

myGrid's Web of Science Data Holdings

J. Zhao, R.D. Stevens, C.J. Wroe, M. Greenwood, C.A. Goble
Department of Computer Science
University of Manchester
Oxford Road
Manchester
United Kingdom M13 9PL

April 8, 2004

It is not enough to be able to run an e-Science experiment, it is also vital to be able to understand the outputs of those experiments. The results of an e-Science *in silico* experiment are of reduced value if other scientists are not able to identify the origin, or *provenance* of those results. Many experiments, especially those in bioinformatics, are run repeatedly, orchestrating many resources to produce sets of data for analysis and validation by a human scientist. Such experiments produce much fragmented data, each from a separate resource and these data need to be co-ordinated with each other. Records of the production of these data enable a scientist to ask questions such as 'which experiments also use this biological entity as input?', 'from which data were these derived?', 'which service was used with which inputs and outputs in order to do so?', etc. In the scientific process, such questions need to be asked from multiple viewpoints across multiple experiments. These are questions of the *provenance* of results and provenance must be a cornerstone of e-Science.

In the myGrid project, we model provenance for e-Science experiments, to provide the most common and expected context (what, which, why, when, where and who), resource (where), and derivation (how) information for projects, experiments, services and data in *in silico* experiments. We supply provenance from four viewpoints:

1. Organisation provenance – Users, projects, institutions, etc.
2. Enactment provenance – the services used, their versions, locations, duration of use, etc.
3. Data provenance – the data from which inputs, intermediates and final outputs were derived during an experiment.
4. Knowledge provenance – concepts and relationships from ontologies of bioinformatics and molecular biology that describe the nature of the bioinformatics and biological context in which the other types of provenance lie.

Currently only sporadic and incomplete provenance logs are available for *in silico* experiments. For e-Science we need systematic and automatic production of provenance records and an informed understanding of the information in these provenance data. In myGrid the production, storage and visualisation of provenance information are important aspects of the added value provided to the e-scientist.

At the core of myGrid's provenance support is a common information model [2]. The myGrid information model provides a standard by which to structure information about bioinformatics experiments and data. In brief, the model splits into two parts, that reflect the views of provenance above: 1) organisational information such as the members of the research group, data access rights, current projects and their experiments; 2) information about the life cycle of a single experiment such as its design, when it has been performed, results it has produced and provenance of those results.

The provenance is produced by the Taverna workbench for creating and running workflows. Taverna does not just produce results, but also metadata about those results. The enactment provenance and data provenance are produced by the workbench's Freefluo enactment engine. Taverna does not just record the data values; it also identifies the specific data instances so that their creation and use

can be tracked. To do this, Taverna uses the Life Sciences Identifier (LSID) scheme, a technique for associating a Universal Resource Identifier (URI) with data [1]. Taverna interacts with an LSID authority and a database to store the provenance information. The authority generates the LSID's required, and the database stores both the data value and RDF statements about these values. This approach has considerable flexibility: the provenance information can be browsed by any software that can work with an LSID authority. Experiments have included using the LSID Launchpad plugin for Internet Explorer as a simple, single-item oriented viewer, and the RDF-browsing capabilities of Haystack for richer views of the provenance graph [2]. Of particular note is the fact the producing, storing and visualising provenance information has changed the workflow descriptions. The workflow descriptions now include RDF templates of relationships that are completed and stored as the workflow executes. This means that users can provide richer provenance statements, describing the semantic relationship between data items, in addition to the default statements that are created for all workflows.

^{my}Grid enables the user to explore the context of experimental data by providing associated metadata with each item. This metadata uses a schema derived from the information model. Items are also marked up with terms from an ontology of bioinformatics and molecular biology to facilitate browsing and querying of this experience base of *in silico* experiments. This cross-linking and mark up enables the user to explore the context of experimental data by providing associated metadata with each item.

These cross-references allow a Web of science to be created, such that a user can navigate to validate results, find other related results and generate different views over his or her body of scientific work. As workflows are enacted, filaments of such Webs are produced, showing the derivation path of results, and the context of the experiment. These are also linked to provenance of the enactment itself, which shows which services were used, at what time, where and with which inputs. The combination of organisation, enactment, data and knowledge provenance means a scientist can perform a wide range of tasks within the scientific process. In this way, ^{my}Grid aims to allow scientists to capitalise on e-Science.

References

- [1] Tim Clark, Sean Martin, and Ted Liefeld. Globally Distributed Object Identification for Biological Knowledgebases. *Briefings in Bioinformatics*, 5(1):59–70, 2004.
- [2] R.D. Stevens, H.J. Tipney, C.J. Wroe, T.M. Oinn, M. Senger, P.W. Lord, C.A. Goble, A. Brass, and M. Tassabehji. Exploring Williams Beuren Syndrome Using ^{my}Grid. Accepted for publication at Intelligent Systems for Molecular Biology (ISMB) 2004, 2004.