Developing a kidney and urinary pathway knowledge base

Simon Jupp, Julie Klein, Joost Schanstra, Robert Stevens

Jupp et al. Journal of Biomedical Semantics 2011, 2(Suppl 2):S7
http://www.jbiomedsem.com/content/2/S2/S7

JOURNAL OF BIOMEDICAL SEMANTICS

PROCEEDINGS Open Access

Developing a kidney and urinary pathway knowledge base

Simon Jupp[1*], Julie Klein[2,3], Joost Schanstra[2,3], Robert Stevens[1*]

From Bio-Ontologies 2010: Semantic Applications in Life Sciences
Boston, MA, USA. 9-10 July 2010

* Correspondence: simon.jupp@manchester.ac.uk; robert.stevens@manchester.ac.uk
[1]School of Computer Science, University of Manchester, UK

Abstract

Background: Chronic renal disease is a global health problem. The identification of suitable biomarkers could facilitate early detection and diagnosis and allow better understanding of the underlying pathology. One of the challenges in meeting this goal is the necessary integration of experimental results from multiple biological levels for further analysis by data mining. Data integration in the life science is still a struggle, and many groups are looking to the benefits promised by the Semantic Web for data integration.

Results: We present a Semantic Web approach to developing a knowledge base that integrates data from high-throughput experiments on kidney and urine. A specialised KUP ontology is used to tie the various layers together, whilst background knowledge from external databases is incorporated by conversion into RDF. Using SPARQL as a query mechanism, we are able to query for proteins expressed in urine and place these back into the context of genes expressed in regions of the kidney.

Conclusions: The KUPKB gives KUP biologists the means to ask queries across many resources in order to aggregate knowledge that is necessary for answering biological questions. The Semantic Web technologies we use, together with the background knowledge from the domain's ontologies, allows both rapid conversion and integration of this knowledge base. The KUPKB is still relatively small, but questions remain about scalability, maintenance and availability of the knowledge itself.

Availability: The KUPKB may be accessed via http://www.e-lico.eu/kupkb.

Introduction

The early detection and better understanding of (chronic) renal disease is important as it will reach pandemic proportions over the next few decades [1]. The biologist's goal in renal disease is to understand the pathological processes and identify disease biomarkers. This requires the analyses of experimental data from multiple biological levels (e.g. genes, proteins and metabolites). These data need to be integrated with existing knowledge from databases and the scientific literature to connect the different levels. In addition, the kidney field is peculiar for at least two reasons:

1. the kidney is highly cellular and compartmentalised and each compartment is involved in many different functions and,

2. most of the large scale data comes from analysis of urine, that needs to be put into the 'kidney' context.

All together this makes the analysis of data, for which integration of data is a pre-requisite, problematic. This paper presents a case-study for developing a knowledge base around a focused domain in the life sciences, namely the kidney and urinary pathway (KUP). The KUP Knowledge Base (KUPKB) is being developed as part of the e-LICO project [2]. e-LICO is developing a data mining platform that supports the semi-automated construction of data mining workflows for data intensive sciences [3]. The e-LICO platform is to be demonstrated with a system biology use case that uses real data encountered in the KUP domain. The data spans multiple -omic levels and is collected from different tissues and from different species. For example, most of the human -omics data originates from urine [4] and needs to be related back to the kidney and its parts. In contrast, multilevel -omics data from animal models is more regularly available. e-LICO aims to develop tools that will mine these large scale disparate experimental findings, link those to existing data and build new predictive models for renal disease.

The KUPKB is built using a Semantic Web approach in order to assess the benefits and feasibility of creating such a resource with this technology. The methodology section guides the reader through the creation of a Kidney and Urinary Pathway Ontology (KUPO), that provides a specialised application ontology for the KUP domain. The KUPO provides the schema for the data held in the KUPKB. Within this methodology we explore the requirements for tools that help engage the biologists in the design and construction of such an ontology. The results section describes the KUPKB with examples of the kinds of queries that can be asked across multiple biological levels. We conclude by discussing the merits and limitations of our approach.

Background

Data integration in the life science is an ongoing challenge in Bioinformatics; problems arise because standards for data formats, identifiers, common vocabularies and agreed semantics between databases are lacking [5,6]. Data in the life sciences are complex and volatile that, when taken with the issues outlined, makes the necessary integration of life sciences data hard work. Another factor is the numerous data resources published by independent groups that leads to an expansion of the heterogeneities that are rife in life science data [7].

Developing new resources that integrate existing data typically involves centralising the external data within new bespoke schemas. This 'warehousing' approach is common in the life sciences and over time leads to an increasing number of resources, each with their own schema [7]. The situation with respect to accessing these data is, however, improving with data providers often offering programmatic access to the data via Web Services or database exports [8,9]. This access affords easier integration opportunities, despite the semantic heterogeneities and the problem of identity of entities within life science's data. The *'identity crisis'* [10] is being addressed through efforts such as shared names [11] and services such as BridgeDB [12], but wide spread compliance has yet to be realised. The adoption of ontologies for the annotation of data is providing new possibilities for data integration that go beyond using primary database entry identifiers alone.

The problem is exacerbated in the life sciences due to the nature of the data being captured. Biological data are complex, heavily inter-related and also often irregular or incomplete [6,13-15]. Relational databases are good in situations where the data are regular and complete, and thus are not always suitable for life science data [16-18]. Ontologies offer a potential solution to this problem as they have been demonstrated to be good at modelling things that are irregular and incomplete [19,20]. Ontologies are designed to be extensible and can be used to build a conceptualisation of a domain. An ontology language, such as the Web Ontology Language (OWL) [21], provides a precise semantics for the language that can be used to check for consistency in data, along with querying and inference over data. Ontologies have become popular in the life sciences [22] for the annotation of data and offer novel ways for the analysis and integration of biological data. Despite the uptake of ontologies for annotating data, these annotations are often held in more traditional database systems that lack the necessary support to fully exploit the benefits an ontology brings.

The Semantic Web encompasses many ideas and technologies, but at its heart is the creation of a connected Web of semantically described data. As opposed to the existing Web of connected documents, the Semantic Web provides a framework to publish statements about entities in those documents [23]. At the core of the Semantic Web is the Resource Description Framework (RDF) [24] that uses Uniform Resource Identifiers (URI) to identify objects on the Web. The RDF data model provides a mechanism to make statements about these objects in the form of Subject, Predicate and Objects; these statements are commonly referred to as triples. This simple triple model can be used to create large connected graphs of data and expose them to application over the Web. The use of URIs and RDF to expose data on the Web enables the syntactic integration of data held in biological databases; this model is easily extensible by the addition of new triples and is not constrained to a particular schema. This light, syntactic reconciliation means that publishing data in RDF is easy. This ease is at least in part due to the lack of a schema. Using these data, however, can be hard due to the lack of a schema; this means the heterogeneity in the data still exist and, while a common syntactic form allows queries across different data represented as RDF, the query formulator still needs to reconcile the naming and conceptualisation of those data in order to formulate that query.

RDF alone provides little in the way of semantics for what these objects are and what the relationships mean. It does, however, offer a means of layering semantic information over its simple data model and data published according to that model. At their heart, ontologies provide a simple service by defining the entities as they appear in a domain's information. 'Knowing what there is' and adopting a common means of naming and inter-relating 'what there is' offers a means of providing a semantic layer over the syntactic integration of RDF. A common understanding of the types of entities and the types of relationships between them provides the means to query those data; that is, ontologies offer a schema-like mechanism for RDF data. The Web Ontology Language (OWL) provides us with an ontology language that can be expressed using RDF. OWL ontologies can add semantics to data captured in RDF, these semantics facilitate a common model for the data, inference and expressive queries over the data.

The work presented in this paper builds on previous efforts to expose life science data on the Semantic Web. Ruttenburg, Antezana and Cheung give a wider overview

of how Semantic Web technologies are being deployed in the life sciences [25-27]. Dhanapalan, Sahoo, Hugo and Antezana provide case studies from different domains for data integration on the Semantic Web [28-31].

Several public databases have been made available as RDF, the Bio2RDF project provides a repository of over forty biological database already available to download as RDF [32]. The W3C health care and life sciences working group (HCLS) provide guidelines and a case-study on Alzheimer's Disease that was built on Semantic Web technology [33].

The KUPKB sits within this tradition. The KUPKB is distinguished by its means of production; we use a series of existing ontologies with a light-weight mechanism for integrating these ontologies to give a backbone of background knowledge for the KUP domain in the form of the KUPO. This 'schema' is populated with workflows that bring in the various pre-existing KUP resources.

Methodology

An initial set of experiments were chosen by the KUP scientists for inclusion in the KUPKB. These experiments span over different biological levels (genes or proteins), different techniques, different species (mouse or human), and different sample type (urine or kidney tissue):

1. The Higgins dataset [34] represents gene expression values from seven dissected compartments of the healthy adult human kidney, analysed by microarray technology.

2. The Chabardés-Garonne [35] dataset represents gene expression values from eight dissected compartments of the healthy adult human kidney, analysed by the Serial Analysis of Gene Expression (SAGE) method.

3. The EuReGene dataset [36] represents gene expression values from all the different renal structures of the healthy adult mouse kidney, analysed by the *in situ* hybridisation technique.

4. The Vlahou [37] and the Mann [38] datasets represents protein expression values from healthy adult human urine, analysed by different mass spectrometry techniques.

For each of these experiments the output is typically a list of genes or proteins of interest. The initial challenge is to identify proteins found in the urine and related them back to genes expressed in the kidney. This is made difficult as the data are coming from a range of experiments on both mice and humans. Background knowledge about genes and proteins is combined with experimental findings to generate new data for further analysis. The use of ontologies to annotate these data adds value, such as the ability to generalise over the observations. This combined data will provide input-data to a series of data-mining experiments within the e-LICO project.

We need the KUPKB to answer a series of queries about biological compounds in the urine and kidney. To achieve these we need basic information about the compounds under analysis. Information about genes, proteins and metabolites can be harvested from publicly available databases. These data must be linked to evidence coming from the experimental analysis and appropriately annotated to distinguish between the different experimental factors, such as the biological material or pathological state. To tie this data together we need ontologies that describe the relationships between the various datasets. These will include an ontology for the kidney and urinary pathway system, experimental analysis and biological databases. We use Semantic Web

technologies to build the KUPKB so we can exploit the ontologies for querying and inference. RDF gives us language to represent the data plus a means to publish it on the Web following Linked Data principles [39]. Ontologies provide both the schema for the KUPKB and a controlled vocabulary for data annotation. We conceptually divide the KUPKB into three overlapping sets of ontologies (Figure 1). The first set of ontologies provides a domain vocabulary for describing the kidney and urinary pathway. This Kidney and Urinary Pathway Ontology (KUPO) describes the cells of the kidney in terms of their function and their anatomical locations. The second set describe the experimental data, this includes descriptions of the experimental results along with meta-data about the experiment, such as the experimental factors under observation. The final set of ontologies describe the data obtained from various external biological databases.

The life sciences are now rich in ontologies to support our task [22,40]. We can take advantage of these efforts and use fragments of these ontologies to build the KUPKB. Re-using existing ontologies in the KUPKB offers many advantages from an integration point-of-view: by adopting standards and exploiting existing annotation efforts [41] we have a greater potential for future integration of the KUPKB with other similar applications.

Kidney and urinary pathway ontology development

The kidney enables the filtration of waste from the blood in the form of urine. Schematically, the kidney can be divided into four major compartments: 1) the glomerular compartment, involved in blood filtration, 2) the tubular compartment, involved in the fine tuning of urine composition, 3) the vascular compartment, involved in renal blood supply and 4) the interstitial compartment that surrounds the other structures. Each kidney compartment is formed from a wide variety of cell types, and the specificity of the compartments relies on these specialised cell functions. Depending on the aetiology, renal diseases may differentially affect the renal cells and the kidney

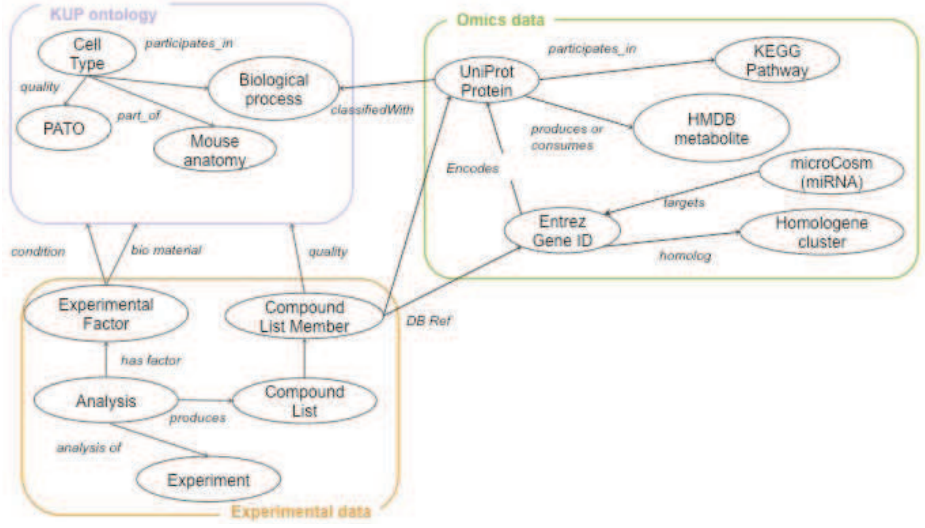

Figure 1 Schema overview. Overview of the KUP KB schema showing experimental data connected to background knowledge and annotated with the KUP ontology

compartments (e.g. diabetic nephropathy affects mainly the glomerular compartment whereas obstructive nephropathy affects mainly the tubular compartment). It is important to link back the disease processes to the anatomical alterations, as it will help not only to better understand the pathological mechanisms, but also to adapt therapeutic strategies. For these reasons, the first version of the KUPO imports ontologies to describe the anatomy, cell types and biological function associated with the cells of the kidney.

Reference ontologies for the components of the kidney and urinary pathways are readily available through resources such as OBO foundry [22] and BioPortal [40]. KUPO is a set of OWL classes that represent the cells of the kidney. The classes are described using logical definitions that use conceptualisations taken from external ontologies relevant to the KUP. This modular approach avoids repeating any previous development efforts, and focuses on extending and enriching pre-existing ontologies where necessary.

Two ontologies were initially considered to describe the anatomy of the kidney: the Foundational Model of Anatomy (FMA) [42] and the Mouse Adult Gross Anatomy Ontology (MAO) [43]. Domain experts inspecting the KUP portion of the FMA found that there was too much detail in some sections and not enough in others. In addition, too many ontological distinctions were made within this portion of the FMA and the consequent dispersal of information made it hard for our collaborating biologists to use. In time, we could have refined views of the FMA to do the job required, but we found that the MAO had all the detail for our needs. Furthermore, despite the *connecting tubule* being absent in mouse and present in humans, we still find this concept in the MAO. Therefore we concluded that the MAO can act as a substitute for the human anatomy, at least as far as the kidney is concerned.

This approach may be acceptable in the short term. There is, however, the issue that we effectively label human entities as being mouse entities. A solution would be a species neutral vertebrate anatomy, so any renal cell from any vertebrate species could be labeled as *renal cell*, and the species recorded elsewhere (see [44]. Efforts such as CARO (Common Anatomy Reference Ontology) [45] are attempting to provide cross species reference ontologies for anatomy. The Vertebrate Bridging Ontology (VBO) project [46] has recently started and we will look to use such efforts in the future.

The Cell Type Ontology (CTO) [47] was the obvious choice for cells, but was found to be lacking a large number of known renal cell types. A list of new cells was therefore generated and cells were described in terms of their anatomical location using the *part_of* relationships from the Relations Ontology (RO) [48]. By exploiting the rich partonomy of the MAO we could use the transitive characteristic of the RO *part_of* relationship to describe the renal cells using equivalent class axioms. The logical definition meant that the complete classification of renal cell types could be computed using an OWL reasoner. The renal cells were further described in terms of the biological processes from the Gene Ontology (GO) [49] in which they participate. These new cell types are to be submitted for inclusion into the CTO.

KUPO development

Populating the KUPO requires detailed domain knowledge about the anatomy and cells of the kidney. We wanted to explore methodologies that engaged the domain expert in

the ontology building process. Modern ontology editors, such as Protégé [50], are powerful, but necessitate a relatively steep learning curve that can be dissuasive for domain-experts. We attempted to construct the ontology using simple templates that could be populated by domain experts to gather knowledge about the kidney. Populated templates were then converted to OWL to produce the KUP ontology.

The templates were generated using a simple spreadsheet, a technique that has recently become popular in other ontology efforts [51,52]. A KUPO day was also held as part of a meeting between domain experts from the EuroKUP consortium [53] to develop and review the knowledge captured in the spreadsheet. These spreadsheets were validated and transformed into OWL using the Populous tool [54]. Figure 2 shows the KUPO template populated in Populous. A detailed description of Populous and the KUPO development methodology is available in [54].

Background knowledge

Background knowledge in the KUPKB is composed of various external databases represented in RDF. Converting existing data into RDF triples is relatively straight forward. Our methodology closely mimics the development of the HCLS knowledge base [55]. Additionally the Bio2RDF project [32] provides a repository of public databases that can be downloaded in RDF. These two efforts provided some of the core datasets for the background biological knowledge represented in the KUPKB.

The initial KUP experiments are concerned with genes and proteins expressed in kidney and urine samples, while future datasets will be generated from metabolomic and microRNA studies. We selected the Entrez Gene database [56] for gene annotations and UniProt Knowledgebase [57] for proteins, including the Uniprot Gene Ontology annotations [41]. KEGG [58] provides the data for biochemical pathways whilst the microCosm database [59] provides microRNA target prediction sites. Given that the

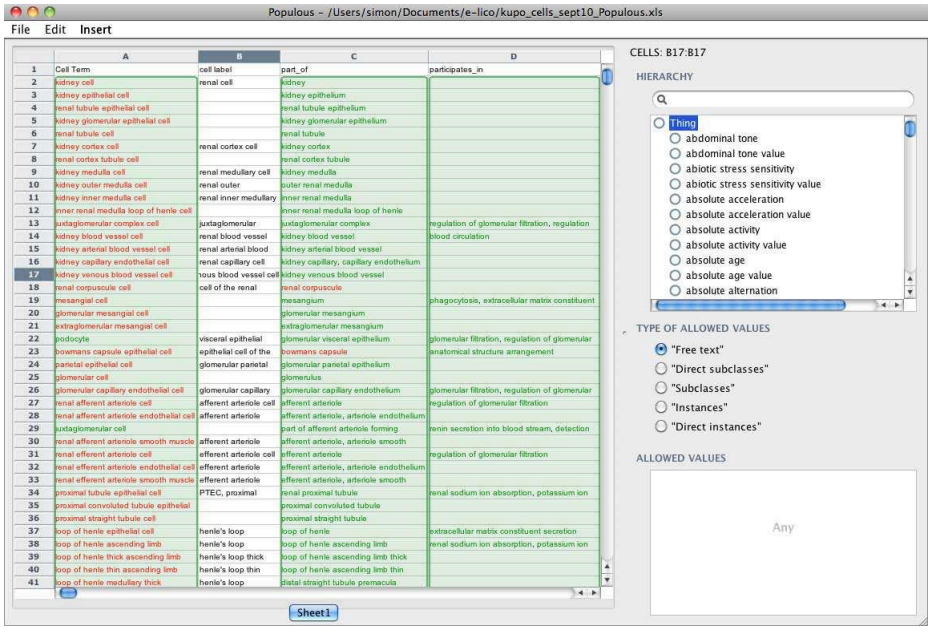

Figure 2 Populous. Screenshot of Populous showing template population of the KUP ontology

experiments are being conducted on multiple species, we also required data about orthologous genes that can be obtained from the Homologene database [60].

We wished to capture the relationships between the various entries in these databases to form a large connected graph of data. Figure 1 gives an outline of how our background knowledge relates to the KUPO and the experiments. Where possible we used the Bio2RDF ontology [61] to provide a simple schema for the different databases.

RDF representations of Entrez Gene, Uniprot, Homologene and KEGG can be obtained from Bio2RDF. Uniprot KB also provide periodic releases of their database in RDF. Whilst it was possible take the complete databases and place them directly into the KUPKB; it was neither necessary or desirable to replicate these databases in their entirety. Instead we reuse URIs from Bio2RDF and Uniprot KB to reference external entities in the KUPKB. By re-using URIs from external resources we can expose the KUPKB as *Linked Data* and harvest additional data if and when we need it. Resources in the KUPKB have URIs that resolves to documents on the Web; these documents provide a description of that resource in RDF and link directly to other documents that describe the external resources using RDF. This linking connects the KUPKB into the growing web of linked life science data and highlights one of the major advantages of a Semantic Web approach.

KUPKB experiments ontology

The KUPKB is required to capture findings from biological experiments such as lists of interesting genes or proteins expressed under certain conditions. Ontologies such as OBI [62] and EFO [63] provide concepts for the annotation of design, protocols, instrumentation, materials, experimental factors and data associated with a biomedical investigation. Despite the extensive coverage of these ontologies, neither provided a clear way to represent information in gene and protein lists. It is particularly difficult to attach the appropriate meta-data to these lists in a way that could accommodate a wide variety of use-cases.

It is beyond the scope of this work to provide a complete modelling solution to this problem, and work is underway through projects like OBI and the BioRDF task force of the HCLS working group to create standard modelling patterns for this kind of data. We opted to build our own ontology to model this data that was simply driven by requirements of the queries we needed to ask. Though, with a view to future integration with external ontologies, we followed patterns and reused concepts from existing ontologies, including OBI, EFO and PATO [64].

Each experiment in the KUPKB is modelled as an instance of a particular *'experimental assay type'*. Each *'experimental assay type'* can have multiple *'experimental analysis'* associated with it. Each *'experimental analysis'* is annotated with an *'experimental factor'* and produces some *'data'*.

For example, the Chabardés-Garonne dataset [35] represents gene expression values from eight dissected compartments of the healthy adult human kidney. This experiment is described as a type of *'Expression Profiling by SAGE'*. We create instances of *'experiment analysis'* to represent each of the factors under investigation. In this case, the analysis represents gene expression sets for dissected compartments of the kidney, as published by the authors. *'Experimental factors'* can have a related

sample (e.g. glomerulus), a related condition (e.g. *'Adult Human'*), and an associated role in the experiment (e.g. *'control' or 'analyte'*). The resulting *'gene list'* for a particular analysis is related to multiple *"gene list members'*. Each member in a gene lists can be annotated with additional meta-data such as an external database reference (e.g. Entrez Gene id for genes). We have adopted terminology from PATO to describe attributes for the *'gene list members'* such as *'postitive regulation'* or *'present'*. Figure 3 is an OWL representation of a single gene in the gene list from the Garonne dataset. We make no claims about our choice of modelling other than it is sufficient for the kinds of queries we need to ask over the current set of proposed experiments in the KUPKB.

KUPKB architecture

There are now many options for developers wishing to deploy RDF and OWL data, the most common approach to date is the use of a triple store. Triple stores provide a framework for storing and querying RDF data, along with support for varying degrees of inferencing. One of the major factors in selecting a triple store is scalability. In recent years triple stores have improve to accommodate large amounts of RDF data (billions of triples). See here [65,66] for a review of some popular triple stores. We opted to deploy the KUPKB data using Sesame [67] backed with a storage and inference layer provided by BigOWLIM [68]. Sesame provides a powerful yet simple framework for managing and interfacing with the RDF data whilst BigOWLIM provides a fast and scalable implementation of the Sesame Storage and Inference Layer (SAIL). In order to expose the KUPKB data as Linked Data, we use the Pubby [69] service. A simple description of every resource in the KUPKB can be retrieved by its URI as either a human readable HTML document or plain RDF.

Results

The KUPKB can be queried over the Web via a SPARQL endpoint [70]. The KUPKB Web page provides some example SPARQL queries to generate interest and harvest requirements from the wider KUP community. The demo shows how one can query across multiple data sources using predicates and terminology from the available ontologies. At the time of writing, the total number of RDF triples in the KUPKB is 10,415,339, whilst the ontologies and experiments describe 24,557 classes and 10,539 individuals. To date no triple stores support the full expressivity of OWL-DL or OWL 2; in order to compensate for this we attempted to compute any inferred knowledge before loading the data into the knowledge base. We attempted to classify the complete knowledge base with three OWL reasoners; Pellet 2.1.2 [71], Fact++ 1.5.0 [72] and HerMiT 1.3.1 [73] on a 2 x 2.66GHz Dual Core Intel Xeon Mac Pro with 16GB or RAM and running OS X server 10.6.4. No reasoner was able to classify the complete knowledge base in our experiment, however, by excluding the large external databases we were able to classify the various ontologies along with the experimental data in reasonable time (4.30 min with Fact++). This classification step provided a level of consistency checking and also enabled the computation of missing subsumptions to form the KUPO class hierarchy.

The first set of demo queries answer simple questions from the KUPO. For example, 'Which biological processes occur in the kidney collecting duct?' is answered by getting

```
Individual: GSE694_Garrone_assay

  Types:
    'Expression Profiling by SAGE'

Individual: GSE694_Garrone_MA_0002628_analysis

  Types:
    SAGEAnalysis

  Facts:
    analysisOf GSE694_Garrone_assay,
    annotated_with GSE694_Garrone_MA_0002628_factor,
    produces GSE694_Garrone_MA_0002628_list

Individual: GSE694_Garrone_MA_0002628_factor

  Types:
    'experimental factor'

  Facts:
    has_role analyte,
    has_bio_condition 'Normal Adult Human',
    has_bio_material MA_0002628


Individual: GSE694_Garrone_MA_0002628_list

  Types:
    GeneList

  Facts:
    hasMember GSE694_Garrone_listmember_10263

Individual: GSE694_Garrone_listmember_10625

  Types:
    GeneListMember

  Facts:
    hasDatabaseRef geneid:10625,
    hasQuality present
```

Figure 3 Gene lists in OWL. Manchester OWL syntax for the asserted information representing a single analysis of the Garonne dataset

all the cells from the CTO and the KUPO that participate in a biological process from GO, and filtering these results on cells that are part of the collecting duct from MAO.

More interestingly, the KUPKB demonstrates how having access to experimental observations along-side existing domain knowledge could be useful for hypothesis generation and experimental evaluation. The query 'Which genes have evidence for upregulation the glomerulus of a normal adult kidney?' can now be answered by getting all genes annotated with *'upregulated'* from all experiments where the condition is *'normal adult human'* and the bio material is *'glomerulus'*. This kind of query demonstrates the added value KUPO brings. Not only do we get answers for genes that are expressed in the glomerulus, but by exploiting the transitive nature of the *part_of* relationship, we infer genes upregulated in experiments on any part of the glomerulus. We can extend this query by exploiting the KUPO to ask for all the cells that are in parts of the kidney expressing protein involved in some biological process or pathway.

Our final query demonstrates how we can begin to answer real questions from the KUP scientists. The initial goal was to place protein expressed in urine back into the context of the kidney. The Vlahou and Mann datasets [37,38] both provide us with sets of protein identified in urine from normal adult humans. We want to compare these results with evidence for gene expression in particular compartments of the kidney. Our example query gets all the proteins found in urine across both experiments. We need to find the intersection of proteins in this list with proteins expressed in both the Higgins [34] and Garonne [35] datasets. To do this we must first map the genes to their respective protein using background knowledge in the KUPKB. We can get the Uniprot identifiers for each gene expressed in the transcriptomics datasets. By combining these two sets of derived protein we find that 183 proteins have evidence for expression in seven separate compartments of the kidney. We can further enrich this list with their appropriate GO annotations to see if there are any patterns observed in this list. We can now extend this query using data from homologene to bring in the EuReGene dataset from mice to collect further evidence.

The flexibility offered by KUPKB provides the e-LICO project with a platform to plan data mining experiments over the KUP data. A recent development from e-LICO was the initiation of a KUP challenge; this is a challenge for the data mining community to learn models from datasets relating to Obstructive Nephropathy (ON) in children [74]. These datasets are of high dimensionality, but extremely small sample size. The datasets represent analysis from multiple biological levels including miRNA, mRNA, proteomics and metabolomics; the challenge is to build a prediction model from these datasets that uses background knowledge in the KUPKB to connect the different levels.

Discussion

We have presented an approach to developing a bespoke knowledge base for integrating biological data relating to the KUP. This KUPKB integrates experimental findings with background knowledge to provide input for data mining experiments on the e-LICO platform. The KUPKB differs from more traditional database approaches by its extensive use of Ontologies and Semantic Web technologies to provide the underlying data model. The KUPKB is an example of the levels of data integration that can be achieved once data can be reduced to a common language. Having access to multiple

connected data sources in a single RDF repository provides a powerful and flexible platform for data gathering and analysis. The additional semantics provided by the ontologies offer new and more flexible ways to query and explore the data. Despite the potential, there are, however, many outstanding issues that continue to stifle development in this area.

We found that taking an approach using RDF to gather our data in one form worked well. As well as taking advantage of existing resources in RDF, conversion of bespoke KUP experimental data was relatively straight forward using simple scripts. The main difficulty was reconciling identifiers across resources; for each experimental dataset we encountered, a considerable effort was required to map the various identification schemes between databases. Whilst this was to be expected for the biological databases, it was a surprise to see that little authority exists within the life sciences on the correct URIs for various ontologies and RDF datasets. For example, we encountered multiple URIs for common predicates such as the *part_of* relationship from the OBO relations ontology.

The KUPKB still has aspects of a warehouse; it gathers resources together in one place in one form. The KUP ontology is used as a schema for the data. Some of the dangers of warehouses are, however, avoided. The Semantic Web technologies are tolerant, indeed suited for, irregular and incomplete data. Production of data in the common form of RDF is distributed and much of the task of forming the KUPKB is the 'gathering'. The emerging ontological *lingua franca* makes mapping to a schema relatively straight-forward. An early requirement for the KUPKB was an ontology that accurately described the kidney and urinary system. Given the widespread availability of anatomical ontologies that describe the renal system, we decided to focus on the cells of the kidney, that previously had little coverage in existing ontologies. The challenge was to engage the domain experts in the development of such an ontology. We managed to achieve this by shielding the experts from the underlying ontology building, and instead gathered knowledge using a simple template approach. The templates were populated in spreadsheets, an application familiar to the domain scientists. This approach led to the description of over 180 renal and urinary cell types by domain experts with little or no ontology building experience.

In order to support the validation of these spreadsheets and their transformation into an ontology, we developed an application called Populous [54]. Populous enables us to separate the axiom pattern in the ontology from its population. The pattern itself was simple, but in a pattern with few axioms and 180 repetitions, we generated an ontology with several hundred asserted and inferred axioms. The Populous approach allows this generation to happen quickly and with utter consistency. If the pattern changes, but with the knowledge the same, then update is equally quick. Any change to the knowledge can be done in the table environment that validates input against constraints.

This first version of KUPO serves its purpose as an application ontology. Questions, however, still remain about our choice of modelling for this ontology. We used *participates_in* to relate a cell to the process in which it is involved, even though we know this is not necessarily true for all cells of a given type at all times. We could model that these cells have a disposition that is realised in the biological process, but this kind of distinction does not add any value to the kinds of queries we want to ask. We have deliberately moved away from a representation of the"truth" of the biology in

order to accommodate the queries our users wish to make. Problems of "truth" continue to plague us in the development of an ontology to describe our experiments.

Developing a common ontology for biomedical experiments and findings is difficult as the semantics of a biological finding often involve a complex set of interactions. If we take our gene lists in the KUPKB as an example; these represent some experimental finding that a biological entity is either present or absent under a certain condition. The validity of such a statement should take into account a host of factors including the type of experiment, which instruments were used, how the data was analysed, along with consideration for less controllable factors such as human error or bias. The development of appropriate ontologies that capture these various attributes are well under way, however, progress is inevitably slow due to the complex nature of the task. For the KUPKB we initially have few requirements for complex descriptions of the experiments. Datasets are being selected by a relatively small community, and much of this data is similar enough for cross-comparison. As the KUPKB expands to include more heterogeneous datasets, more accurate description of the data may well be required in order to ask more complex queries of the data. Developments in this area is ongoing and we will contribute experiences from this exercise to the appropriate ontology efforts. We believe that a relatively simple ontology, similar to our experiments ontology, could provide a model that is "just enough" to achieve more wide spread integration of this kind of data.

One of the attractive features of adopting ontologies and technologies like OWL is the ability to exploit the semantics of the language to check for consistency and draw inferences from the data. Where possible we exploited OWL semantics to drive inferences and reduce the number of manual assertions we needed to make in our ontology. This has many advantages from an ontology maintenance point of view and provides us with with advanced querying capabilities, such as asking for parts of the renal cortex and returning all the parts right down to the cells. The ability to traverse transitive relationships is one of the benefits an ontology language can bring; it enables us to ask general queries such as asking for all cells that participate in cytokine production, and by inference would also return cells that participate in specialisations of cytokine production, such as B cell cytokine production.

Despite the expressive power of OWL queries, we hit a limit when we try to populate these ontologies with instance data; current OWL reasoners do not easily scale to the volumes of data presented in the KUPKB, which only represents a very small dataset in comparison to other efforts such as the Alzeimers knowledge base. As a practical solution to this problem we use an RDF database, which can scale to many billions of triples, and exploit the limited, but adequate, inferencing capabilities currently on offer.

By moving to an RDF store we lose the ability to ask queries that use higher level OWL constructs. An RDF query language like SPARQL enables us to explore the structure of an OWL ontology, but queries soon become complex when working with class level descriptions. Our chosen RDF store BigOWLIM provides inferencing up to the level of RDFS along with a fragment of OWL that can be expressed using rules, such as transitivity and *same as*. We take the attitude that some answers (that may be incomplete) are better than no answers from a sound and complete query solution. The reduction in expressivity within the RDF domain means that the queries asked may not be as precise; to date we have not yet found this to cause significant

problems, though this may not be the case in the future. Work to increase the expressivity of SPARQL with respect to OWL semantics is underway through current proposals such as SPARQL 1.1 entailment regimes [75,76] and SPARQL-DL [77].

The technical limitations associated with querying a knowledge base like the KUPKB has an impact on how we build our ontologies and model our data. Most ontologies for the life science model biological phenomena at the class level. For example, the class *'cytokine production'* describes all instances of the process of cytokine production. If we treat a particular protein record from Uniprot as an instance, then to represent the relationship in OWL we would type this instance as a member of the class of things that participate in cytokine production. Representing OWL axioms such as this in RDF soon becomes verbose and impacts on the complexity of the query. In order to simplify the query we treat the classes as individuals and create a binary relationship between the two.

Future work

The KUPKB is to be extended with several datasets from experiments about specific types of renal disease. These include proteomic analysis of urine from children suffering obstructive nephropathy. The KUP ontology will be extended to incorporate the appropriate ontologies that describe the diseases under investigation. These data will serve as input to data mining experiments within the e-LICO project, with the ultimate goal of generating new predictive models for renal disease. Early experiments have shown that we can find useful and interesting correlations in the data pulled from the KUPKB.

A further issue to tackle is the update of resources. Experiments change; new data emerge; the ontologies describing these resources change. Keeping resources like the KUPKB up-to-date will be a struggle. Tools such as Populous will help with re-generating the ontology, but the mappings to the ontology all have to be managed, exposed and recorded. In addition, the provenance of the input from all the resources needs to be recorded to aid scrutiny.

At present we have not recorded much information about the experimental protocols. Currently, the experiments we include in the KUPKB have been chosen by our participating biologists. In future we will need to describe the experiments such that scientists can have access to a wide variety of experiments and make the choice themselves. This final point of scientists interacting with the KUPKB highlights the need for good interfaces for querying and browsing such knowledge bases.

Conclusion

We have presented an approach to developing a knowledge base to serve a community of scientists working on kidney and urinary pathway diseases. We have demonstrated how a knowledge base such as this can be rapidly developed using state-of-the-art SW tools. In the KUPKB we have taken advantage of the *de facto* integration described as an aim of ontologies such as GO [78], by using the ontologies themselves along with data annotated with those ontologies to provide an integrated resource of data about KUP for a community of biologists. The KUPKB can provide useful querying facilities to biologists and has been relatively easy to produce. There remains much to do, but we count the work so far as a success.

Acknowledgements
SJ built the workflows and tools to construct the KUPKB and KUPO. The design of KUPKB and KUPO was lead by SJ, JK and RS. JK and JS provided the biological use-cases and selected the initial set of data to be included in the KUPKB. SJ drafted the manuscript whilst all authors contributed and approved the final manuscript. The authors would like to thanks the EuroKUP members for participating in the KUPO day, in particular A. Vlahou and J. Frokiaer for reviewing the KUPO and their useful suggestions; Michael Wicks and Duncan Davidson for supplying the EuReGene dataset. This work is funded by the EU/FP7/ICT-2007.4.4 e-LICO project.
This article has been published as part of *Journal of Biomedical Semantics* Volume 2 Supplement 2, 2011: Proceedings of the Bio-Ontologies Special Interest Group Meeting 2010. The full contents of the supplement are available online at http://www.jbiomedsem.com/supplements/2/S2.

Author details
[1]School of Computer Science, University of Manchester, UK. [2]Institut National de la Santé et de la Recherche Médicale (INSERM), U858, Toulouse, France. [3]Université Toulouse III Paul-Sabatier, I2MR, IFR150, Toulouse, France.

Competing interests
The authors declare that they have no competing interests.

Published: 17 May 2011

References

1. Nahas AME, Bello AK: **Chronic kidney disease: the global challenge.** *Lancet* 2005, **365(9456)**:331-340.
2. e-LICO:[http://www.e-lico.eu].
3. Serban F, Kietz JU, Bernstein A: **An Overview of Intelligent Data Assistants for Data Analysis.** *Proc. of the 3rd Planning to Learn Workshop (WS9) at ECAI'10* 2010, 7-14.
4. Decramer S, de Peredo AG, Breuil B, Mischak H, Monsarrat B, Bascands JL, Schanstra JP: **Urine in Clinical Proteomics.** *Molecular and Cellular Proteomics* 2008, **7(10)**:1850-1862.
5. Davidson SB, Overton C, Buneman P: **Challenges in Integrating Biological Data Sources.** *Journal of Computational Biology* 1995, **2**:557-572.
6. Goble C, Stevens R: **State of the nation in data integration for bioinformatics.** *Journal of Biomedical Informatics* 2008, **41(5)**:687-693, Semantic Mashup of Biomedical Data.
7. Galperin MY: **The Molecular Biology Database Collection: 2008 update.** *Nucleic Acids Research* 2008, **36(suppl 1)**: D2-D4.
8. Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A: **BioMart Central Portal–unified access to biological data.** *Nucleic Acids Res* 2009, **37(Web Server issue)**:W23-W27.
9. Bhagat J, Tanoh F, Nzuobontane E, Laurent T, Orlowski J, Roos M, Wolstencroft K, Aleksejevs S, Stevens R, Pettifer S, Lopez R, Goble CA: **BioCatalogue: a universal catalogue of web services for the life sciences.** *Nucleic Acids Research* 2010, **38(suppl 2)**:W689-W694.
10. Zhao J, Goble C, Stevens R: **An Identity Crisis in the Life Sciences.** *Nucleic Acids Research* 2006, 254-269.
11. **Shared Names.** [http://sharedname.org].
12. van Iersel M, Pico A, Kelder T, Gao J, Ho I, Hanspers K, Conklin B, Evelo C: **The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services.** *BMC Bioinformatics* 2010, **11**:5.
13. Venkatesh T, Harlow H: **Integromics: challenges in data integration.** *Genome Biology* 2002, **3(8)**:REPORTS4027.
14. Akula SP, Miriyala RN, Thota H, Rao AA, Gedela S: **Techniques for integrating omics data.** 2009.
15. Al-Daihani B, Gray A, Kille P: **Bioinformatics Data Source Integration Based on Semantic Relationships Across Species.** In *Data Mining and Bioinformatics, Volume 4316 of Lecture Notes in Computer Science.* Springer Berlin / Heidelberg;Dalkilic M, Kim S, Yang J 2006:78-93.
16. Töpel T, Kormeier B, Klassen A, Hofestädt R: **BioDWH: a data warehouse kit for life science data integration.** *J Integr Bioinform* 2008, **5(2)**[http://www.biomedsearch.com/nih/BioDWH-Data-Warehouse-Kit-Life/20134070.html].
17. Achard F, Vaysseix G, Barillot E: **XML, bioinformatics and data integration.** *Bioinformatics* 2001, **17(2)**:115-125.
18. Hsing M, Cherkasov A: **Integration of Biological Data with Semantic Networks.** *Current Bioinformatics* 2006, **1(18)**:273-290.
19. Rector AL, Bechhofer S, Goble CA, Horrocks I, Nowlan WA, Solomon WD: **The GRAIL concept modelling language for medical terminology.** *Artificial Intelligence in Medicine* 1997, **9(2)**:139-171.
20. Stevens R, Aranguren ME, Wolstencroft K, Sattler U, Drummond N, Horridge M, Rector A: **Using OWL to model biological knowledge.** *International Journal of Human-Computer Studies* 2007, **65(7)**:583-594, Knowledge representation with ontologies: Present challenges - Future possibilities.
21. Bechhofer S, van Harmelen F, Hendler J, Horrocks I, McGuinness DL, Patel-Schneider PF, Stein LA: **OWL Web Ontology Language Reference.** *W3C Recommendation* 2004 [http://www.w3.org/TR/owl-ref/].
22. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.** *Nature Biotechnology* 2007, **25(11)**:1251-1255.
23. Berners-Lee T, Hendler J, Lassila O: **The Semantic Web.** *Scientific American* 2001, **284(5)**:34-43.
24. W3C Recommendation, World Wide Web Consortium: In *RDF Primer* Manola F, Miller E 2004 [http://www.w3.org/TR/rdf-primer/].
25. Ruttenberg A, Rees JA, Samwald M, Marshall MS: **Life sciences on the Semantic Web: the Neurocommons and beyond.** *Briefings in Bioinformatics* 2009, **10(2)**:193-204.
26. Antezana E, Kuiper M, Mironov V: **Biological knowledge management: the emerging role of the Semantic Web technologies.** *Briefings in Bioinformatics* 2009, **10(4)**:392-407.

27. Cheung KH, Frost HR, Marshall MS, Prud'hommeaux E, Samwald M, Zhao J, Paschke A: **A journey to Semantic Web query federation in the life sciences.** *BMC Bioinformatics* 2009, **10(Suppl 10)**:S10.
28. Dhanapalan L, Chen JY: **A case study of integrating protein interaction data using semantic web technology.** *Int J Bioinform Res Appl* 2007, **3(3)**:286-302.
29. Sahoo SS, Bodenreider O, Rutter JL, Skinner KJ, Sheth AP: **An ontology-driven semantic mashup of gene and biological pathway information: Application to the domain of nicotine dependence.** *J. of Biomedical Informatics* 2008, **41**:752-765.
30. Lam H, Marenco L, Clark T, Gao Y, Kinoshita J, Shepherd G, Miller P, Wu E, Wong G, Liu N, Crasto C, Morse T, Stephens S, Cheung KH: **AlzPharm: integration of neurodegeneration data using RDF.** *BMC Bioinformatics* 2007, **8(Suppl 3)**:S4.
31. 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* 2009, **10(Suppl 10)**:S11.
32. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J: **Bio2RDF: Towards a mashup to build bioinformatics knowledge systems.** *Journal of Biomedical Informatics* 2008, **41(5)**:706-716, Semantic Mashup of Biomedical Data.
33. Ruttenberg A, Rees J, Stephens S, Samwald M, Cheung KH: **A Prototype Knowledge Base for the Life Sciences.** *W3C Interest Group Note* [http://www.w3.org/TR/hcls-kb/].
34. Higgins JP, Wang L, Kambham N, Montgomery K, Mason V, Vogelmann SU, Lemley KV, Brown PO, Brooks JD, van de Rijn M: **Gene Expression in the Normal Adult Human Kidney Assessed by Complementary DNA Microarray.** *Mol. Biol. Cell* 2004, **15(2)**:649-656.
35. Chabardés-Garonne D, Méjean A, Aude JC, Cheval L, Di Stefano A, Gaillard MC, Imbert-Teboul M, Wittner M, Balian C, Anthouard V, Robert C, S□urens B, Wincker P, Weissenbach J, Doucet A, Elalouf JM: **A panoramic view of gene expression in the human kidney.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100(23)**:13710-13715.
36. **EuReGene - European Renal genome Project.** [http://www.euregene.org/].
37. Mischak H, Schanstra JP, Vlahou A: **Comprehensive human urine standards for comparability and standardization in clinical proteome analysis.** *Prot. Clin. Appl* 2010, **4(4)**:464-478.
38. Adachi J, Kumar C, Zhang Y, Olsen J, Mann M: **The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins.** *Genome Biology* 2006, **7(9)**:R80.
39. Bizer C, Heath T, Berners-Lee T: **Linked Data - The Story So Far.** *International Journal on Semantic Web and Information Systems (IJSWIS)* 2009.
40. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, Musen MA: **BioPortal: ontologies and integrated data resources at the click of a mouse.** *Nucleic Acids Research* 2009, **37(suppl 2)**: W170-W173.
41. **Gene Ontology Annotations.** [http://www.ebi.ac.uk/GOA].
42. **A reference ontology for biomedical informatics: the Foundational Model of Anatomy.** *Journal of Biomedical Informatics* 2003, **36(6)**:478-500, Unified Medical Language System.
43. Hayamizu T, Mangan M, Corradi J, Kadin J, Ringwald M: **The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data.** *Genome Biology* 2005, **6(3)**:R29.
44. Mungall C, Gkoutos G, Smith C, Haendel M, Lewis S, Ashburner M: **Integrating phenotype ontologies across multiple species.** *Genome Biology* 2010, **11**:R2.
45. Haendel MA, Neuhaus F, Osumi-Sutherland D, Mabee PM, Mejino JLJ, Mungall CJ, Smith B: **CARO - The Common Anatomy Reference Ontology.** In *Anatomy Ontologies for Bioinformatics Principles and Practice Volume.* Springer;Albert Burger, Duncan Davidson and Richard Baldock 2007:.
46. **Vertebrate Anatomy Ontology.** [http://www.ebi.ac.uk/ebiwiki/VBO/index.php/Main_Page].
47. Bard J, Rhee S, Ashburner M: **An ontology for cell types.** *Genome Biology* 2005, **6(2)**:R21.
48. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector A, Rosse C: **Relations in biomedical ontologies.** *Genome Biology* 2005, **6(5)**:R46.
49. Consortium GO: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Research* 2004, **32(suppl 1)**:D258-D261.
50. Horridge M, Tsarkov D: **Supporting Early Adoption of OWL 1.1 with Protege-OWL and FaCT++.** *OWL: Experiences and Directions (OWLED 2006)* Athens, GA; 2006.
51. Peters B, Ruttenberg A, Greenbaum J, Courtot M, Brinkman R, Whetzel P, Schober D, Sansone SA, Scheuerman R, Rocca-Serra P: **Overcoming the Ontology Enrichment Bottleneck with Quick Term Templates.** *International Conference on Biomedical Ontology* 2009.
52. O'Connor MJ, Halaschek-Wiener C, Musen MA: **Mapping Master: a Spreadsheet to OWL Mapping Language.** *International Semantic Web Conference (ISWC)* 2010.
53. **EuroKUP.** [http://www.eurokup.org].
54. Jupp S, Horridge M, Iannone L, Klein J, Owen S, Schanstra J, Stevens R, Wolstencroft K: **Populous: A template tool for populating OWL ontologies.** *Semantic Web Application and Tools for the Life Sciences* 2010.
55. **Semantic Web Health Care and Life Sciences (HCLS) Interest Group.** [http://www.w3.org/2001/sw/hcls].
56. **Entre Gene.** [http://www.ncbi.nlm.nih.gov/gene].
57. **UniProt Knowledgebase.** [http://www.uniprot.org].
58. **KEGG.** [http://www.genome.jp/kegg].
59. **microCosm.** [http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/].
60. **Homologene.** [http://www.ncbi.nlm.nih.gov/homologene].
61. **Bio2RDF ontology.** [http://bio2rdf.org].
62. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, Malone J, Parkinson H, Peters B, Rocca-Serra P, Ruttenberg A, Sansone SAA, Soldatova LN, Stoeckert CJ, Turner JA, Zheng J, OBI consortium: **Modeling biomedical experimental processes with OBI.** *Journal of biomedical semantics* 2010, **1 Suppl 1(Suppl 1)**:S7+.
63. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H: **Modeling sample variables with an Experimental Factor Ontology.** *Bioinformatics* 2010, **26(8)**:1112-1118.

64. Gkoutos G, Green E, Mallon AM, Hancock J, Davidson D: **Using ontologies to describe mouse phenotypes.** *Genome Biology* 2004, **6**:R8.
65. Rohloff K, Dean M, Emmons I, Ryder D, Sumner J: **An Evaluation of Triple-Store Technologies for Large Data Stores.** In *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops, Volume4806 of Lecture Notes in Computer Science.* Springer Berlin / Heidelberg;Meersman R, Tari Z, Herrero P 2007:1105-1114, 10.1007/978-3-540-76890-6_38.
66. **Comparison of Triple Stores.** [http://www.bioontology.org/wiki/images/6/6a/Triple_Stores.pdf].
67. Broekstra J, Kampman A, van Harmelen F: **Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema.** In *The Semantic Web - ISWC 2002, Volume 2342 of Lecture Notes in Computer Science.* Springer Berlin / Heidelberg;Horrocks I, Hendler J 2002:54-68.
68. Kiryakov A, Ognyanov D, Manov D: **OWLIM - A Pragmatic Semantic Repository for OWL.** In *WISE Workshops, Volume 3807 of Lecture Notes in Computer Science.* Springer;Dean M, Guo Y, Jun W, Kaschek R, Krishnaswamy S, Pan Z, Sheng QZ 2005:182-192[http://dblp.uni-trier.de/db/conf/wise/wise2005w.html#KiryakovOM05].
69. **Pubby.** [http://www4.wiwiss.fu-berlin.de/pubby/].
70. **Kidney and Urinary Pathway Knowledge Base (KUP KB).** [http://www.e-lico.eu/kupkb].
71. Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y: **Pellet: A practical OWL-DL reasoner.** *Web Semant* 2007, **5(2)**:51-53[http://dx.doi.org/10.1016/j.websem.2007.03.004].
72. Tsarkov D, Horrocks I: **FaCT++ Description Logic Reasoner: System Description.** *Proc. of the Int.Joint Conf. on Automated Reasoning (IJCAR 2006), Volume 4130 of Lecture Notes in Artificial Intelligence* Springer; 2006, 292-297.
73. Motik B, Shearer R, Horrocks I: **Hypertableau Reasoning for Description Logics.** *Journal of Artificial Intelligence Research* 2009, **36**:165-228.
74. **ON KUP Challenge.** [http://tunedit.org/challenge/ON].
75. Glimm B, Krötzsch M: **SPARQL Beyond Subgraph Matching.** In *Proceedings of the 9th International Semantic Web Conference (ISWC'10), Volume 6496 of LNCS.* Springer;Patel-Schneider PF, Pan Y, Glimm B, Hitzler P, Mika P, Pan J, Horrocks I 2010:241-256.
76. **SPARQL 1.1 Entailment Regimes.** [http://www.w3.org/TR/2010/WD-sparql11-entailment-20100126/].
77. Sirin E, Parsia B: **SPARQL-DL: SPARQL Query for OWL-DL.** *OWLED 2007: Proceedings of the Third International Workshop on OWL: Experiences and Directions* Innsbruck, Austria; 2007 [http://ceur-ws.org/Vol-258/paper14.pdf].
78. The Gene Ontology Consortium: **Gene Ontology: Tool for the Unification of Biology.** *Nature Genetics* 2000, **25**:25-29.

doi:10.1186/2041-1480-2-S2-S7
Cite this article as: Jupp *et al.*: Developing a kidney and urinary pathway knowledge base. *Journal of Biomedical Semantics* 2011 **2**(Suppl 2):S7.