33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Using Semantic Web Technologies to Manage Complexity and Change in Biomedical Data

Robert Stevens, Simon Jupp, Julie Klein and Joost Schanstra

Abstract— **Data in biomedicine are characterised by their complexity, volatility and heterogeneity. It is these characteristics, rather than size of the data, that make managing these data an issue for their analysis. Any significant data analysis task requires gathering data from many places, organising the relationships between the data's entities and overcoming the issues of recognising the nature of each entity such that this organisation can take place. It is the inter-relationship of these data and the semantic confusion inherent in the data that make the data complex. On top of this we have volatility in the domain's data, knowledge and experimental techniques that make the processing of data from the domain a distinct challenge, even before those data are organised. In this article we describe these challenges with respect to a project that is using data mining techniques to analyse data from the kidney and urinary pathway (KUP) domain. We are using Semantic Web technologies to manage the complexity and change in our data and we report on our experiences in this project.**

I. Introduction

In this paper we describe our approach and experiences in processing the complex data associated with the kidney and urinary pathway (KUP) [1], [7]. This domain's data are typical of those found in biomedicine and the types of analysis needed are similar to those in other sub-domains of biology and medicine. We have used Semantic Web technologies to manage the complexity of these data. We have developed the KUP knowledgebase (KUPKB) [5] as a resource description framework (RDF) [6] store that uses a KUP ontology (KUPO) as a flexible schema to capture the breadth of the domain's knowledge and to use it as a basis for data mining experiments.

Several aspects of working to understand an organism's biology have come together to make the processing of biology's data a science in its own right. Biology is itself highly complex; an organism is made of many entities that form complex relationships with each other to create the biological system in question. In the data we have about biological systems, each of these entity's relationships with other entities are regulated in complex patterns, with many states being contingent on other events. Representing and managing this complexity when processing biological data is hard [2], [3].

On top of this aspect, we also have the practice of the discipline itself that adds further complexity. Whilst some large projects produce large amounts of data, much data are still produced from an individual laboratory's high-throughput experiments. The experiments and the entities they produce can be described in highly heterogeneous ways that make organising and comparing data difficult. Common identifiers for the same entity are rare across the numerous databases in biology. The move towards ontologies for common vocabularies and minimal information standards helps [4], [9], but compliance is not great, except in some of the larger public resources [4].

Added to this we have change; molecular biology as a science changes rapidly. What was a 'fact' can change on an almost daily basis. Added to this are the rapid changes in experimental technique and variations upon experimental themes. In biology we now have experimental techniques that will let biologists explore the genome, the expression of the genes it contains, the proteins within a cell, the metabolism of a cell, and so on—all at a system rather than individual entity level [5]. All these experimental techniques produce new knowledge about a domain, but knowledge that has to be interpreted in the context of the factors used in the experiment and the broader biological context in which the experiment was performed. All these factors need to be taken into account in managing a domain's knowledge to make it suitable for data mining.

Over the past decade and a half, biology has worked hard to create a range of ontologies to overcome barriers of semantic heterogeneity. Most notable of these is the Gene Ontology (GO) [10] that gives a common vocabulary for describing the major functional attributes of gene products. GO is now used by over 40 data resources and has over 24 000 concepts. The Open biomedical Ontologies (OBO) is an umbrella organisation for GO and many other ontologies that cover biological phenomena from genotype to phenotype [8], and to investigations and clinical trials. Data are now being described with these ontologies, so biology is in principle now rich with coherently semantically described data that can be inter-linked to enable complex descriptions of biological phenomena. Now these data are described, we should be able to exploit those consistently semantically described data to explore those data more deeply than ever before.

II. The KUP Knowledgebase

In respect to its data and its questions, the KUP domain is a microcosm of the field of biology. KUP biologists wish to understand the kidney and urinary pathway, both its biology and the diseases in which the KUP goes wrong. To do so, KUP biologists study the genes, proteins, metabolism and all their

This work was funded by the e-LICO project—EU/FP7/ICT-2007.

R. Stevens and S. Jupp are with School of Computer Science, University of manchester, Oxford Road, Manchester, United Kingdom

Julie Klein and Joost Schanstra are with Institut National de la Santé et de la Recherche Médicale, Toulouse, France

regulations in the highly compartmentalised organ that is the kidney and related urinary tract. Data on the KUP domain resides in many public databases and in many individual data files about experiments, typically spreadsheets available as supplementary information for published articles. All these data need to be integrated to give appropriate contextual knowledge about each entity for data mining to take place.

We have developed the KUP ontology (KUPO) in the Web Ontology Language (OWL) to organise these data and populated it with genes, proteins, metabolites, and experiments, covering the transcription and protein complements of cells and urine across the KUP domain [5]. The KUPO uses ontologies for the anatomy of the kidney, cells, gene product functionality, disease, metabolism and investigations to cover the KUP domain. We have used proteins described using GO and added some 163 proteomic and transcriptomic experiments (at the time of writing) to the KUPKB, using the KUPO as a schema for describing these resources. The KUPKB forms an resource description framework (RDF) graph of all these data, integrated by the common naming scheme provided by the KUPO. In RDF terms, the KUPKB is not large (only some 20 million triples), but is complex, relating many kidney related entities together in complex ways.

RDF, such as the KUPKB, naturally forms a graph data structure that can be a target for analysis by graph-based learning algorithms. However, many data mining preprocessing and analysis algorithms expect a feature-vector based representation. By generating feature-vector based representation from RDF data we can explore a wider range of machine learning techniques over our data. The simple approach is to generate a data-table where a node in the graph becomes the identifier for a row entry, and the edges represent the attributes. Complications arise when a node in the RDF graph has multiple related objects along a given edge. This is a common case in RDF graphs; for example, a single gene annotated with multiple molecular functions from the GO. There are multiple approaches to deal with set-valued features, that involve some preprocessing in order to make the data suitable for data mining. In the KUPKB we can use a SPARQL query to generate a table of all the genes along with their GO annotations (See Figure 1). When a term has multiple related values along a single predicate we split these into individual attributes and create a true-false table for a gene against its given GO annotation. From this generated table we can begin to explore a wider range of data mining algorithms, such as building an association matrix. We have now started using the KUPKB to pull out data into such tables suitable for data mining experiments.

III. Experiences in building and Using the KUPKB

Our creation of the KUPKB has in essence worked for our needs. It has, however, not entirely been ‘plain sailing’. Some observations we have made on our experiences are:

- Availability of ontologies that cover the KUP domain was good, but when working with combinations of OBO and OWL, there are different mappings depending on the converter use. For example, we found three different URIs for the OBO relationship ‘part of’ depending on how and where the OBO to OWL translation was made. This type of deviation can make it harder to integrate across resources. For this reason, we have had to normalise existing representations to conform to the way we have chosen to model. The latest release of the OBO’s OBO to OWL converter is expected to address this problem by providing a reference mapping between the two languages.
- The KUP ontology covers many biological sub domains relating to KUP and re-uses vocabulary from existing ontologies wherever possible. There are many benefits to building an ontology using such a modular approach; however, current ontology editing environments, such as Protégé, do not provide an adequate level of support for managing the collaborative construction and maintenance of such an ontology.
  Problems also exist when it comes to publishing such an ontology. The KUPO specific axioms make little sense when viewed without considering the imports closure of KUPO and classification by an OWL reasoner. Conversely, when viewing KUPO with the imports closure and post reasoning the KUPO specific region forms only a small fraction of imported ontologies, and is thus not immediately obvious to the viewer. We have been exploring techniques in ontology modularisation to try and extract the specific KUPO portion of the ontology from the imports closure, but to date, whilst the extracted modules are often complete from a logical point of view, the output is still not desirable from a visualisation point of view.
- The KUPKB makes extensive use of external RDF datasets. To date no scalable approaches exist to support federated queries across multiple SPARQL endpoints. For now the KUPKB acts as a data warehouse where external linked datasets are manually imported into the KUPKB. This approach provides good scalability for our queries, but raises issues of concurrency with the imported data.
- The adoption of linked data for many life science databases is slow. There is still no authority on which URIs should be used for entities from many of the databases. This means that some published RDF datasets are not properly linked, even though they describe the same entities. For databases that have no RDF content, we are forced to generate RDF representations of the data that are not authoritative and likely to be incompatible for integration in the future.
- Much of the relevant proteomic and genomic data for KUP is only found in supplementary material of published research paper. This data comes in a variety of formats and is often difficult to extract, for example from PDFs, before we can convert into our RDF representation.
- We have to make many compromises with respect to the

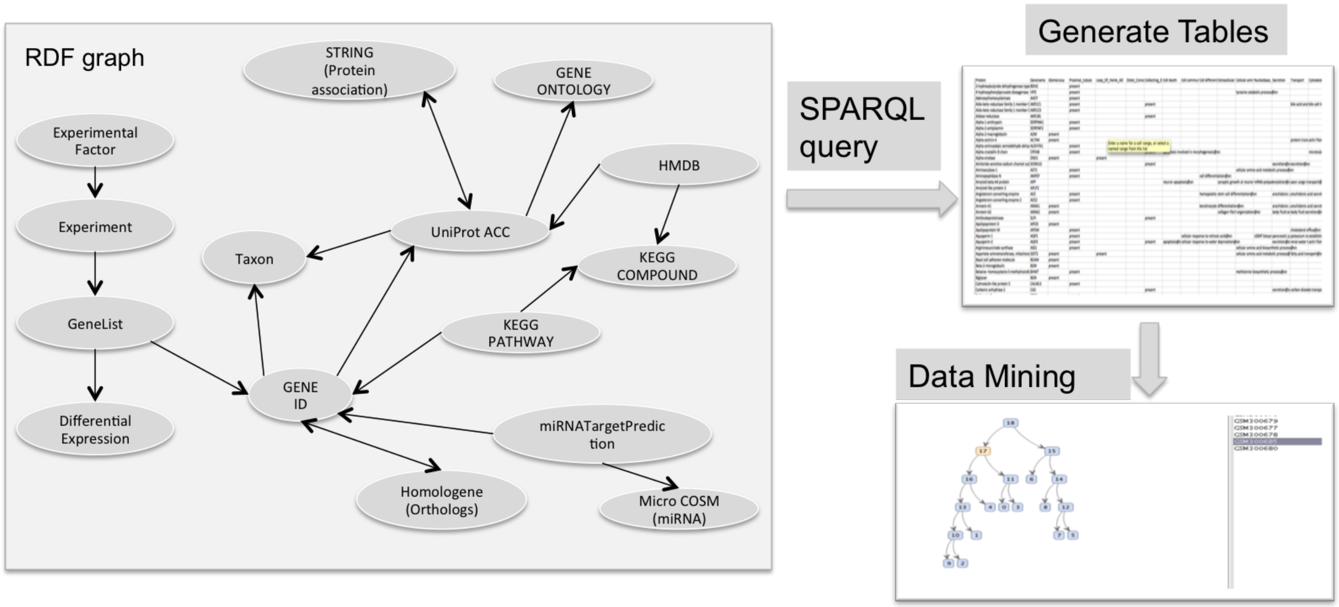

Fig. 1. The KUPKB RDF graph provides a flexible data model for querying. The queries generate data tables that feed into data mining workflows for further analysis

way we model the biological data. Our ideal solution would be to use all of OWL's features and expressivity to capture the data and use automated reasoning to make inferences and query the data. With the current size of the KUPKB (30,000 classes, and millions of individuals) we, however, face scalability problems. As a result we have used simpler modelling to ease the writing of SPARQL queries. We still use automated reasoners to pre-compute inferences over subsets of the data for computing class hierarchies, but much of the inferential power in querying is lost inside the RDF triple store.

IV. Discussion

The main issue with data mining biology's data is gathering and organising the data to be mined. This is not to diminish the complexities of the data mining itself, but the nature of bioinformatics data is hard. It is not necessarily a massively technical problem; there are many possible solutions, but the main issues are sociological. As a community, development of widely accepted standards that are responsive to change are needed. We know (by and large) how to do this from a technical point of view, but the community needs to change to adopt and co-operate to produce data in a certain way. As computer scientists, we need to make it easier to follow a (*de facto*) standard, rather than 'making up' one's own data standard.

We have made the KUPKB following various 'de facto' standards in bioinformatics, as well as having made some of our own deviviations—usually due to the incompleteness or changing nature of current efforts; as these develop the KUPKB will move to meet them. The KUPKB is, however, in a state that it is useful for both our human and computer users. Data will continue to be added to the KUPKB, especially more experiments and more types of experimenmt; this should involve extensions to the ontology that forms the KUPKB's schema.

We have already started using the KUPKB to gather data for data mining analyses. At present these are straightforward frequency counts and correlations. Even these simple analyses do, however, show us some interesting results (see, for example, the myExperiment pack at `http://www.myexperiment.org/packs/184.html`). Our main area of development, however, lies not in the KUPKB, but in exploiting this background data within the current table based data mining operations. As already mentioned, data mining operators tend to use ** aranged in tables. this can be done, but it is rather lumsy. We would like to be able to have data mining tools that exploited these graph based forms of background knowledge in their native form. At the simplest level, the graph of background knowledge can be exploited to abstract over the entities being mined. The other relationships in the graph can also provide much of an entity's context—processes in common, participating entities; stages and lifecycle parts in which these events happen; common forms of regulation. All these background concepts and relationships are in the KUPKB, but as yet are under-exploited in our data mining, but not by our human users. Future efforts wil be directed in this area.

The bioinformatics community is moving in the right direction. Metadata standards are being developed. Uptake is, however, patchy. The KUPKB is a start of an example of the kind of data organisation that can be made by exploiting Semantic Web languages and the semantically compliant data that are now available. There remains, however, much to do.

References

[1] Danielle Chabardés-Garonne, Arnaud Méjean, Jean-Christophe Aude, Lydie Cheval, Antonio Di Stefano, Marie-Claude Gaillard, Martine Imbert-Teboul, Monika Wittner, Chanth Balian, Véronique Anthouard, Catherine Robert, Bátrice Sǵurens, Patrick Wincker, Jean Weissenbach, Alain Doucet, and Jean-Marc Elalouf. A panoramic view of gene expression in the human kidney. *Proceedings of the National Academy of Sciences of the United States of America*, 100(23):13710–13715, 2003.

[2] S. B. Davidson, C. Overton, and P. Buneman. Challenges in integrating biological data sources. *Journal of Computational Biology*, 2:557–572, 1995.

[3] Paul Fisher, Cornelia Hedeler, Katherine Wolstencroft, Helen Hulme, Harry Noyes amd Stephen Kemp, Robert Stevens, and Andrew Brass. A Systematic Strategy for Large-Scale Analysis of Genotype-Phenotype Correlations: Identification of candidate genes involved in African Trypanosomiasis. *Nucleic Acids Research*, 2007.

[4] Carole Goble and Robert Stevens. State of the nation in data integration for bioinformatics. *Journal of Biomedical Informatics*, 41(5):687 – 693, 2008. Semantic Mashup of Biomedical Data.

[5] 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.

[6] Frank Manola and Eric Miller, editors. *RDF Primer*. W3C Recommendation. World Wide Web Consortium, February 2004.

[7] A Meguid El Nahas and Aminu K Bello. Chronic kidney disease: the global challenge. *The Lancet*, 365(9456):331 – 340, 2005.

[8] 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.

[9] C.F. Taylor, D. Field, S.A. Sansone, J. Aerts, R. Apweiler, M. Ashburner, C.A. Ball, P.A. Binz, M. Bogue, T. Booth, Alvis Brazma, Ryan R Brinkman, Adam Michael Clark, Eric W. Deutsch, Oliver Fiehn, Jennifer Fostel, Peter Ghazal, Frank Gibson, Tanya Gray, Graeme Grimes, John M. Hancock, Nigel W. Hardy, Henning Hermjakoband Randall K. Julian Jr, Matthew Kane, Carsten Kettner, Christopher Kinsinger, Eugene Kolker, Martin Kuiper, Nicolas Le Novre, Jim Leebens-Mack, Suzanna E. Lewis, Phillip Lord, Ann-Marie Mallon, Nishanth Marthandan, Hiroshi Masuya, Ruth McNally, Alexander Mehrle, Norman Morrison, Sandra Orchard, John Quackenbush, James M Reecy, Donald G. Robertson, Philippe Rocca-Serra, Henry Rodriguez, Heiko Rosenfelder, Javier Santoyo-Lopez, Richard H. Scheuermann, Daniel Schober, Barry Smith, Jason Snape, Christian J. Stoeckert Jr, Keith Tipton, Peter Sterk, Andreas Untergasser, Jo Vandesompele, and Stefan Wiemann. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature biotechnology*, 26:889–896, 2008.

[10] The Gene Ontology Consortium. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, 25:25–29, 2000.