Accessing Existing Distributed Science Archives As RDF Models

Alasdair J. G. Gray[1], Norman Gray[2], and Iadh Ounis[1]

[1] Computing Science, University of Glasgow, Glasgow, UK.
[2] Physics and Astronomy, University of Leicester, Leicester, UK.

Due to the ease with which information can be published on the Internet, scientists potentially have access to more data than at any time in history. The Internet makes it viable for research laboratories to publish the results of their experiments for others to use, as text files, XML documents, or a database. There are also numerous archive centres which gather information from various sources and make it available from one place, often as a query endpoint to a relational database due to the size and highly structured nature of the data. In order to access this data, a scientist must be able to: (i) locate those data sources with information relevant to their research; (ii) understand the data model of each relevant data source in order to compose a query to extract the required data.

The resource description framework (RDF) [5] has been developed by the W3C with the aim of sharing and linking data on the Web. RDF is a graph based data model that makes the semantics of the data explicit, and can be queried using the SPARQL query language [6]. One of the benefits often stated for RDF is the ease with which data can be integrated from distributed RDF sources. To take advantage of this feature of RDF, tools for exposing relational databases as *virtual* RDF graphs have been developed including D2R [2] and SquirrelRDF [7]. In this work, we have considered these tools for exposing relational databases as SPARQL endpoints in the context of astronomical data archives.

There are many astronomical data archives available, each of which stores data according to their own relational schema. While work on developing a consensus model for accessing these archives is underway within the International Virtual Observatory Alliance (IVOA)[1], the current Characterisation Data Model [4] is far from expressing everything an astronomer needs. We are considering an alternative approach of using Semantic Web tools such as RDF and OWL [1] to help astronomers to locate and retrieve the data that they require.

Initially, we have focused on testing the database to RDF conversion tools with one astronomical archive, the SuperCOSMOS Science Archive (SSA) [3] whose schema was developed with 20 scientifically significant queries in mind. We found that only seven of these queries can be expressed in SPARQL due to the need for more powerful features, e.g. aggregates and mathematical functions[2]. Of these seven, only four were able to finish executing when posed over a 0.1% extract of the SSA[3] which contains 569,819 records. The execution times for these four queries when using different tools to access the data are shown in Fig. 1. The result show that D2R access will not scale to the size of databases used in astronomy. While in general SquirrelRDF performed reasonably well in comparison to D2R, it was at best about twice as slow than accessing the database through JDBC and could perform badly with certain queries, e.g. Query 2. These results are in-line with benchmark tests conducted on these tools [8].

Due to the limitations for expressing scientifically significant queries in SPARQL, and the poor efficiency performance of the database to RDF conversion tools, we suggest an alternative architecture for locating and extracting relevant data. The approach exploits the strengths of both the relational data model (storing and accessing vast quantities of structured data) and the RDF model (understanding the data semantics and integrating distributed ragged data). This architecture uses the RDF model of the data to help the scientist to locate and understand the data models of the distributed data sources. However, queries using the more powerful SQL query language are constructed by the scientist and posed directly to the data source. Another advantage of this approach is the ability to continue using existing astronomical data processing tools, which expect their input in a tabular format, as data is returned from the data sources in a relational format.

[1] `http://www.ivoa.net` Accessed 8 May 2008

[2] While it is possible to extend SPARQL with user defined functions, this will have further impact on the performance of these tools.

[3] `http://surveys.roe.ac.uk/ssa/pssa.html` Accessed 8 May 2008

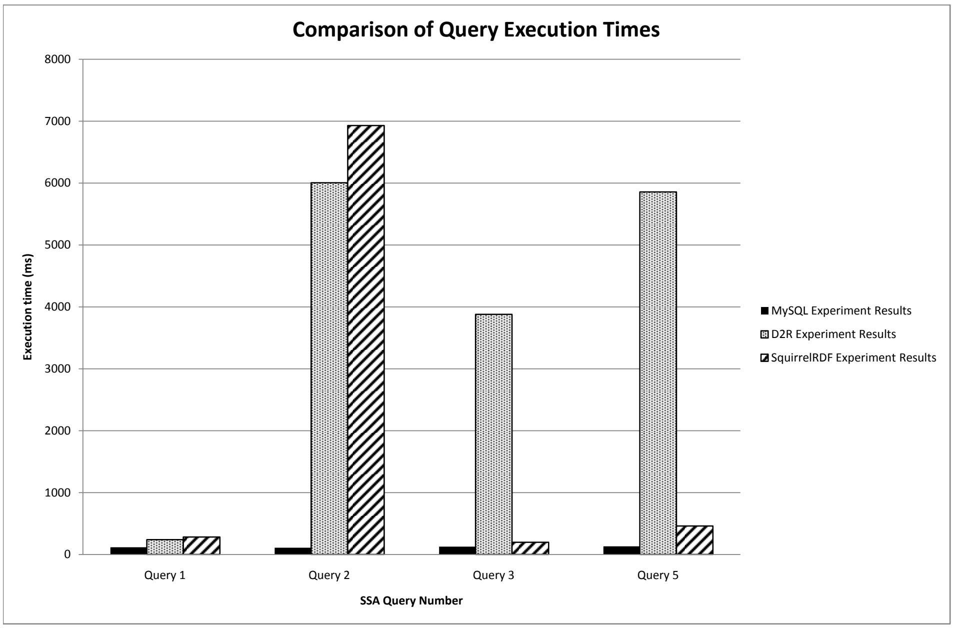

Fig. 1. Results of running queries overs the Personal SuperCOSMOS Science Archive using various data access tools.

Acknowledgements This research has been funded by the EPSRC grant number EP/E01142X/1.

This research has made use of data obtained from the SuperCOSMOS Science Archive, prepared and hosted by the Wide Field Astronomy Unit, Institute for Astronomy, University of Edinburgh, which is funded by the UK Particle Physics and Astronomy Research Council.

References

1. S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D.L. McGuinness, P.F. Patel-Schneider, and L. Andrea Stein. OWL web ontology language reference. Recommendation, W3C, 10 February 2004.
2. C. Bizer. D2R MAP - A database to RDF mapping language. In *WWW (Posters) 2003*, Budapest (Hungary), May 2003.
3. N. Hambly, M. Read, B. Mann, E. Sutorius, I. Bond, H. MacGillivray, P. Williams, and A. Lawrence. The SuperCOSMOS science archive. In *Astronomical Data Analysis Software and Systems XIII*, volume 314 of *ASP Conference Series*, pages 137–140, San Francisco (CA, USA), 2003.
4. M. Louys, A. Richards, F. Bonnarel, A. Micol, I. Chilingarian, and J. McDowell (eds). Data model for astronomical dataset characterisation. Recommendation, IVOA, 8 November 2007.
5. F. Manola and E. Miller (eds). RDF primer. Recommendation, W3C, 10 February 2004.
6. E. Prud'hommeaux and A. Seaborne (eds). SPARQL query language for RDF. Recommendation, W3C, 15 January 2008.
7. A. Seaborne, D. Steer, and S. Williams. SQL-RDF. In *W3C Workshop on RDF Access to Relational Databases*, Cambridge (MA, USA), October 2007.
8. M. Svihla and I. Jelinek. Benchmarking RDF production tools. In *DEXA 2007*, number 4653 in LNCS, pages 700–709, Regensburg, Germany, September 2007. Springer.