Provenance Management for Neuroimaging Workflows in neuGrid

Ashiq Anjum, Nik Bessis, Richard Hill
School of Computing and Mathematics
University of Derby, UK
Email: a.anjum@derby.ac.uk

Richard McClatchey, Irfan Habib, Kamran Soomro,
Peter Bloodsworth, Andrew Branson
Faculty of Environment and Technology, UWE Bristol, UK
Email: richard.mcclatchey@uwe.ac.uk

Abstract— **An increased amount of large scale, collaborative biomedical research has recently been conducted on e-Science infrastructures. Such research typically involves conducting comparative analysis on large amounts of data to search for biomarkers for diseases. Running these analysis manually can often be quite cumbersome, labour-intensive and error-prone. Significant work has been invested into automating such analysis with appropriately configured workflows. It is also important for biomedical researchers to validate analysis outcomes, to ensure the reproducibility of the results and to ascertain the ownership of specific scientific results. The detailed, traceable information required for this is often referred to as provenance data. Developing suitable methods and approaches to managing provenance data in large-scale distributed e-Science environments is another important area of research currently being investigated. We present an approach that has been adopted in the neuGRID project, which aims to develop an infrastructure to facilitate research into neurodegenerative disease studies such as Alzheimer's. To facilitate the automation of complex, large-scale analysis in neuGRID, we have adapted CRISTAL, a workflow and provenance tracking solution. The use of CRISTAL has provided a rich environment for neuroscientists to track and manage the evolution of both data and workflows in the neuGRID infrastructure**.

Keywords; Provenance, Workflows, Biomedical Analysis, Grid Computing, Neuroimaging

I. INTRODUCTION

It is common that scientific analysis on e-Science infrastructures consists of sequences of computations and data transformations that may process single or multiple, often large, data sets. Such scientific analysis is often based on scientific workflows [1]. As the use of collaborative distributed computing infrastructures for hosting this analysis becomes more common, many challenging issues are beginning to emerge. State of the art scientific workflows are increasing both in terms of the number of computations they perform and the size of the data they consume or produce [2]. Due to this increase in complexity, it is essential to ensure the reproducibility of analysis and also to confirm the correctness of the resulting outcomes [3]. Additionally, the knowledge required to author any scientific analysis must also be captured. In a collaborative research environment, where researchers use each others' results and methods, traceability of the data generated, stored and used must also be maintained. All these forms of knowledge are collectively referred to as forms of 'provenance' information.

In any analysis system where there are multiplicities of data-sets, and versions of workflows operating upon those data-sets, particularly when the analysis is carried out repetitively and/or in collaborative teams, it is imperative to retain a record of who did what, to which sets of data, on which dates, as well as recording the outcome(s) of the analysis. This information needs to be logged as records of particular users' analysis so that they can be reproduced or amended and repeated as part of a robust research process. All of this information, normally generated through the execution of scientific workflows, can be termed provenance data and it enables the traceability of the origins of data (and processes) and, perhaps more importantly, their evolution between different stages of their usage. This provenance data arises from the definition of candidate data sets, workflow activities, roles and actors, research outcomes and results sets and data derived from image analysis and other research processes. Capturing and managing this provenance data will enable users to query analysis information, automatically generate analysis workflows and to detect errors and exceptional behaviour in past analysis. This can then be utilised to validate analysis processes. Generally, in a scientific research infrastructure, multiple forms of provenance information are required; we shall explore these types in the following subsections.

Typically, data that is captured from different sensors or devices may need to be normalised and pre-processed before it can be analysed. The details of how the data has been transformed from the initial process which captured the raw data, to the final data product which makes it suitable for processing, is important in a scientific research infrastructure. For instance, in neuroimaging research different MRI scanners work at different resolutions, and different manufacturers can use different encoding schemes. Typically MRI brain scans are normalised into a single image format such as MINC [19] to assist data analysis. This normalisation process may result in the loss of some data. Similarly, due to confidentiality requirements, data sets must be pseudo-anonymised in biomedical research environments to ensure patient privacy. The pseudo-anonymisation processing may alter the image and the information related to this transformation of images must also be captured.

The availability of provenance information about a scientific analysis is regarded to be as important as the results of the scientific analysis itself [4]. Without provenance information, the correctness of an analysis cannot be clearly determined which could render the results to be scientifically

978-0-7695-4531-8/11 $26.00 © 2011 IEEE
DOI 10.1109/3PGCIC.2011.20

questionable. Here, provenance means the history, ownership and usage of data and its processing in some domain of interest For example, the tracking of engineering samples in the construction of aerospace engines, or the logging of data and process execution in the study of High Energy Physics (HEP) experiments [5]. The Large Hadron Collider at CERN will produce large data sets in the range of hundreds of terabytes, to be analysed by teams of researchers geographically distributed across the globe. To verify and subsequently interpret the results produced by the scientific analysis of all this HEP data, researchers require reliable provenance information [6]. Similarly, in order to assist research into neurodegenerative diseases such as Alzheimer's disease, researchers require scientific workflows (also termed pipelines) to process brain scans for various biomarkers. These biomarkers include the mean cortical thickness of a brain, the thinning of which has been linked to the onset of Alzheimer's disease. Researchers can track the progression of the disease by employing complex image analysis algorithms into neuroimaging scientific workflows. The knowledge acquired from executing these neuroimaging workflows must be validated using provenance information. The outputs from several rounds of analysis and the associated provenance information may be combined to provide a comprehensive picture of a diseased brain and thereby to determine a patient's likely prognosis.

In the health informatics community great emphasis has been placed upon the provision of suitable infrastructures to support biomedical researchers for the purposes, amongst others, of data capture, image analysis, and the processing of scientific workflows and the sharing of diagnoses. Many projects have reported on the customisation of Grid middleware such as gLite [7] and GRIA [8], and on the provision of access through portals to data distributed on the Grid between centres of biomedical research, for example, those studying degenerative brain diseases. Lately, neuroscience projects such as NeuroLog [9], NeuroGrid [10] and neuGRID [11] have considered services based on service-oriented architectures to facilitate the complex studies required to analyse Magnetic Resonance Imaging (MRI) and Computerised Tomography (CT) images. This may include querying and browsing data samples and specifying and executing workflows (or pipelines) of algorithms required for neurological image analysis. To date few have considered how such analysis can be tracked over time, between researchers and over varying data samples and analysis workflows.

In this article we outline the provenance management approach developed in the neuGRID project for the purposes of capturing and preserving the data that is collected in the specification and execution of (stages in) analysis workflows and in the definition and refinement of data samples used in studies of Alzheimer's disease (AD) [12]. The neuGRID project adopts a provenance tracking system for the purposes of tracking neurological analysis of AD called CRISTAL [13] to manage the construction of large-scale HEP detectors for the Large Hadron Collider. CRISTAL is a highly configurable system that was found to be suitable for neuGRID. Its design enables the rapid reconfiguration and adaption required to meet constantly evolving provenance requirements in biomedical research infrastructures. Existing state-of-the-art provenance management systems are not completely generic and reconfigurable. For instance, most workflow provenance management services are designed only for data-flow oriented workflows. CRISTAL, on the other hand, was initially designed for control-flow oriented workflows. Due to its extensible and reconfigurable nature it was ideally adapted to managing data-flow oriented workflows in the neuGrid project.

This article proceeds as follows: Section II summarises scientific workflows, provenance data management and the background for this in medical imaging studies, followed by a discussion of the domain requirements for workflow-generated provenance data capture, using a practical use-case from the neuGRID project. Later, in section III, we describe the provenance service developed in neuGRID to capture and manage provenance. Section IV presents the use of CRISTAL for provenance tracking in the neuGrid project. The paper concludes in section V with discussion of possible research directions for future.

II. Requirements Analysis: Workflow and Provenance Management in the neuGRID project

The neuGRID infrastructure is based on a set of generalised services that enables the European neuroscience community to carry out research that is necessary for the study of degenerative brain diseases. This set of generalised infrastructure services provides workflow management, provenance management, querying and a service that encapsulates the specifics of the underlying Grid infrastructure from more generalised services.

The user requirements gathering process in neuGRID involved working closely with the clinical researcher community. Meetings focused initially on the description of high-level scenarios and usage patterns that would later be used to cross check system functionality during final system testing. As these were produced a range of use-cases were created and then prioritised. This provided a framework upon which more detailed individual requirements could be based, and this has been of considerable benefit in terms of describing the project and ensuring that important components were not overlooked. This also led to a clear hierarchical conceptual framework being identified, that links high-level scenarios to more finely grained use-cases and to individual users' requirements. The primary focus of this work was on the production of easily understandable models that are meaningful to both clinical researchers and software developers.

In neuGRID a scenario (see Figure 1) was identified during requirements gathering that illustrates the valuable role that accurate workflow and provenance information can play in the research process. Consider the situation in which a given analysis workflow yields some surprising and potentially significant results. A researcher may wish to confirm that the results are accurate and to identify any errors that may have been made in their analysis. By analysing all the intermediary

image sets and workflow execution logs, the user would be able to manually verify that the results were correct or incorrect. It may be found that an error was due to a specific group of images interacting poorly within the workflow. The user annotates the workflow so that other users are warned if they attempt a similar analysis. A data provenance system would enable the capture of the workflow specification, the outcomes of each run of that workflow on a specified data sample to be gathered, together with the annotation provided by researchers on the execution of that workflow to be collected and managed. Consequently, the provenance database would begin to act as a shared knowledge base for the community of researchers.

Neuroscience relies heavily upon the use of statistical analysis techniques to process the output from a workflow and thereby test a given research hypothesis. A key factor in being able to draw meaningful conclusions from data is the size of the study sample. Generally speaking, the greater the size of the study set the more significance that can be given to the results that are produced. It may also be that a larger sample size will allow more precise questions to be asked which can lead to the discovery of new correlations between variables. A potential problem that arises when working with large numbers of scans is that the highly sensitive image processing algorithms may often fail in a proportion of cases. Such errors may have a significant impact on the analysis results and potentially allow incorrect conclusions to be reached. Such information may play an important role in the development of new treatments and the evaluation of the efficacy of experimental drugs and it is therefore important that errors are discovered before research outcomes are published. A provenance database can track the evolution of data samples as produced by the researchers and consequently robust data and process provenance is a high priority for researchers.

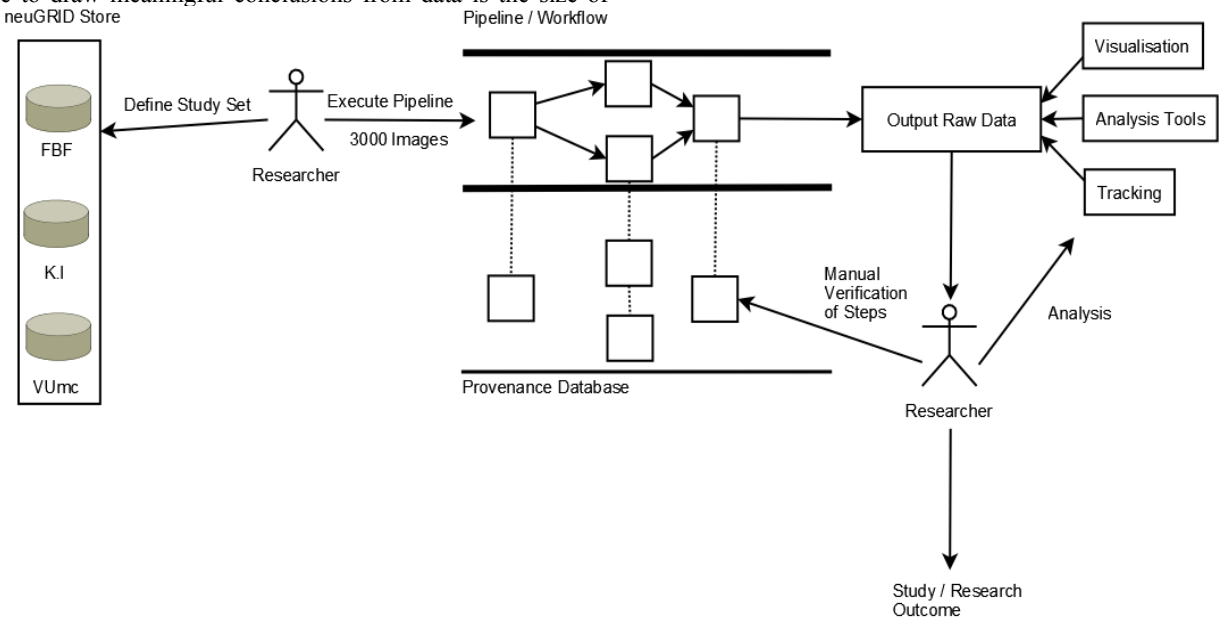


Figure 1 Validation of results using a provenance data scenario

III. PROVENANCE SERVICE: ARCHITECTURE AND COMPONENTS

As previously mentioned, the provenance management service developed in neuGRID is based on the CRISTAL system. CRISTAL captures provenance data that emerges in the specification and execution of the stages in analysis workflows. The provenance management service also keeps track of the origins of the data products generated in an analysis and their evolution between different stages of research analysis. Provenance querying facilities are provided for the so-called Provenance Service in neuGRID. Users can retrieve past analysis, retrieve specific versions of a workflow and examine the results of each individual computation and track ownership of workflows. A key ability of the CRISTAL system is its ability to adapt to changing requirements in terms of provenance storage. The domain of neuroscience is constantly changing as new workflows, algorithms and research studies are developed. The underlying CRISTAL model allows the system to evolve to handle such challenges whilst retaining provenance information in a consistent and traceable manner. Further details of CRISTAL are outlined in section IV of this paper. The Provenance Service captures the following information.

- Workflow specifications.;
- Data or inputs supplied to each workflow component;
- Annotations added to the workflow and individual workflow components;

- Links and dependencies between workflow components;
- Execution errors generated during analysis;
- Output produced by the workflow and each workflow component.

NeuGRID is an example of an infrastructure that has been designed to provide researchers with a shared set of facilities through which they can perform their research. At the heart of the platform is a distributed computation environment which was designed to efficiently handle the running of image processing workflows such as the cortical thickness-measuring algorithm, CIVET [14]. This is not enough on its own, however, as users require more than simply processing power. They need to be able to access a large distributed library of data and to search for a group of images with which they want to work. A set of common image processing workflows is also necessary within the infrastructure. A significant proportion of clinical research involves the development of customised workflows and image analysis techniques. The ability to edit existing scientific workflows and to construct new workflows using established tools is therefore important to researchers. Another vital aspect is the traceability of the analysis data that is produced using a workflow. Researchers need to be able to examine each stage in the processing of an analysis workflow in order to confirm that it is accurate. Overall users expect a well-tailored research infrastructure to support them in their research. Provenance information therefore plays a crucial role in achieving this by bringing together and storing information about how individual users have interacted with the underlying data infrastructure.

The set of neuGRID infrastructure services is depicted in Figure 2 and includes a Workflow Management Service, a so-called Gluing Service (or Infrastructure Abstraction Service), a Querying Service and the Provenance Service. The Workflow Management Service is a generic service that orchestrates the planning and enactment of workflows and manages the retrieval of provenance from executing workflows. Its design enables the specification of workflows in several formats which includes pipelines and basic scripts. The Workflow Management Service transforms these services into a neuGRID standard format, before orchestrating the enactment with the Gluing Service and subsequent retrieval of provenance from the Glueing Service to the Provenance Service.

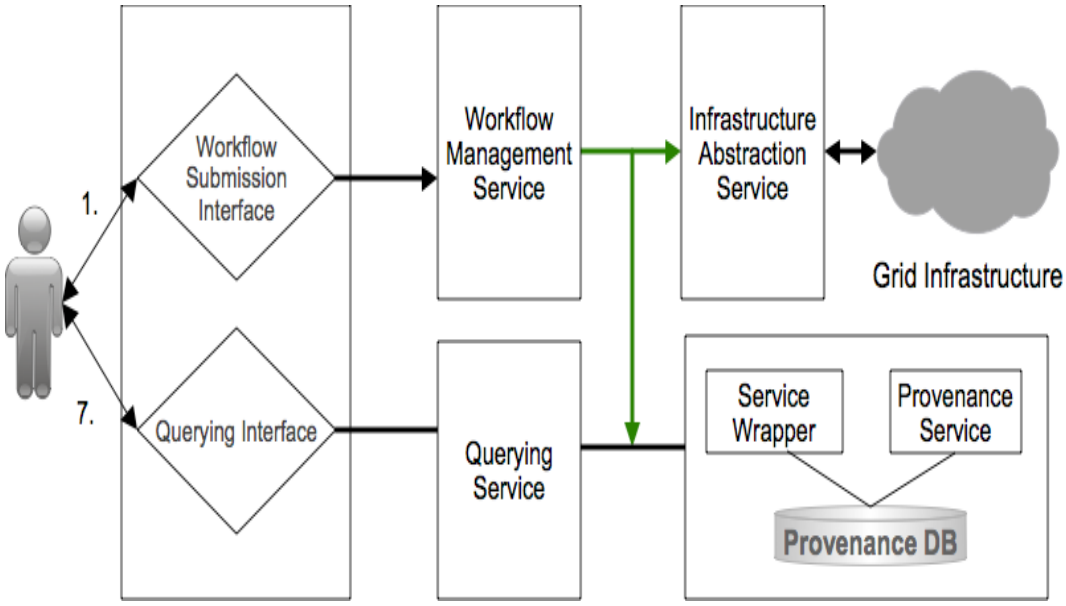


Figure 2: The architecture of a workflow-based research infrastructure.

The Gluing Service is a generic service that encapsulates the specific infrastructure's functionality. It shields the other services from the specifics of the infrastructure, allowing them to be developed against a standard set of features and APIs. It exposes methods to submit workflows to the infrastructure and to retrieve its status and provenance. It also exposes methods to upload and download files from the infrastructure, as well as to provide access to the authentication mechanisms. The Querying Service is designed to provide a querying gateway to users of the neuGRID infrastructure. It provides querying interfaces to the provenance repository as well as to other image and data stores within neuGRID. The Querying Service uses the querying API provided by the Provenance Service to query the provenance repository. The Provenance Service has

been developed to capture and manage provenance for the entire neuGRID infrastructure.

The Provenance Service has been designed based on the initial set of requirements described in section II. It consists of a web service, a CRISTAL core, and a relational database. The web service acts as a single point of entry to the Provenance Service's functionality in a distributed and heterogeneous manner. It consists of methods to store workflow definitions, to create workflow instances, to update workflow status and to store workflow provenance. It also provides methods to query the stored provenance. CRISTAL is the main provenance storage repository in the Provenance Service. It manages and orchestrates the execution of the stored workflows as well as the collection of their provenance information. The relational database provides a rich querying interface to the stored provenance.

The Provenance Service works by creating a virtual instance of the workflow within the service. Once created, the virtual instance must be kept synchronised with the actual workflow execution. Therefore, to store provenance in the Provenance Service, a client goes through the following steps:

1. Store workflow definition.
2. Create workflow instance.
3. Update workflow status.
4. Store workflow provenance.

Once these steps have been completed, the stored provenance can be queried from the Provenance Service.

IV. CRISTAL AS A PROVENANCE MANAGEMENT PLATFORM

CRISTAL is a data and workflow tracking (and provenance management) system. It is a process modelling and product data capture tool that addresses the harmonisation of processes by the use of the so-called CRISTAL Kernel software, so that potentially multiple heterogeneous processes can be integrated with each other and have their workflows tracked in the database. Using the facilities for description and dynamic modification in CRISTAL in a generic and reusable manner, CRISTAL is able to provide dynamically modifiable and reconfigurable workflows. It uses the so-called description-driven nature of CRISTAL models to act dynamically on process instances already running, and can thus intervene in the actual process instances during execution. These processes can be dynamically (re)-configured based on the context of execution without compiling, stopping or starting the process and the user can make modifications directly and graphically upon any process parameter, whilst preserving all historical versions so they can run alongside the new version. In the neuGRID Provenance Service, we have used CRISTAL to provide the provenance needed to support neuroscience analysis and to track individualised analysis definitions and usage patterns, thereby creating a practical knowledge base for neuroscience researchers. This section describes how CRISTAL fits into the overall neuGRID architecture for capturing and coordinating provenance data.

As shown in Figure 3 the interaction starts with the authoring of a workflow, which the user wants to execute on the Grid (shown as step (1)). Authoring can be carried out via several tools, the prototype being implemented in neuGRID using the LONI Workflow [15] and Kepler [16] as examples of authoring environments. The workflow management service as created for neuGRID translates the workflow specification into a standard neuGRID format (step (2) in Figure 2). The translated workflow, as shown in figure 3, is forwarded to the CRISTAL enabled Provenance Service, which then creates an internal representation of this workflow and stores the workflow specification into its schema [17].

The workflow specification is enriched by the inclusion of provenance actors for provenance collection. The Pipeline Service translates the workflow specification into a standard format and plans the workflow. The planned workflow, as shown in figure 3, is forwarded to the CRISTAL enabled provenance service, which then creates an internal representation of this workflow and stores the workflow specification in its schema (step (3)). This schema has sufficient information to track the workflow during subsequent phases of a workflow execution. The workflow activity is represented as a tree-like structure (a directed acyclic graph) and all associated dependencies, parameters, and environment details are represented in this tree.
Once the workflow specification is stored in CRISTAL, the Pipeline Service creates an instance of it. This initialises a virtual copy of the workflow within CRISTAL that records the execution of the actual workflow (steps (4) and (5)). As the workflow executes on the grid infrastructure, the workflow management service updates the status of the virtual copy within CRISTAL, with the status information retrieved from the grid through the Gluing Service. Any intermediate outputs generated by the activities within the workflow are uploaded to the grid. The locations of these intermediate outputs are retrieved by the Pipeline Service and passed along with the status updates to CRISTAL (step (6)). Once the execution of the workflow has completed, the provenance of the workflow is ready to be queried by users for analysis.

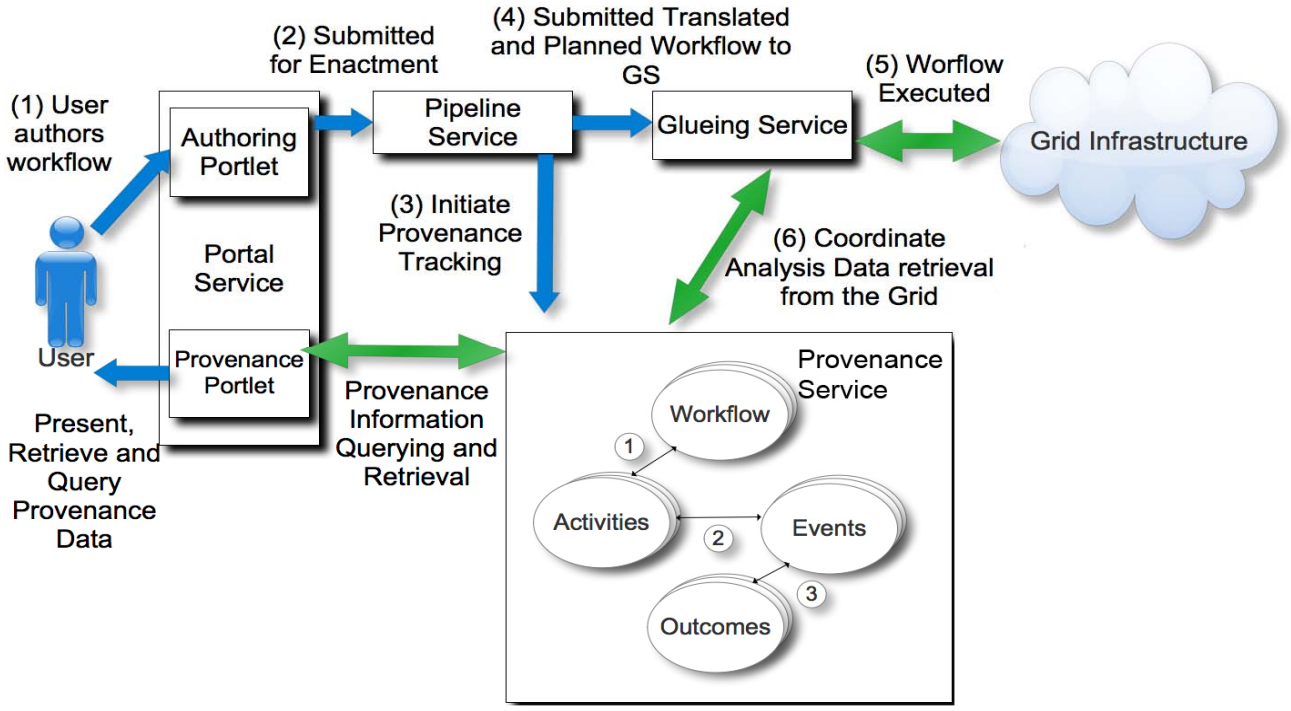


Figure 3: The Provenance Service in neuGRID

The separation between description and instance imposed by CRISTAL is instrumental in enabling workflow reuse. Re-enacting a workflow in the future becomes as simple as creating a new instance of a previously stored workflow specification. The workflow specification can also be retrieved, modified and executed by a user. In this case, the modified workflow is stored as a new version of the previous workflow specification in CRISTAL, maintaining a link between them. In this manner, CRISTAL keeps track of the evolution of a workflow. The provenance data is stored as a provenance graph; the nodes of the graph are atomic data transformation operations similar to the ones used by Taverna [18]. Nodes and edges in the provenance graph are tuples of the form:

1) $xform([X_1 : \tau_1 / x_1 ... X_n : \tau_n / x_n], Y_j : \sigma_j / y_j, P/p])$ (1)

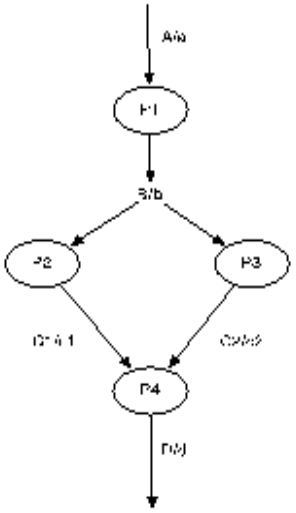
and

2) $xfer(X : \tau/x, Y : \tau/y)$

(1) records a transformation operation on some input variable X of type τ bound to value x. The input is to process a P bound to a process instance p. The output of p is the variable Y bound to a value y of type σ. These types of records are the nodes of the provenance graph. (2) records a transfer of value x of output X to input Y through a datalink. These records form the edges of the provenance graphs. The example workflow in figure 4 shows the provenance trace of a sample workflow execution.

In the above example, the trace can be used to infer the dependencies of each data product. We have kept the *xfer* relations implicit for simplicity. In the Provenance Service, these tuples are mapped and stored in a relational database. The Provenance Service employs a 2-pass translation mechanism. In the first pass, the workflow is mined for information about the each activity such as *TaskName*, *Executable*, *Priority* etc. In the second pass, the CRISTAL workflow is constructed using information mined during the first pass. The following rules are followed for each activity when constructing the CRISTAL workflow:

- All activities are mapped using one-to-one mapping into the CRISTAL workflow.
- If current activity has multiple successor activities, a succeeding AND Split is inserted into the CRISTAL workflow.
- If current activity has multiple predecessor activities, a preceding JOIN (3) inserted into the CRISTAL workflow.
- If current activity has multiple successor activities, all successor activities are connected to the succeeding AND Split of the current activity in the CRISTAL workflow.
- If current activity has multiple predecessor activities, all predecessor activities are connected to the preceding JOIN of the current activity in the CRISTAL workflow.



$$xform(A/a,B_1/b_1,P_1)$$

$$xform(B/b,C_1/c_1,P_2)$$

$$xform(B/b,C_2/c_2,P_3)$$

$$xform([C_1/c_1,C_2/c_2],D/d,P_4)$$

Figure 4: Example workflow trace

Once the workflow has been created, two additional steps are performed:

- If workflow has multiple starting activities, an AND Split preceding all starting activities is inserted in the CRISTAL workflow.
- If workflow has multiple ending activities, a JOIN succeeding all ending activities is inserted in the CRISTAL workflow.

These two steps are required for creating a correct CRISTAL workflow. Once a workflow execution starts in the neuGRID infrastructure, a parallel workflow simulation is created within CRISTAL. This allows clients to send incremental updates to the Provenance Service. The virtual workflow within CRISTAL simulates the actual execution of the workflow on the grid infrastructure. Adapting CRISTAL for the Provenance Service involved creating the appropriate Item descriptions and factories within CRISTAL.

V. Conclusions and Future Directions

In this paper we have outlined the approach that has been developed in the neuGRID project to use provenance management for the purposes of capturing and preserving the provenance data that emerges in the specification and execution of (stages in) analysis workflows, and in the definition and refinement of data samples used in studies of Alzheimer's disease (AD). In the neuGRID project a so-called Provenance Service has been designed and implemented that is primarily intended to capture the workflow information needed to populate a project-wide provenance database from the execution of scientific workflows. The Provenance Service can keep track of the origins of the data and its evolution between different stages of research analysis. The Provenance Service can allow users to query analysis information, to regenerate analysis workflows, to detect errors and any unusual behaviour in prior analysis, and to validate analysis. The Provenance Service has been based on the CRISTAL software [17], which is a data and workflow tracking system. CRISTAL is a process modelling and provenance capture tool that addresses the harmonisation of processes by the use of a kernel, so that potentially multiple heterogeneous processes can be integrated with each other, and have their workflows tracked in the database. Using the facilities for description and dynamic modification in CRISTAL in a generic and reusable manner, the Provenance Service is able to provide modifiable and reconfigurable workflows for a wide variety of Health applications. The Provenance Service also has pluggable data storage to store and retrieve provenance information and can be extended, to support a particular provenance database, by replacing its default storage mechanism.

In the long run we intend to research and develop a User Analysis module. This will enable applications to learn from their past executions and improve and optimise new studies and processes based on the previous experiences and results. Using machine learning approaches, models can be formulated that can derive the best possible optimisation strategies by learning from the past execution of experiments and processes. These models will evolve over time and will facilitate decision support in designing, building and running the future processes and workflows in a domain. A provenance analysis mechanism will be built on top of the data that has been captured in the Provenance Service. It will employ approaches to learn from the data that has been produced, find common

patterns and models, classify and reason from the information accumulated and present it to the system in an intuitive way. This information will be delivered to users while they work on new processes or workflows and will be an important source for their future decision-making.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support of the Framework Programme 7 of the EC through the Grant Agreement number 211714. In addition they thank the partners in neuGRID for their contributions: to this paper from Fatebenefratelli (Brescia, Italy), UWE (Bristol, UK), Maat Gknowledge (Archamps, France), VUmc (Amsterdam, Netherlands), HealthGrid (Clermont, France), Prodema (Bronschhofen), Switzerland), CFconsulting (Milan, Italy) and the Karolinska Institute (Stockholm, Sweden).

VI. REFERENCES

[1] I. J. Taylor et al., Workflows for e-Science, Springer-Verlag London Limited, 2007.

[2] Y. Gil et al., "Examining the challenges of scientific workflows," Computer, vol. 40, pp. 24-32, 2007.

[3] S. Miles et al., "Connecting scientific data to scientific experiments with provenance," IEEE International Conference on e-Science and Grid Computing, IEEE, pp. 179-186. 2007.

[4] S. Miles et al., "Provenance: The bridge between experiments and data," Computing in Science & Engineering, vol. 10, pp. 38-46, 2008.

[5] CERN's Large Hadron Collider, http://lhc.web.cern.ch/lhc/, Last accessed on 7/11/2010.

[6] A. Dolgert et al., "Provenance in high-energy physics workflows," Computing in Science & Engineering, vol. 10, pp. 22-29, 2008.

[7] A. Kretsis, P. Kokkinos, and E. Varvarigos, "Developing Scheduling Policies in gLite Middleware," Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, pp. 20-27. 2009.

[8] M. Surridge et al., "Experiences with GRIA – Industrial Applications on a Web Services Grid," Proceedings of the First International Conference on e-Science and Grid Computing, vol. pp. 98–105, 2005.

[9] J. Montagnat et al., "NeuroLOG: a community-driven middleware design," HealthGrid 2008, Chicago, 2008.

[10] S. Joseph, "NeuroGrid: Semantically Routing Queries in Peer-to-Peer Networks," in Web Engineering and Peer-to-Peer Computing, vol. 2376 of Lecture Notes in Computer Science, pp. 202–214, Springer Berlin / Heidelberg, 2002. 10.1007/3-540-45745-3_18.

[11] The neuGRID Project, http://www.neugrid.eu, Last accessed on 7/11/2010.

[12] G. B. Frisoni, "Structural imaging in the clinical diagnosis of Alzheimer's disease: problems and tools," Journal of Neurology, Neurosurgery & Psychiatry, vol. 70, pp. 711-718, 2001.

[13] R. McClatchey et al., "A distributed workflow and product data management application for the construction of large scale scientific apparatus," NATO ASI series. Series F: Computer and systems sciences, vol. 164, pp. 18-34-XVII, 524, 1998.

[14] A. P. Zijdenbos, R. Forghani, and A. C. Evans, "Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis," IEEE Transactions on Medical Imaging, vol. 21, pp. 1280-1291, 2002.

[15] M. Pan, D. Rex, and A. Toga, "The LONI Workflow Processing Environment: Improvements for Neuroimaging Analysis Research," 11th Annual Meeting of the Organization for Human Brain Mapping, 2005.

[16] I. Altintas et al., "Kepler: an extensible system for design and execution of scientific workflows," Proceedings 16th International Conference on Scientific and Statistical Database Management, pp. 423-424. 2004.

[17] A. Branson et al., "Coping with Evolving Requirements – A Case Study in Designing for Change," Under review at the Journal of Software and Systems Modelling, special issue on Models and Evolution.

[18] P. Missier et al., "Data Lineage Model for Taverna Workflows with Lightweight Annotation Requirements," in vol. 5272 of Lecture Notes in Computer Science, pp. 17-30, Springer Berlin / Heidelberg, 2008.

[19] P. Neelin, D. MacDonal, and D. L. Collins, "The MINC File Format: From Bytes to Brains," Fourth International Conference on Functional Mapping of the Human Brain, Montreal, 1998.