BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 14 2011, pages 2021–2022
doi:10.1093/bioinformatics/btr312

Databases and Ontologies Advance Access publication May 26, 2011

RightField: embedding ontology annotation in spreadsheets

Katy Wolstencroft[1,*], Stuart Owen[1], Matthew Horridge[1], Olga Krebs[2], Wolfgang Mueller[2], Jacky L. Snoep[3], Franco du Preez[3] and Carole Goble[1]

[1]School of Computer Science, [2]Heidelberg Institute of Theoretical Studies gGmbH, Schloß-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany and [3]MIB, University of Manchester, Kilburn Building, Oxford Road, Manchester, M13 9PL

Associate Editor: Martin Bishop

ABSTRACT

Motivation: In the Life Sciences, guidelines, checklists and ontologies describing what metadata is required for the interpretation and reuse of experimental data are emerging. Data producers, however, may have little experience in the use of such standards and require tools to support this form of data annotation.

Results: RightField is an open source application that provides a mechanism for embedding ontology annotation support for Life Science data in Excel spreadsheets. Individual cells, columns or rows can be restricted to particular ranges of allowed classes or instances from chosen ontologies. The RightField-enabled spreadsheet presents selected ontology terms to the users as a simple drop-down list, enabling scientists to consistently annotate their data. The result is 'semantic annotation by stealth', with an annotation process that is less error-prone, more efficient, and more consistent with community standards.

Availability and implementation: RightField is open source under a BSD license and freely available from http://www.rightfield.org.uk

Contact: kwolstencroft@cs.man.ac.uk

Received on January 26, 2011; revised on April 29, 2011; accepted on May 6, 2011

1 INTRODUCTION

The quantity and complexity of biological data produced during standard laboratory projects has increased with the availability of high-throughput technologies. Advances 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 required.

The biocuration and data standards communities have been addressing this issue by providing recommendations for annotation requirements. Minimum information models (Taylor *et al.*, 2008) with associated controlled vocabularies or ontologies that define the terms that should be used to describe metadata elements are being developed. In some cases, 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 allow data curators (informaticians and ontologists) to augment standards compliant spreadsheet templates with embedded ontology term selection. Once a RightField-enabled template has been prepared, laboratory scientists can use it in its native form to annotate their data using simple drop-down boxes.

RightField was produced by the SysMO-DB project. SysMO-DB is developing a platform for data management and exchange in Systems Biology, primarily for the pan-European SysMO consortium (Systems Biology of Micro-Organisms). SysMO-DB supports a consortium of over 300 scientists, from a variety of Life Science backgrounds, with little experience of metadata management, ontologies and standardization. Data is standardized by developing spreadsheet templates for different types of experiment and embedding ontology term selection, using RightField, within these templates. The combination of templates and RightField, provides an infrastructure that promotes and encourages compliance and consistent data annotation.

2 DATA ANNOTATION AND REUSE

RightField is an application developed to operate within spreadsheets since spreadsheet use is ubiquitous in laboratory sciences.

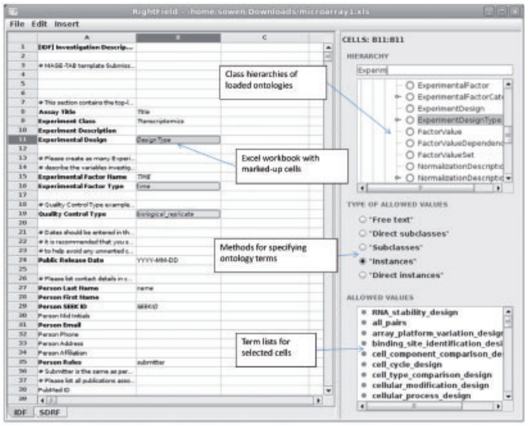
To use RightField, curators either upload an existing Excel spreadsheet, or create metadata entries in a new sheet. When the structure of the template is fixed, ranges of ontology terms can be embedded by either uploading ontologies from the BioPortal (Noy *et al.*, 2009), or from a local file in OWL, OBO, RDFS or RDF formats, Figure 1 shows a template being marked-up using RightField.

Individual cells, whole columns or whole rows can be marked up 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.

RightField is intended for use by informaticians and ontologists because setting up templates requires some knowledge of which ontologies to use and when. The spreadsheets, however, are intended for use by laboratory scientists as well as informaticians. Once templates are marked-up and saved, users only interact with them in the native spreadsheet form.

RightField-enabled spreadsheets display simple drop-down lists of terms in any marked-up cells (Fig. 2). Users of the spreadsheet are restricted to select from the term labels in the dropdown boxes, but no information from the ontology is lost. The terms displayed are linked to encapsulated hidden sheets that contain

*To whom correspondence should be addressed.



RightField - /home/sowen/Downloads/microarray1.xls
File Edit Insert
Class hierarchies of loaded ontologies
Excel workbook with marked-up cells
Methods for specifying ontology terms
Term lists for selected cells
IDF
SDRF

	A	B	C
1	[IDF] Investigation Descrip...		
2			
3	# MAGE-TAB template Submiss...		
4			
5			
6			
7	# This section contains the top-l...		
8	Assay Title	Title	
9	Experiment Class	Transcriptomics	
10	Experiment Description		
11	Experimental Design	Design Type	
12			
13	# Please create as many Experi...		
14	# describe the variables investig...		
15	Experimental Factor Name	TIME	
16	Experimental Factor Type	time	
17			
18	# Quality Control Type example...		
19	Quality Control Type	biological_replicate	
20			
21	# Dates should be entered in th...		
22	# it is recommended that you s...		
23	# to help avoid any unwanted c...		
24	Public Release Date	YYYY-MM-DD	
25			
26	# Please list contact details in s...		
27	Person Last Name	name	
28	Person First Name		
29	Person SEEK ID	SEEKID	
30	Person Mid Initials		
31	Person Email		
32	Person Phone		
33	Person Address		
34	Person Affiliation		
35	Person Roles	submitter	
36	# Submitter is the same as per...		
37	# Please list all publications asso...		
38	PubMed ID		
39			

CELLS: B11:B11

HIERARCHY

Experim

○ ExperimentalFactor
○ ExperimentalFactorCat
○ ExperimentDesign
○ ExperimentDesignType
○ FactorValue
○ FactorValueDependenc
○ FactorValueSet
○ NormalizationDescriptio
○ NormalizationDescriptio

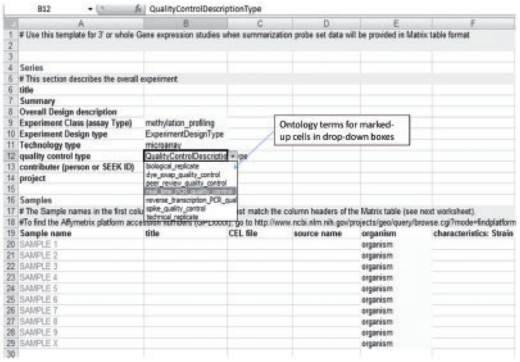
TYPE OF ALLOWED VALUES

○ "Free text"
○ "Direct subclasses"
○ "Subclasses"
● "Instances"
○ "Direct instances"

ALLOWED VALUES

- RNA_stability_design
- all_pairs
- array_platform_variation_design
- binding_site_identification_desi
- cell_component_comparison_de
- cell_cycle_design
- cell_type_comparison_design
- cellular_modification_design
- cellular_process_design

Fig. 1. RightField template generation, embedding ontology terms from the MGED ontology into an MS Excel spreadsheet for describing a transcriptomics experiment.



B12 fx QualityControlDescriptionType

	A	B	C	D	E	F
1	# Use this template for 3' or whole Gene expression studies when summarization probe set data will be provided in Matrix table format					
2						
3						
4	Series					
5	# This section describes the overall experiment					
6	title					
7	Summary					
8	Overall Design description					
9	Experiment Class (assay Type)	methylation_profiling				
10	Experiment Design type	ExperimentDesignType				
11	Technology type	microarray				
12	quality control type	QualityControlDescriptionType				
13	contributer (person or SEEK ID)					
14	project					
15						
16	Samples					
17	# The Sample names in the first colu...		...st match the column headers of the Matrix table (see next worksheet)			
18	#To find the Affymetrix platform acce...		...go to http://www.ncbi.nlm.nih.gov/projects/geo/query/browse.cgi?mode=findplatform			
19	Sample name	title	CEL file	source name	organism	characteristics: Strain
20	SAMPLE 1				organism	
21	SAMPLE 2				organism	
22	SAMPLE 3				organism	
23	SAMPLE 4				organism	
24	SAMPLE 5				organism	
25	SAMPLE 6				organism	
26	SAMPLE 7				organism	
27	SAMPLE 8				organism	
28	SAMPLE 9				organism	
29	SAMPLE X				organism	
30						

- biological_replicate
- dye_swap_quality_control
- peer_review_quality_control
- real_time_PCR_quality_control
- reverse_transcription_PCR_qual
- spike_quality_control
- technical_replicate

Fig. 2. A RightField-enabled spreadsheet showing drop-down list selection.

the full Internationalized Resource Identifiers (IRIs) for each term. Ontology IRIs must be captured as they are vital for data provenance and for extracting and processing data from the spreadsheets, but they do not need to be shown to users. Future versions will allow users to switch between labels and identifiers, providing an alternative *informatics view*. To view information in the hidden sheets, spreadsheets can be uploaded into the RightField Inspector application (http://inspect.rightfield.org.uk/).

The encapsulation of ontology information into the spreadsheet is crucial. It allows users to complete annotation offline when required and it ensures that a series of experiments can be annotated with the same versions of the same ontologies, even if the live ontologies change during this time. If users wish to update to a new version, the spreadsheet must be edited in RightField.

The recommended operational protocol is that RightField-enabled templates are created before a series of experiments, so that laboratory scientists can select metadata descriptions as and when they perform the experiments.

RightField is in use in the SysMO consortium. Examples can be found at http://www.rightfield.org.uk. As anticipated, RightField-enabled spreadsheets showed a marked increase in the consistency of annotation and the majority of SysMO evaluators found RightField useful.

The use of RightField, however, highlighted a wider issue with the structure of annotation ontologies. Some ontologies have shallow hierarchies. If users are expected to make a choice from a long list of classes (or individuals) in order to follow a community standard, perhaps this is an indication that a particular area of the ontology requires further development. RightField can address this issue from a technical perspective by providing auto-complete functionality for marked-up cells, but standardization should be encouraged by making the process straightforward. Defining more specific classes and/or splitting individuals between them would be of benefit to the whole community of ontology users.

3 RIGHTFIELD IMPLEMENTATION

RightField is an open-source, cross-platform Java application. It combines the use of the Apache POI library to read and manipulate spreadsheets and the Protégé OWL API to read and process ontology files. When a set of ontology terms are applied to cells, the identifiers and label for these elements of the ontology are stored within a hidden sheet in the spreadsheet file. The terms are applied to cells using data validation as provided by Apache POI. RightField-enabled spreadsheets can be opened in Microsoft Excel or in Open Office.

4 DISCUSSION

RightField provides a simple and unobtrusive way of using semantic web technologies to improve and standardize data annotation. The novel part of this work lies in shielding the use of ontologies from laboratory scientists, allowing them to continue working in familiar native spreadsheets. Other applications, such as the ISA Tools (Rocca-Serra *et al.*, 2010), provide similar functionality, but operate from bespoke client applications and are designed for more expert users. The Anzo commercial platform has similar goals (http://www.cambridgesemantics.com/).

The introduction of high-throughput, omics technologies has revolutionized experimental biology. The data produced in these experiments, however, is difficult to interpret or reuse without the use of common vocabularies. RightField is an application that bridges this gap. It restricts the choices of annotation terms to a manageable set, which is accessible and understandable to the scientists. It also lays the foundations for the structured extraction of spreadsheet data into Resource Description Framework (RDF) and Open linked Data (http://linkeddata.org/).

Ongoing development in RightField will explore the issues surrounding structured data extraction and greater validation of supplied annotations.

Funding: Biotechnology and Biological Sciences Research Council; SysMO-DB: Supporting Data Access and Integration (BBG0102181).

Conflicts of Interest: none declared.

REFERENCES

Noy,N.F. *et al*. (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.*, **37**, W170–W1733.

Rocca-Serra,P. *et al*. (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, **6**, 2354–2356.

Taylor,C.F. *et al*. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotech.*, **26**, 889–896.