Provenance of e-Science Experiments - experience from Bioinformatics

Mark Greenwood, Carole Goble, Robert Stevens, Jun Zhao, Matthew Addis,
Darren Marvin, Luc Moreau, Tom Oinn
EPSRC e-Science Pilot Project myGrid
http://www.mygrid.org.uk

Abstract

Like experiments performed at a laboratory bench, the data associated with an e-Science experiment are of reduced value if other scientists are not able to identify the origin, or provenance, of those data. Provenance information is essential if experiments are to be validated and verified by others, or even by those who originally performed them. In this article, we give an overview of our initial work on the provenance of bioinformatics e-Science experiments within myGrid. We use two kinds of provenance: the derivation path of information and annotation. We show how this kind of provenance can be delivered within the myGrid demonstrator WorkBench and we explore how the resulting Webs of experimental data holdings can be mined for useful information and presentations for the e-Scientist.

1 Introduction

Like experiments performed at a laboratory bench, the data associated with an e-science experiment are of reduced value if other scientists are not able to hold the provenance of those data. Provenance is a kind of metadata, recording the process of biological experiments for e-Science, the purpose and results of experiments as well as annotations and notes about experiments by scientists [1]. This information is essential if experiments are to be validated and verified by others, or even by those who originally performed them. It is also important in assessing the quality, and timeliness of results.

Our investigation of provenance is part of the myGrid project, which is reviewed in Section 2. We then describe our understanding of provenance data for *in silico* experiments and how it is generated. We then present how we think these data can be exploited in Section 4 and conclude with a discussion in Section 5.

2 The myGrid project

e-Science is the use of electronic resources – instruments, sensors, databases, computational methods, computers - by scientists working collaboratively in large distributed project teams in order to solve scientific problems. An *in silico* experiment is a procedure that uses computer-based information repositories and computational analysis to test a hypothesis, derive a summary, search for patterns, or demonstrate a known fact. The myGrid project[1] [5] is developing high-level service-based middleware to support the construction, management and sharing of data-intensive *in silico* experiments in biology. The generation and production of provenance data is seen as a central requirement for the myGrid project.

[1] http://www.mygrid.org.uk

3 Generating Provenance in myGrid

Within myGrid we want to generate and exploit provenance that is intended for machine consumption. The large amount of provenance data that is required to allow scientists to verify the experiments of others, means that as much of the collection as possible must be automatically and systematically generated. There is considerable potential for programs that process collections of provenance records and provide users with higher-level views of what has happened within their virtual organisation.

There are two major forms of provenance: First, the *derivation path* records the process by which results are generated from input data. This could include a database query, a program and its parameters, or a workflow that orchestrates a number of services. Knowledge of the derivation path is essential for effective change management or knowing when an experiment needs to be re-run in the light of new information [6]. Second, *annotations* are attached to objects, or collections of objects. There are standard annotations such as when an object was created, last updated, who owns it and its format. In addition, there can be annotations that describe an object with concepts related to the scientific domain. For example, a database entry is a protein sequence in FASTA format and it describes an enzyme with a glucose substrate

and a kinase function. All annotations add context information that is essential to the interpretation of the raw scientific data.

The workflows that represent *in silico* experiments in myGrid describe the orchestration of bioinformatics data and analysis services that are used to derive outputs. The outputs of one workflow may form the inputs to another so that a complete *in silico* experiment includes a network of related workflow invocations. The myGrid workflow enactment system ensures that any output can be associated with its corresponding workflow invocation record and the associated provenance data. In this way, the detail of how an output is related to its inputs is available when required [7].

The myGrid environment uses the open source Freefluo workflow enactment engine[2]. Coupled with the myGrid WorkBench [4], this enactment engine produces a provenance log that records what events have been performed during the enactment. Figure 1 shows the myGrid schema for workflow or *derivation path* provenance information myGrid generates. This schema is being further extended to include more *annotation* provenance.

The workflow provenance log provides a record of the *derivation path* of the workflow. Thus the provenance log of a workflow enactment is the recording of the start time, end time and service instances operated in this workflow. At the end of the workflow execution, the resulting data, metadata about the workflow and the provenance logs are held in the myGrid Information Repository (mIR) [2]. When the myGrid WorkBench is used, there is additional *annotation* about the context of the workflow. In addition to the derivation path, a set of metadata is associated with the workflow invocation instance: the input and output relationships between the workflow instance and data items, the 'is defined by' relationship between the workflow instance, the semantic description document and the workflow template. Other annotations regarding the hypothesis of the experiment, thoughts and opinions by the scientist and quality of results are also stored as XML in the mIR or as regular web documents[3].

3.1 Beyond workflow provenance

It is also clear that there is much more than just data provenance of workflow outputs to consider. The scientists using myGrid may want to record information about the provenance of data that they load directly into the mIR. For example if the data is a DNA sequence, then scientists might want to store:

- some note on where it came from;
- some other biological information such as species, function, etc;
- comments on why this data was being used;
- standard mIR metadata (who input the data, when, its syntactic, semantic and display(MIME) type).

The workflow descriptions are themselves objects within the environment that can have their own provenance data. Bioinformatics data and analysis services may themselves have provenance information indicating:

- their versions;
- default parameters;
- resource versions - For services that involve searching databases, it can be important to know what version of the database was used.

This is of particular relevance in bioinformatics, where there are a large number of secondary databases whose content is based both on automatic annotation of the primary data, and on the additional annotations of a small number of human experts (curators). This means that there is an inevitable time lag between changes to the primary data, and the corresponding changes in secondary databases.

Running workflows is not the only activity that goes on within the myGrid environment. Other activities might include: Tracking what users have selected; tracking the texts that users refer to; and understanding what colleagues have done. Potentially all actions of the user can be recorded and act as annotations; indeed, these can be themselves annotated with the input of the e-Scientist.

All this information provides a 'work context' for *in silico* experiments. We would like to build a web of related pages relevant to an experimental investigation, marked up with, and linked together using annotations drawn from shared ontologies (see Figure 2). This web includes not only the provenance record of a workflow run, but also links to other provenance records of other directly or indirectly related workflow runs, diagrams of the workflow specifications; web pages about people who ran the workflow or have related study in provenance; literatures relevant to provenance study; notes of the experiment and so on. This is the idea behind a "web of science" as proposed by Hendler [3].

4 Exploiting Provenance Records

There are a potentially wide number of uses for this provenance data:

[2] `http://freefluo.sourceforge.net/`

[3] An example of a myGrid provenance record may be seen at `http://twiki.mygrid.org.uk/twiki/bin/view/Mygrid/GDProvenanceExample`. An example of the additional information available through the mIR may be seen at `http://twiki.mygrid.org.uk/twiki/bin/view/Mygrid/ProvenanceData\#provenance_from_mIR_metadata`.

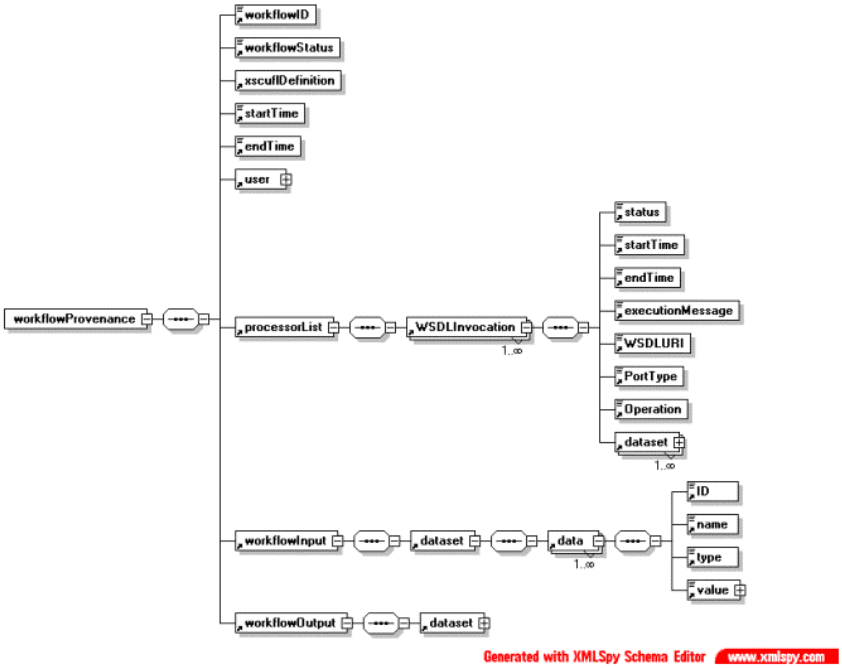

Figure 1: The myGrid information model for provenance information.

- The derivation path provenance should be enough to allow others to repeat and validate an experiment. If exactly the same conditions are available (same database version, same program, same algorithm, etc.) then it may be possible to repeat the *in silico* experiment.

- The provenance data also provides a store of know-how, a means of learning from history and disseminating best practise.

- There is also substantial benefit in using provenance to give scientists a view across multiple experiments. Who in my community has worked on similar data, in the last six months, and can I adapt their workflow descriptions to my current experimental context? This will give e-Scientists the ability to share community knowledge, best practice and know-how.

- The web of experiment data-holdings enabled by provenance information allows a variety of personalised views to emerge: An experiment, user, data, organisation, project, etc. centric view are all possible from this kind of information.

- Scientists always wish to know if the experiment they wish to run or the hypothesis they wish to test has been performed or explored before – *in silico* work should be no different.

- Provenance information can be useful from a management point of view; 'what are the most used bio-services in my organisation/group/project?'; 'is is worth renewing my subscription to this expensive service?'. Such queries obviously have security and privacy implications – another fundamental topic for e-Science.

- In the volatile world of bioinformatics such information could be used to automatically re-run an *in silico* experiment in the light of a notification of change in data, analysis tool or third-party data repository.

5 Discussion

All empirical scientists know that keeping a good logbook is vital. This is also true for the empirical e-Scientist. One of the big efforts within the myGrid project is to make the gathering of the *provenance* information from *in silico* experiments both automatic and systematic. These metadata can then be placed within a user's information repository for additional uses. These data can also be supplemented by the e-Scientist him- or herself.

The ability to provide this information is dependent upon bioinformatics service providers providing the relevant information. myGrid envisages the design of a provenance port type - placing obligations on service providers to give provenance information. This Port Type would have to be implemented in order to be

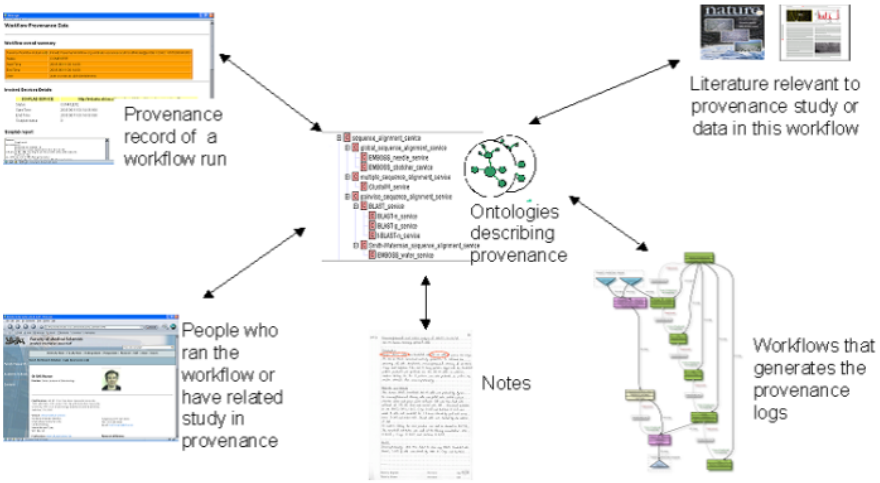

Figure 2: the web of annotated documents envisaged from the use of myGrid provenance information.

myGrid compliant.

The work presented here is very much an initial exploration of an important area in e-Science. Our derivation paths and annotations of data holdings are prototypes to be used as a tool to explore further user requirements. We envisage the provenance data to be a resource in its own right, driving the personalisation of e-Science through the idea of a 'web of science'. Although our experience is from bioinformatics, the issues of provenance range across e-Science, and we expect this work to encourage others to compare and contrast their experiences with ours.

Acknowledgements: This work is supported by the UK e-Science programme EPSRC GR/R67743. We would also like to acknowledge the assistance of the whole myGrid consortium.

References

[1] P. Buneman, S. Khanna, K. Tajima, and W-C Tan. Archiving Scientific Data. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, 2002.

[2] Carole Goble, Chris Wroe, Robert Stevens, and the myGrid consortium. The myGrid Project: Services, Architecture and Demonstrator. to appear in Proc UK e-Science programme All Hands Conference, 2-4 Sept 2003, Nottingham, UK, 2003.

[3] J. Hendler. Science and The Semantic Web. *Science*, page 24, Jan 2003.

[4] Robert Stevens, Kevin Glover, Chris Greenhalgh, Claire Jennings, Simon Pearce, Peter Li, Melena Radenkovic, and Anil Wipat. Performing *in silico* Experiments on the Grid: A Users' Perspective. to appear in Proc UK e-Science programme All Hands Conference, 2-4 Sept 2003, Nottingham, UK, 2003.

[5] Robert D. Stevens, Alan J. Robinson, and Carole A. Goble. myGrid: personalised bioinformatics on the information grid. *Bioinformatics*, 19:i302–i304, 2003.

[6] 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)*, lncs, Catania, Sicily, Italy, November 2003.

[7] J. Zhao, C.A. Goble, M. Greenwood, C. Wroe, and R. Stevens. Annotating, linking and browsing provenance logs for e-science. submitted to: 2nd Intl Semantic Web Conference (ISWC2003) Workshop on Retrieval of Scientific Data, Florida, USA, October 2003, 2003.