Solving the Problem: SPARQL Access to Distributed Data Sources.

Position Paper: W3C Workshop on RDF Access to Relational Databases
Submitter: Dr Philip Ashworth
Email: philip.ashworth@ucb-group.com
Organisation Type: Life Sciences

Position

UCB has spent 12 months investigating semantic technologies with a view to understanding how operational issues within the organisation could benefit from these technologies.

Three areas of interest were quickly identified

1. Operational orchestration
2. Data Integration
3. Knowledge Understanding and Creation

Data Integration was further identified, as a key initial goal and has been the main focus of our efforts. This work has been directed towards the research organisation looking at internal data integration of scientific information from disparate data sources.

The data within UCB, as with most companies, is distributed across various structured and unstructured data sources.
Our initial aim was to focus on integration of data from a small number of relational data sources, for a select group of users within a particular scientific domain.
The data sources in question comprised of vendor and company devised schemas.

The users in question had the ability to register data into systems using a variety of client applications, but often had no way to view or interrogate the data from a single source, let alone bring data together from the various sources for decision making.

Our approach into semantic technologies has been to use vendor applications rather than in house development.
Our learning curve into semantic technologies has been steep, making many mistakes, learning about limitations and suffering many frustrations along the way. We have employed a key semantic consulting company to help and enhance our efforts in this area. However one of our main concerns, the actual data integration of the relational data sources into a semantic environment, remains a barrier still to be overcome.

We have investigated several avenues for the relational data integration problem, two of which are outlined below.

1. The concept of RDF data extraction and managing of the data within triple store/s, seemed a potentially viable option. Whilst this could be achieved from a short-term perspective, the realisation as a long-term solution could prove difficult. The

problems of data synchronisation (ETL) between relational databases and triple stores and subsequent remodelling of the concepts using domain and sub-domain ontologies followed by reasoning over the data set would prove time consuming. We wanted to achieve a real time solution to the integration problem, in order to show a distinct advantage over traditional warehousing techniques commonly used in the relational world.
In our project the volume of data has purposely been kept to a minimum. Yet we are still facing issues with the number of triples being generated and the current capacity of triple stores. This in turn led to issues with efficient reasoning over the data especially if this had to be done across data in separate triple stores.

2. We have also investigated querying of relational data directly using SPARQL in an attempt to remove the conversion issues mentioned above. Our efforts focussed on the use of D2RQ and the creation of mapping files for the conversion of SPARQL queries into SQL. In our hands this has also proved troublesome, due to problems from both the relational data sources and D2RQ limitations.

One problem we see with using current SPARQL to SQL conversion technologies is the subsequent loss of the potential power of semantic modelling and reasoning. This comes from the aspect of forward chaining and the static read only nature of the mapping files.
A further problem foreseen is that the SQL access to different data sources would be sequential, which could have a major impact on data access times.

Our project currently uses the triple store solution, however our hopes for the future lie with a SPARQL to SQL conversion utility that does not lose the capability for semantic understanding and has distributed asynchronous data access. In an attempt to solve this problem we have engaged with a vendor, with 15 years experience in distributed query arena, to look into the possibility of providing a SPARQL end point directly on top of their distributed server environment. We believe this could solve many of the issues we, and others, have today but also solve issues we see for the future. We have also brought together our semantic consultants with this company so that the tools are available within one of the leading ontology development environments. To date the work is on going, but is looking very promising. Both companies will be free to use the products of this collaboration however they see fit, so should this data access solution prove viable, the semantic community as a whole will benefit directly.

We believe RDF Access to Relational Databases is the biggest hurdle for adoption of semantic technologies within organisations, and would therefore like to participate in the W3C Workshop. By sharing our experiences and frustrations with others and vice-versa we believe that this hurdle can be overcome.