The Origin and History of *in silico* Experiments

J. Zhao & R.D. Stevens & C.J. Wroe & M. Greenwood & C.A. Goble

Department of Computer Science
University of Manchester
{zhaoj, robert.stevens, wroec, markg, carole}@cs.man.ac.uk

Abstract

It is not enough to be able to just run an e-Science *in silico* experiment; it is also vital to be able to understand and interpret the outputs of those experiments. The results have little value if other scientists, or even the same scientist at a later date, are unable to identify their origin, or *provenance*. In myGrid, *in silico* experiments are run as workflows; these produce three kinds of results: data outcomes, knowledge outcomes and provenance about the experiment. These results have a complex interlinking relationship between each other, within the context of the workflow that gave rise to them, as well as across workflows executed in the same or a different study. This poster describes the kind of provenance data recorded in myGrid during a workflow. It introduces myGrid's provenance data model and the Semantic Web-based technology used to support provenance-based tasks. These tasks include the verification and validation of results; the sharing and annotation of results; and the management of resources. For e-Science to succeed it must have provenance data support as its cornerstone.

1 Introduction

Many experiments, especially those in bioinformatics, are run repeatedly, orchestrating many resources to produce sets of data for analysis and validation by scientists. Such experiments produce a great deal of fragmented data, each from a separate resource and these data need to be coordinated with each other.

myGrid is a UK e-Science pilot project which is developing Grid middleware infrastructure for *in silico* experiments in biology. myGrid regards each *in silico* experiment run as a workflow. These workflows automate experiments by coordinating the services that process data. These data outcomes, or indeed the workflows themselves, have little use unless they are accompanied by their origin or *provenance*. The provenance gives scientists the metadata to allow them to interpret the experimental context and the experiment's outcomes. In the myGrid project, we model provenance to provide the most common and expected context (what, which, why, when, where and who), resource (where), and derivation (how) information about *in silico* experiments.

2 Provenance in myGrid

A myGrid workflow carries metadata describing its origins: where it came from, who designed it and for what purpose; what hypothesis is being test and why. Broadly speaking, it consumes data inputs and produces three kinds of results: data outcomes, knowledge outcomes and provenance records about the experiment. Data outcomes are the intermediate and final results; knowledge outcomes are the findings in terms of the scientific domain and the overall experiment; and provenance provides the repeatable record of the workflow (who ran it, when, over what resources and data), the derivation path of the data and the evidence to back the findings. These results have a complex interlinking relationship, within the context of the workflow that gave rise to them, as well as across workflows executed in the same or a different study.

Figure 1 gives a four part framework for differentiating between classes of provenance:

Process level, collects how, when and where the workflow is run, what data are used and generated, which computational services are invoked, and the input and output data for service invocation. An example is in Figure 2A, the BLAST version 3.1 run on 20040504 at13:34 GMT was invoked with a nucleotide sequence URN:lsid:taverna:datathing:13, and successfully executed in 2.1 minutes;

Data level, primarily inferred from the process level provenance, describes the

derivation path of the data final and intermediate results. For example (again in Figure 2A), a collection of nucleotide sequences URN:lsid:taverna:datathing:15 from Genbank are derived from URN:lsid:taverna:datathing:13 by BLAST version 3 etc;

Organization level, records the workflow user and creator, their organization, project, the hypothesis for this experiment/project, the experiment design, etc. It can be attached to any computational resources (including data, services, workflows, etc) recorded in the process and data level provenance, as shown in Figure 2B. For example, H. Tipney of the University of Manchester, designed workflow W21 whose purpose is to characterise a newly submitted sequence with its predicted genomic function; this workflow is part of a greater experiment to identify the genes associated with Williams-Beuren Syndrome;

Knowledge level, links the knowledge outcome of the workflow with the process, data and organization provenance that provide the evidence to support it. For example, the knowledge outcome of the workflow fragment in Figure 2A is that a nucleotide sequence (datathing:13) has *similar sequences to* a collection of sequences in a BLAST report (datathing:15). The provenance (origin) of this knowledge, which gives its evidential backing, is the process and data provenance, and a template that the scientist has used to extract this finding from the workflow. The template can be workflow, experiment and scientist specific. Other knowledge outcomes, such as a scientist's personal conclusions, are captured as semi-structured annotations.

The organization provenance is gathered either explicitly or implicitly as a side effect of signing onto a portal. The process and data provenance are mechanistic and straightforwardly gathered as side effects of the workflow enactment engine. The knowledge level is more complex. The template was developed in order to (a) bridge across long workflows where there may be some distance between the objects the scientist is interested in; (b) to filter out the SHIM adapter services [3] that play no part in the biological intention of the workflow but are essential for its successful execution and (c) to allow a semantic interpretation of the data objects (nucleotide sequence rather than id:13) and their relationship ("similar sequence to" rather than "derived from"). This knowledge is not provenance itself - it is a first class experimental outcome, albeit a subjective interpretation by the scientist - but the knowledge level ties it to the process/data provenance records from which it was derived.

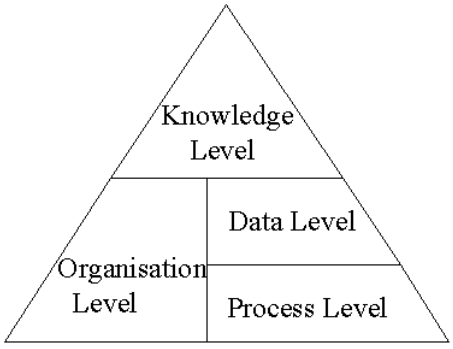

Figure 1: The provenance pyramid.

3 Provenance Generation and Presentation

Provenance generation is supported by the provenance data model, part of the myGrid information model [6], which provides a standard by which to structure information about bioinformatics experiments. In brief, the model splits into two parts, which reflect the three lower levels of provenance: (a)**organisational information** such as the members of the research group, data access rights, current projects and their experiments; (b)**experimental information** about the life cycle of a single experiment such as its design, when it was performed, results it produced and provenance of those results.

This process, data and organisation provenance is automatically generated when creating and running workflows with the Taverna[1] toolkit [4], including the Freefluo[2] workflow enactment engine. The Life Sciences Identifier (LSID) [2] scheme is used to uniquely and persistently identify and resolve data resource and its associated metadata. The use of LSIDs is an important part of the openness of the provenance within myGrid. The adoption of LSIDs means that any tool which can interact with an LSID resolver can be used to browse provenance data. An example tool is the LSID Launchpad[3] plug-in for Internet Explorer, a simple, single-item oriented viewer for resources identified by LSIDs. Also, the provenance model is not tied to a specific workflow language or system. Any workflow, or service invocation tool, that can interact with the myGrid LSID authority can create provenance data.

The Resource Description Framework (RDF) is used to represent the provenance graph, with LSIDs as the Unique Resource Identifiers (URIs) to link resources together. In order to present a rich view of this RDF provenance graph, Haystack [5] is applied.

[1] http://taverna.sourceforge.net
[2] http://freefluo.sourceforge.net
[3] http://www-124.ibm.com/developerworks/oss/lsid/

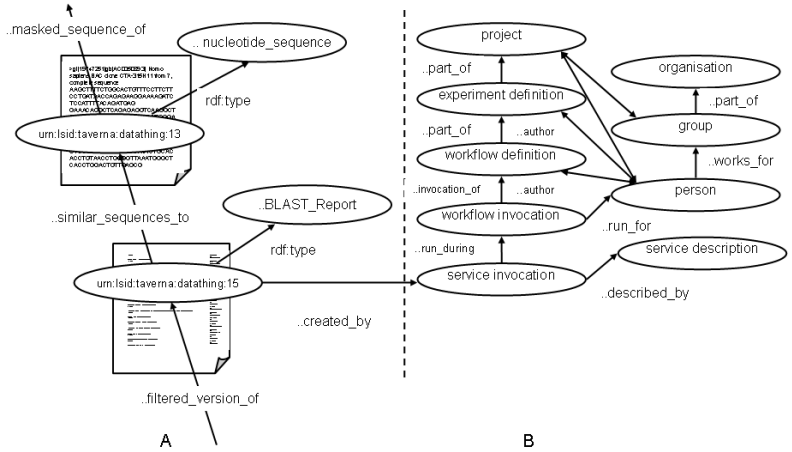

Figure 2: (A) shows relationships between data items and the knowledge outcomes that this can generate; (b) shows the organisational and contextual provenance for the data/process/knowledge graphs in (A).

Figure 3 shows an example RDF graph viewed in Haystack, which displays how resources are linked together in four different views (process, data, organisation and knowledge views).

4 Knowledge from Provenance

The provenance record of one experiment is crucial to its verification, its ability to be repeated and its credibility. However, our intention is not to treat these logs in isolation but to gather, link, query, integrate and exploit them collectively. For example, scientists would like to discover the experiments that use the same service invocation, where are the similar DNA sequences to their data result, or the conceptual connection between computational data resources from a set of experiments, etc, which can be gleaned from all levels of provenance. Service providers would like to know the quality, performance, reliability of data and services, which can be gleaned from the process and organization levels.

To better achieve an interpretation of the provenance we create a Semantic Web for provenance, which connects across the logs and across the layers through their semantic relationships. A first attempt to discover this knowledge from provenance is by annotating concepts from ontologies and linking these logs conceptually. The result from this experience demonstrates the feasibility of building a knowledge web of provenance resources via a semantic annotation and linking approach. However, the automation of this annotation of scientific data resources is challenging [8]. Thus, RDF and LSIDs are applied to represent provenance data resources and their semantic relationships, with an expectation of more flexibility for building a knowledge web of provenance.

As shown in Figure 3, provenance logs across experiments and across layers are linked together using semantic relationships (e.g. *derived from*, *similar sequence to*, etc) defined in RDF. From the Figure 3, one can tell that the FASTA format output is a *protein sequence*, *created by* a RETRIEVE service, *owned* by Hannah, and is a *similar sequence* to another data, etc. Thus, users can choose to discover knowledge from provenance across experiments through the four different views.

5 Related work

The emphasis of the myGrid provenance is on statements as glue that are used to bind experimental information together, and thus providing a richer web of science for users to browse, filter and query as required. To do this we concentrate on ”unsecured”, coarse-grain provenance. It is coarse grain in that it only depends on recording the inputs and outputs of services, and the externally available service information. For example, where a service provides an interface to a genomic database, the coarse-grain provenance typically records a name used to identify the database (and perhaps a version number) but no details of the actual database entries. In contrast fine-grain provenance records the actual database entries that contributed to an output of a database query [1].

The myGrid provenance is ”unsecured” in that it assumes that the provenance generator is trustworthy and will not attempt to forge provenance for any data. Non-repudiation is required to provide the secure provenance needed for a scientist to prove that their results were obtained in the manner and time described. This is what the project PASOA (Provenance Aware Service Oriented Architecture) [7] proposes to provide provenance aware services so that the provenance information can be recorded from

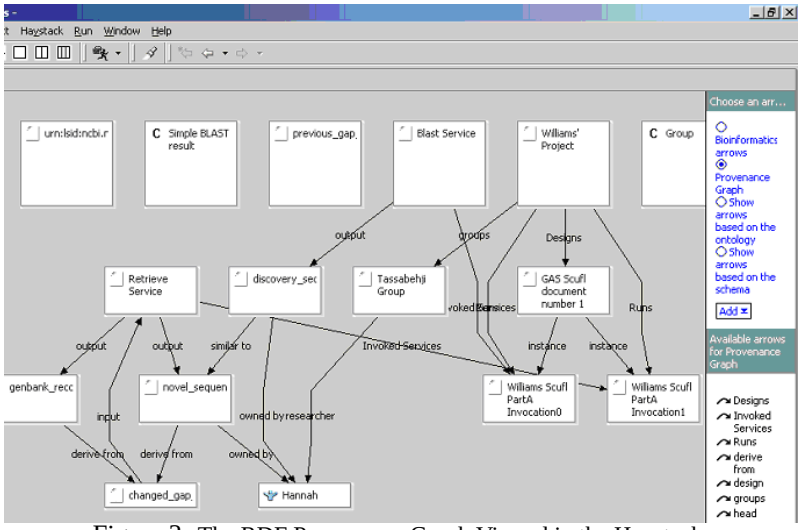

Figure 3: The RDF Provenance Graph Viewed in the Haystack.

the perspective of the service client/user and from the perspective of the service itself. It also requires a trusted service for storing provenance data.

6 Conclusion

The myGrid provenance graph model provides a rich view of provenance resources about experiments from four different levels. In addition, it supports the discovery of the knowledge level relationship between provenance resources. The cross-references to provenance resources across experiments allow a Web of science to be created, such that a user can share, discover & re-use resources, navigate to validate results & draw conclusions, find other related results and generate different views over his or her body of scientific work. The combination of process, data, organisation and knowledge provenance means a scientist can perform a wide range of tasks within the scientific process. In this way, myGrid aims to allow scientists to capitalise on e-Science.

Acknowledgement

The authors would like to acknowledge the assistance of the whole myGrid consortium. This work is supported by the UK e-Science programme EPSRC grant GR/R67743.

References

[1] Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and Where: A Characterization of Data Provenance. In *International Conference on Database Theory (ICDT)*, 2001.

[2] Tim Clark, Sean Martin, and Ted Liefeld. Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics*, 5(1):59–70, 2004.

[3] Phillip Lord Carole Goble Duncan Hull1, Robert Stevens2. Integrating bioinformatics resources using shims. In *Poster Accepted in Intelligent Systems for Molecular Biology (ISMB)*, Glasgow, UK, August 2004.

[4] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat, and Peter Li. Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 2004.

[5] Dennis Quan and David R. Karger. How to make a semantic web browser. In *Thirteenth International World Wide Web Conference*, New York, USA, 2004.

[6] Robert Stevens, Hannah J.Tipney, Chris Wroe, Tom Oinn, Martin Senger, Phil Lord, Carole Goble, Andy Brass, , and May Tassabehji. Exploring williams-beuren syndrome using mygrid. In *Accepted in Intelligent Systems for Molecular Biology (ISMB)*, Glasgow, UK, August 2004.

[7] Martin Szomszor and Luc Moreau. Recording and reasoning over data provenance in web and grid services. In *International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE'03)*, volume 2888, pages 603–620, Catania, Sicily, Italy, November 2003.

[8] Jun Zhao, Carole Goble, and Robert Stevens. Semantically linking and browsing provenance logs for e-science. In *First International Conference on Semantics of a Networked World*, pages 157–174, 2004.