Title: *myGrid: An EU Provenance Case Study*

Author: *Simon Miles, Victor Tan* University of Southampton
Daniele Turi, Katy Wolstencroft, Jun Zhao University of Manchester

Version: *1.0*

Date: *29th May 2006*

Status: *Public*

Summary

This document describes a particular scenario developed as part of the myGrid project and provenance questions apparent within that scenario. This analysis should be a useful source for any project, including Provenance and myGrid attempting to provide a useful provenance infrastructure.

Members of the PROVENANCE consortium:

- IBM United Kingdom Limited United Kingdom

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

- University of Southampton United Kingdom
- University of Wales, Cardiff United Kingdom
- Deutsches Zentrum fur Luft- und Raumfahrt e.V. Germany
- Universitat Politecnica de Catalunya Spain
- Magyar Tudományos Akadémia Számítástechnikai és Automatizálási Kutató Intézet Hungary

Introduction

This document describes a particular scenario developed as part of the myGrid project and provenance questions apparent within that scenario. This analysis should be a useful source

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

for any project, including Provenance and myGrid attempting to provide a useful provenance infrastructure. We then apply the Provenance methodology to determine how the provenance questions would be addressed. While this analysis may also be useful regardless of the provenance system used, it is most applicable to the PASOA/Provenance architecture.

Overview of Application

A scientist is studying Williams Beuren Syndrome (WBS) and has encoded one in-silico experiment they perform as part of this study as a workflow. The workflow aims to find genes that may be relevant to WBS and that they have not examined already. This workflow is regularly enacted, as new gene data will be added to databases over time.

Physical Distribution

The deployment of people and resources in the scenario is as follows. The scientist (User) enacts a workflow using a Workflow Enactor. The workflow makes use of Databases at EBI, NCBI and possibly other providers. It also makes use of external tools provided by EBI, such as BLAST. Results from the workflow are stored in a Data Store. The arrows in the figure show communication between these entities during the experiment. There are three locations (also corresponding to three security domains): the scientist's local lab and one location for each of the providers.

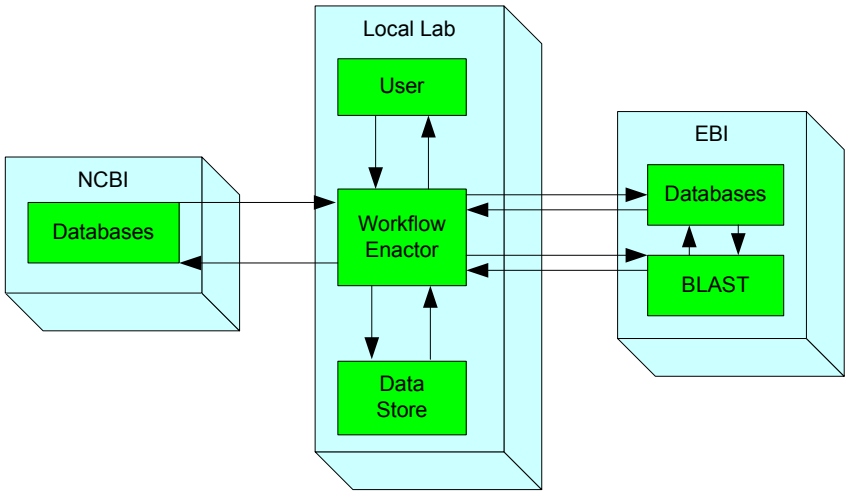


Figure 1: The high-level actors in the scenario and their distribution

Workflow

The workflow is depicted below in a reduced form for the purposes of explanation (the full workflow is given in the Appendix). The fixed inputs to the workflow are the name of the default database in which to search for potentially relevant gene sequences (Database Name), a partially known sequence thought to be relevant to WBS (Masked Sequence) and the set of sequences already studied (Old Sequences). The default database name is passed to the user to confirm or change to another value (User Interaction), and then used as an input to BLAST, along with the Masked Sequence, to find similar sequences which may provide the rest of the details of the partially known sequence. The BLAST produces a report which is kept. The sequences output from BLAST are filtered to include only those on the relevant chromosome, human chromosome 7 (Filter). The filtered results are then

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

compared with the previously studied sequences to remove those already studied (Compare). As a final step (NCBI Converter) the GenBank database accession IDs for the remaining sequences are determined and output (GenBank IDs). The scientist will later use these IDs to find the sequence data to further analyse. All services apart from BLAST are deployed local to the lab, but NCBI Converter communicates with NCBI's GenBank database.

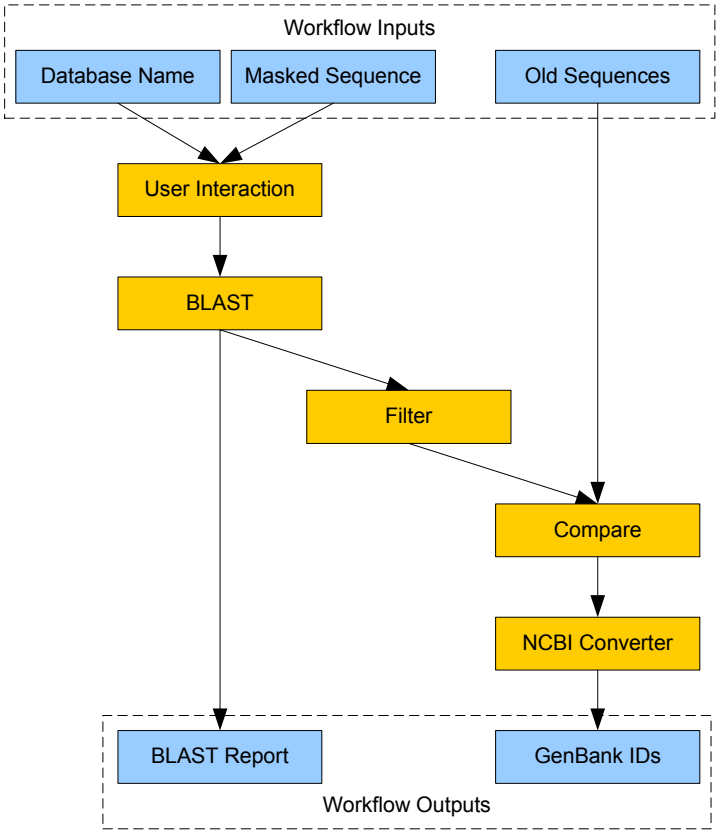


Figure 2: The actors, data and data flow making up the workflow

Provenance Questions

After performing the experiment several times, the scientist asks the following provenance-related questions. The provenance questions are all with respect to the results, GenBank IDs, of one or more workflow runs.

PQ1: *Were the sequences output from the Filter step actually restricted to chromosome 7?*

This is a test of whether the Filter step is working correctly.

PQ2: *In which experiments did BLAST output sequences from species X?*

Sequences from non-human species will not be apparent in the results, but it would be interesting to see whether, over the course of several experiment, BLAST has often returned sequences from another species.

PQ3: *Did the user change the database name?*

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

By examining how often the user changes the default database, we can determine whether it is a good default (if changed often, it is unlikely to be a good default). Ultimately, this may be useful in establishing good defaults in a workflow.

Security Issues

The local lab has a security domain in which only the lab scientists can access and modify the data and process documentation. EBI and NCBI have databases, metadata stores which only they can write to but which can be read by everyone.

Scalability Issues

The distribution is as shown in Figure 1. The application data is of manageable size.

Information Items

An *information item* is a piece of information about a process that needs to be known before the above provenance questions can be answered.

Provenance Question 1

In order to answer PQ1, we need to know information items II1, II2, II3, II4 and II5.

II1: *The sequences output from Filter.*
II2: *The chromosome that the sequence was discovered in.*
II3: *The format that the sequences were in.*

We need II3 in order to process the sequence data correctly and determine II2, because every format is different.

II4: *The database from which the sequences were taken.*

We need II4 in order to know II3, as the format depends on the database.

II5: *The connection between one instance of II4 and one instance of II1.*

Finally, in order to determine the database from which sequences came over multiple experiments, we need to determine which output from Filter corresponds to which database query, i.e. they are part of the same experiment.

Provenance Question 2

In order to answer PQ2, we need to know information items II3, II4, II6, II7 and II8.

II6: *The sequences output from BLAST.*
II7: *The species of those sequences.*

In addition, as above, we need to know II3, to interpret the sequences and extract II7, and II4 to determine II3.

II8: *The connection between one instance of II6 and one instance of II4.*

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

Finally, in order to determine the database from which sequences came over multiple experiments, we need to determine which output from BLAST corresponds to which database query, i.e. they are part of the same experiment.

Provenance Question 3

In order to answer PQ3, we need to know information items II7, II8 and II3.

II9: *The default database name used in the workflow.*
II10: *The output of the User Interaction.*
II11: *The connection between an instance of II9 and an instance of II10.*

In order to determine that the default database name used in one experiment was changed (or not), we need to know the explicit connection between them.

All Provenance Questions

Finally, because each of the provenance questions will be asked with regard to one or more experiment results, for each of the above information items we need to know the following:

II12: *The connection between each instance of the above information items and the result of the same experiment as it occurred in.*

Actors

For the purposes of asserting and recording process documentation, we may model the application at different levels of granularity. For instance, Figure 1 shows a coarse level of granularity with only 6 interacting actors. Figure 2 shows a finer granularity for one of the high-level actors, the Workflow Enactor, in Figure 1, where the workflow is broken down into a further 5 actors. The level used for process documentation depends on what information we need to be recorded for the use cases.

If the information items are all apparent in the communication between actors in the coarse granularity model, then we could just use that model in process documentation. For instance, information item II6 is apparent in the communications between the Workflow Enactor and BLAST.

However, some information items, such as II1, are not made explicit in the coarse-grained model, so we have to use the finer-grained model and treat each step in the workflow as a separate actor.

Knowledgeable Actors

A *knowledgeable actor* is an actor in a process that has access to a required information item. The *primary knowledgeable actor* for an information item is the one that first possesses that information. We determine the knowledgeable actors so that we can ensure the required information is recorded in process documentation.

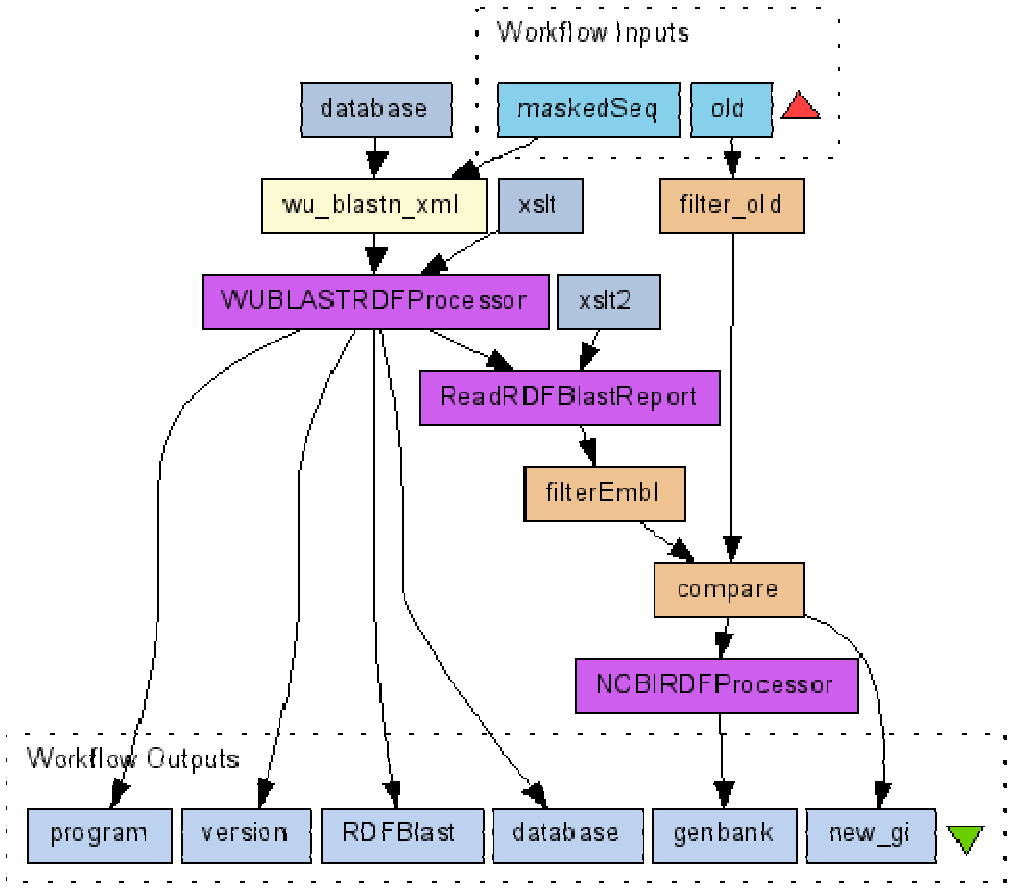
Some of the information items above are derived from others, and so the knowledgeable actor is not within the process: II2, II3 and II7. For the others we list the knowledgeable actors:

II1: Filter (in its output)
II4: User Interaction (in its input)
II5: User Interaction, BLAST, Filter and Workflow Enactor combined (in the relationships between their inputs and outputs and data flow in the workflow)

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

II6: BLAST (in its output)
II8: User Interaction, BLAST and Workflow Enactor combined (in the relationships between their inputs and outputs and data flow in the workflow)
II9: User Interaction (in its input)
II10: User Interaction (in its output)
II11: User Interaction (in the relationship between its inputs and outputs)
II12: User Interaction, BLAST, Filter and Workflow Enactor combined (in the relationship between its inputs and outputs and data flow in the workflow)

Appendix: Full Workflow



The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme