RightField: Embedding Ontology Term Selection into Spreadsheets for the Annotation of Biological Data

Katy Wolstencroft[1], Matthew Horridge[1], Stuart Owen[1], Wolfgang Mueller[2], Finn Bacall[1], Jacky Snoep[1], Olga Krebs[2], Carole Goble[1]

[1] School of Computer Science, University of Manchester, Manchester, M13 9PL, UK
[2] HITS gGmbH, Schloss-Wolfsbrunnenweg 35, Heidelberg, Germany
<given.family@manchester.ac.uk, given.family @h-its.org>

Abstract. RightField is an open source application that provides a mechanism for embedding ontology annotation support for Life Science data in Microsoft Excel spreadsheets. Individual cells, columns, or rows can be restricted to particular ranges of allowed classes or instances from chosen ontologies. Informaticians, with experience in ontologies and data annotation prepare RightField-enabled spreadsheets with embedded ontology term selection for use by a wider community of laboratory scientists. The RightField-enabled spreadsheet presents selected ontology terms to the users as a simple drop-down list, enabling scientists to consistently annotate their data without the need to understand the numerous metadata standards and ontologies available to them. The spreadsheets are self-contained and remain "vanilla" Excel so that they can be readily exchanged, processed offline and are usable by regular Excel tooling. The result is semantic annotation by stealth, with an annotation process that is less error-prone, more efficient, and more consistent with community standards. RightField has been developed and deployed for a consortium of some 300 Systems Biologists. RightField is open source under a BSD license and freely available from http://www.sysmo-db.org/RightField.

Keywords: ontology annotation, biology, metadata standards, spreadsheets

1 Introduction

In the post-genomic era, the quantity and complexity of biological data produced during standard laboratory projects has increased. New techniques and technologies, in areas such as transcriptomics and proteomics, enable scientists to produce high volumes of data in single experiments. In order to compare and reuse this data, however, rich metadata annotation is also required. The cost of this annotation is high and it is a time-consuming and undervalued process.

In the biological sciences, guidelines and checklists describing what metadata is required for the interpretation and reuse of data are emerging. They are often specified as *minimum information models* [1] with associated controlled vocabularies or ontologies that define the terms that should be used to describe these metadata elements. In some cases, for example, for microarray data, publication submissions are

not accepted unless the accompanying data is compliant with the relevant minimum information model (for microarrays, this is MIAME, the Minimum Information about a Microarray Experiment). However, despite this drive to standardization, there are few tools to help scientists manage this process. RightField was created to lower the barrier of uptake by providing a mechanism for scientists to produce ontology annotation from *within the software environments they already use*.

RightField was developed as part of the SysMO-DB project, which supports a consortium of more than 300 Systems Biologists with data management and exchange. SysMO is a pan-European project to study the Systems Biology of Micro-Organisms, which involves a mixture of high-throughput '*omics* experiments, such as microarray analysis or proteomics, as well as traditional molecular biology and enzyme reaction kinetics. In SysMO-DB, data is standardized by providing spreadsheet templates for different types of experiment to conform to the "Just Enough Results Model" (JERM). The JERM is the SysMO-DB internal structure that describes what type of experiment was performed, who performed it, and what was measured. For experiment types with an established minimum information model, the JERM also complies with this. By combining JERM templates and embedded ontology terms with RightField we provide an infrastructure that promotes and encourages compliance and standardization.

2 Data Generation, Annotation and Reuse

RightField was designed to support a community of laboratory scientists with little experience of metadata management, ontologies or standardization. The primary objective was to provide an application that would allow consistent annotation without changing working practices. Understanding the life-cycle of data generation, annotation and reuse is vital in this process. Capturing experimental metadata at the time of the experiment increases accuracy and increases the likelihood that the annotation is provided by the person performing the experiment. Using the same versions of ontologies for a series of experiments is also vital for accurate comparisons. RightField was designed to be a spreadsheet annotation tool because spreadsheets, particularly MS Excel, are ubiquitous in the laboratory science community for organizing and managing experimental data. Embedding annotation terms in the spreadsheets ensures that term selection occurs at the time of the experiment within the application already in use.

RightField is an open-source, cross-platform Java application which uses Apache-POI for interacting with Microsoft documents. It does not require any special macros, visual basic code or platform specific libraries to run it. It enables users to upload Excel spreadsheets along with ontologies from their local file systems, or from the BioPortal [2] (a repository of biological ontologies available at http://bioportal.bioontology.org/). RightField supports OWL, OBO and RDFS ontologies and RDF vocabularies. In the uploaded spreadsheet, individual cells, or whole columns or rows can be marked with the required ranges of ontology terms. For example, they could include all subclasses from a chosen class, direct subclasses only, all individuals, or only direct individuals. Each spreadsheet can be annotated with terms from multiple ontologies.

Once marked-up and saved, the RightField-enabled spreadsheet contains embedded worksheets with information concerning the origins and versions of ontologies used in the annotation. *This encapsulation stage is crucial*. With everything embedded in the spreadsheet, scientists do not require any new applications to use it and they can complete annotation offline should they wish. This also makes the spreadsheets readily exchangeable and enables a series of experiments to be annotated with the same versions of the same ontologies even if the live ontologies change during this time.

3 RightField Annotation – A Case Study

A research group is studying the impact of changes in flux for different nutrient limitation conditions in *Saccharomyces cerevisiae*. They perform transcriptomics, metabolomics and proteomics experiments and integrate the results. Each experiment is high-throughput; generating complex data which needs to be annotated with rich metadata concerning the experimental conditions, the methods and equipment.

The transcriptomics experiments represent a series which will be compared and analysed together. For publication, this data must conform to the MIAME standard and be deposited in a public microarray repository, such as ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/) or GEO (http://www.ncbi.nlm.nih.gov/geo/). ArrayExpress provides a service to auto-generate a MIAME compliant template suitable for a particular experiment, but this does not include annotation terms for use in the template. Uploading this auto-generated template into RightField enables an informatics expert to preset ranges of values for annotation.

Figure 1A shows the RightField annotation tool being used to preset annotation values from the MGED ontology (http://www.mged.org/) into the auto-generated MIAME-compliant template, and 1B shows the resulting template with drop-down lists of ontology annotation terms. The marked-up spreadsheet is distributed to the experimentalists to standardize the information that can be recorded and the terms that can be used for annotation. The result is ontology annotation by stealth. The experimentalists do not require specialist knowledge of the ontology resources used.

4 Discussion

RightField is a tool with *light-touch* use of semantic web technologies. The novel part of this work lies in disguising the use of semantics from the end users. Simplicity and unobtrusive embedding with a widely used application is its strength. To compare with similar efforts: ISA Creator (http://isatab.sourceforge.net/isacreator.html) is a bespoke spreadsheet tool designed for experts not end users; the Anzo platform is a commercial product with similar goals (http://www.cambridgesemantics.com/).

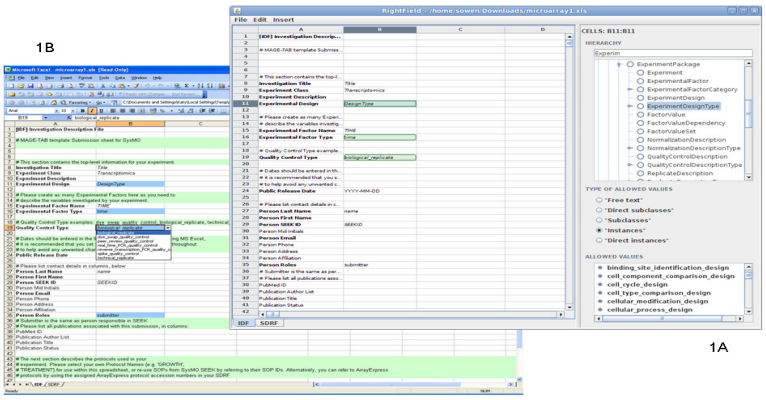

Figure 1 A&B: RightField and a resulting ontology term-embedded spreadsheet

Many experimental biologists have no interest or experience in the use of ontologies and terminologies, but the data they produce is difficult to interpret or reuse without a shared understanding that can be gained from the use of common vocabularies. RightField is the application that bridges this gap. Data can be annotated accurately and at source by the laboratory scientists. This reduces errors and it reduces the time it takes to annotate the data to comply with community standards. Crucially, RightField restricts the choices of annotation terms to a small and manageable set and to make that set accessible and understandable to the scientists.

The next steps for SysMO-DB are to develop methods to fully exploit the corpus of semantically annotated data from RightField Spreadsheets. Compliance with community ontologies means that we can already use Web Services from the BioPortal for term lookup and visualization. In addition, we are developing mechanisms to extract the structured Excel data in RDF to provide further means for searching across the content of spreadsheets and will allow SysMO data to be provided as open linked data. Early investigations using XLWrap [3] are promising.

Acknowledgments
This work was funded by the UK's BBSRC award BBG0102181.

References

1. Taylor, C.F., et al. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nature biotechnology, 26, 889-896.
2. Noy, N.F., et al. (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research, 37, W170-1733
3. Langegger A, Wöß W (2009): XLWrap – Querying and Integrating Arbitrary Spreadsheets with SPARQL. 8th Intl Semantic Web Conf, Washington D.C. LNCS 5823, Springer, 2009.