# Ontologies Come of Age with the iKUP Browser

Simon Jupp[1], Julie Klein[2], Panagiotis Moulos[2], Joost Schanstra[2], and Robert Stevens[1]

[1] School of Computer Science, The University of Manchester, UK
[2] Institut National de la Santé et de la Recherche Médicale, Toulouse, France
first.last@manchester.ac.uk

The iKUP browser, a web-based interface to the kidney and urinary pathway knowledge base (KUPKB) [4], shows ontologies coming of age by enabling biologists to ask questions of integrated data that form hypotheses that are being tested in the laboratory. That is, *semantically* annotated data have been delivered to the target users in a form that they can use to change how they undertake their job. iKUP uses a browsing approach to query the KUP data as its users are usually not bioinformaticians that will design and use sophisticated scripts or workflows, and they are almost certainly not users familiar with semantic web technologies. The KUPKB contains data from high-throughput experiments on the kidney and the urinary system. The experimental data is richly *interconnected* to other biological data to form a single integrated repository for querying and exploration. The KUPKB uses multiple biomedical ontologies that act as a controlled vocabulary for standardised annotation of the datasets. These ontologies' semantics are used to ask queries that return *intelligent* answers that form part of the biologists' hypothesis generation process. By reducing the data to common representation languages like the Web Ontology Language (OWL) and the Resource Description Framework (RDF), we have shown semantic web technologies offering novel opportunities for data analysis.

The iKUP browser supports renal biologists in finding data integrated from many disparate data sources, providing a simple interface to survey a large set of the KUP domains 'omics experiments simultaneously. At present, biologists must gather and integrate many data sets by hand and this integration is vital as genes, proteins, and small molecules have to be co-ordinated across many 'omic levels through investigations reported by many people. By hand, this kind of integration and querying is long, tedious and error prone. Users can use the iKUP browser to search for a molecule (mRNA, miRNA, Protein) or list of molecules. The query box exploits the label and synonym tags for molecules to guide the user with their search *via* a dynamic suggestion box that pops up as users type. Once users confirm the molecules they are looking for are present in the KUPKB, the application performs a SPARQL query based on these search terms to generate the results. The results show known information about each molecule from the range of datasets in the KUPKB. For each result, the user can see where anatomically the molecule is active and under what conditions.

The metadata for each experiment is captured using ontologies. Datasets are collected in spreadsheet based templates that have ontologies embedded inside them. By using a spreadsheet template we provide a lightweight and familiar mechanism for the community of renal biologist to submit data. By embedding

the ontologies inside the spreadsheets, we are able to collect ontological annotations from the users by stealth. These *semantic spreadsheets* are created using RightField [7] which provides a mechanism to embed ontology based restrictions on values inside the spreadsheet. These consistent and precise metadata help to integrate the different experiments and are presented to the user on querying the KUPKB *via* iKUP. Every search result is accompanied by a navigable polyhierarchy that is generated from the ontological metadata. This hierarchy provides a faceted browser that exploits the semantics of the ontology to perform *intelligent* filtering of the data. For example, filtering the results on experiments on the glomerulus of the kidney will include in the results any experiments where the sample is glomerular or any part of the glomerulus, thus spanning from gross anatomy to the cellular level.

The KUPKB is public and open-access, and biologists are encouraged to submit new datasets to the KUPKB. So far, over 160 experiments have been sourced from the literature and public databases. All biological identifiers are normalised to a common identification scheme based on *de facto* URIs sourced from linked data resources, such as Bio2RDF [1] and OBO ontologies. To provide deeper querying and more flexibility, the knowledge base is also accessible through a SPARQL endpoint. The iKUP application is online at http://www.kupkb.org and the SPARQL endpoint is at http://www.e-lico.eu/kupkb/sparql. Many of the datasets and ontologies are sourced from public repositories, however, a specific KUP Ontology (KUPO) was developed and is available from http://www.e-lico.eu/kupo. The source code for the KUPKB website is open source and available from http://code.google.com/p/kupkb-dev. Both the KUPO and the KUPKB were built by engaging our non-Semantic Web savvy users in the building process using tools such as Populous [3] and RightField. Using such tools enables iKUP's users to engage with the building process, delivering the annotated data for the KUPKB and extensions to the KUPO without having to have knowledge of OWL, RDF, SPARQL and related tools. We have taken this approach in an attempt to make the KUPKB sustainable and scalable.

The KUPKB was designed to fulfil specific requirements from members of the KUP community. These requirements include the integration of experimental datasets with existing biological databases about genes, proteins, miRNAs, metabolites and diseases. This type of data integration provides a constant challenge for bioinformaticians. One goal of the KUPKB was to show how many of these challenges can be overcome when data are aligned to a common representation and reuse is maximised. To do this we have reused portions of Bio2RDF and made the KUPO as an application ontology, reusing and extending many OBO ontologies [6].

The KUPKB utilises the state of the art in semantic web technology wherever possible. At its core is the RDF triple store that stores the data. After an evaluation we decided to use a Sesame framework [2] backed with a BigOWLIM storage and inference layer [5]. The iKUP web application is built using the Google Web Toolkit (GWT). Whilst the primary search on the iKUP website is powered by SPARQL, we also exploit the ontologies using the Java based OWL

API to classify results and handle the faceted browsing. We load all the ontologies into the OWL API and use the OWL-DL reasoner HermiT for classification. This provides us with the necessary inference to generate the hierarchy and drive the dynamic filtering of results on the web site. We can develop the whole application using a single programming language as Java interfaces are provided for Sesame, the OWL API and GWT. The importance of having such APIs in place can not be underestimated. The iKUP browser is a demonstration that semantic technologies are sufficiently matured so that they can deliver competitive data integration solutions.

The KUPKB and iKUP show ontologies coming of age by fulfilling some of their promise. The KUPKB has used ontologies to provide a common semantic framework for a broad range of previously semantically heterogeneous data. The use of Semantic Web technologies provides the means to integrate and query these data. The key to the coming of age is the iKUP user interface; without a simple means to access these integrated data, our biologist users would not and could not use the KUPKB; ontologies come of age when they deliver meaningful use to their intended users. This has now happened with the KUPKB, with biologists testing hypotheses generated *via* the iKUP in laboratories. Though the KUPKB is relatively small, it does what semantic technologies are supposed to do and show what is possible with biology's rich resource of data once issues of heterogeneity are taken away and the means of delivery to its users is taken into account.

**Acknowledgments:**This work is funded by the e-LICO project—EU/FP7/ICT-2007.4.4.

## References

1. Franois Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706 – 716, 2008. Semantic Mashup of Biomedical Data.
2. Jeen Broekstra, Arjohn Kampman, and Frank van Harmelen. Sesame: A generic architecture for storing and querying rdf and rdf schema. In Ian Horrocks and James Hendler, editors, *The Semantic Web - ISWC 2002*, volume 2342 of *Lecture Notes in Computer Science*, pages 54–68. Springer Berlin / Heidelberg, 2002.
3. Simon Jupp, Matthew Horridge, Luigi Iannone, Julie Klein, Stuart Owen, Joost Schanstra, Robert Stevens, and Katy Wolstencroft. Populous: A tool for populating ontology templates. *CoRR*, abs/1012.1745, 2010.
4. Simon Jupp, Julie Klein, Joost Schanstra, and Robert Stevens. Developing a kidney and urinary pathway knowledge base. *Journal of Biomedical Semantics*, 2(Suppl 2):S7, 2011.
5. Atanas Kiryakov, Damyan Ognyanov, and Dimitar Manov. Owlim - a pragmatic semantic repository for owl. In Mike Dean, Yuanbo Guo, Woochun Jun, Roland Kaschek, Shonali Krishnaswamy, Zhengxiang Pan, and Quan Z. Sheng, editors, *WISE Workshops*, volume 3807 of *Lecture Notes in Computer Science*, pages 182–192. Springer, 2005.

6. Smith, Barry, Ashburner, Michael, Rosse, Cornelius, Bard, Jonathan, Bug, William, Ceusters, Werner, Goldberg, Louis J., Eilbeck, Karen, Ireland, Amelia, Mungall, Christopher J., Leontis, Neocles, Rocca-Serra, Philippe, Ruttenberg, Alan, Sansone, Susanna-Assunta, Scheuermann, Richard H., Shah, Nigam, Whetzel, Patricia L., and Lewis, Suzanna. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, November 2007.
7. Katy Wolstencroft, Stuart Owen, Matthew Horridge, Olga Krebs, Wolfgang Mueller, Jacky L. Snoep, Franco du Preez, and Carole Goble. Rightfield: embedding ontology annotation in spreadsheets. *Bioinformatics*, 27(14):2021–2022, 2011.