EMBL-Heidelberg, Germany. Proteomics **11** 4284-4290


## ACKNOWLEDGMENT


We thank Matthew Horridge for his initial work on early versions of RightField, and the SysMO-DB team and SysMO PALS for their valuable feedback and testing. This work was funded as part of the SysMO-DB2 grant awarded by the BBSRC (BB/I004637/1).