

Providing Traceability for Neuroimaging Analyses

R. McClatchey, A. Branson, A. Anjum, P. Bloodsworth, I. Habib, K. Munir, J. Shamdasani and K. Soomro

Centre for Complex Cooperative Systems, CEMS Faculty, Univ. of the West of England,
Coldharbour Lane, Frenchay, Bristol BS16 1QY, United Kingdom

Telephone: +44 117 328 3761, FAX: +44 117 344 3155

Email: {Richard.McClatchey, Andrew.Branson, Ashiq.Anjum, Peter.Bloodsworth, Irfan.Habib,
Kamran.Munir, Jetendr.Shamdasani, Kamran.Soomro}@cern.ch

Correspondence to:

Professor Richard McClatchey
Centre for Complex Cooperative Systems
University of the West of England,
Frenchay, Bristol BS16 1QY, UK

Email: Richard.McClatchey@cern.ch

Abstract: With the increasingly digital nature of biomedical data and as the complexity of analyses in medical informatics increases, the need for accurate information capture, traceability and accessibility has become crucial to medical researchers in the pursuance of their research goals. Grid- or Cloud-based technologies, often based on so-called Service Oriented Architectures (SOA), are emerging as potential solutions for managing and collaborating distributed resources in the medical domain. Few examples exist, however, of successful implementations of Grid-enabled medical systems that provide the traceability or provenance of research data needed to facilitate complex analyses and even fewer, if any, have been deployed for evaluation in practice. Over the past decade, we have been working with mammographers, paediatricians and neuroscientists in three generations of projects to provide the data management and provenance services now required for 21st century medical informatics research. This paper proposes a software solution that provides the foundation for such support. It introduces the use of the CRISTAL software to provide provenance management as one of a number of services delivered on a SOA, deployed to manage neuroimaging projects that have been studying biomarkers in the identification of the onset of Alzheimer's disease. In the neuGRID project a provenance service has been designed and implemented that is intended to capture, store, retrieve and reconstruct the workflow information needed to facilitate users in conducting neuroimaging analyses. The software enables neuroscientists to track the evolution of workflows (or pipelines) and datasets. It tracks the outcomes of various analyses and provides provenance traceability throughout the lifecycle of their studies. The paper also comments on the suitability of such an 'analysis service' in the wider context of medical research, discusses the lessons learned in a live deployment and reflects on its application in other forms of medical informatics research.

1. Introduction and Background

The past decade has witnessed orders of magnitude increases in computing power and data storage capacity, giving birth to new applications that can handle increasingly complex data in large volumes. Similar increases in network speed and availability pave the way for applications distributed over the web, carrying the potential for better resource utilization and on-demand resource sharing for bringing meaningful insights from the volumes of data. Medical informatics is one of the areas in which these technological advances could bring significant benefit both for scientists' research study and clinicians' everyday work. With the arrival of the deluge of data and information that has resulted from the advances in the medical domain, medical informatics is faced with increasing problems of data analysis and particularly of data traceability or provenance in the analysis of those data.

Over the past two decades, Grid Computing has emerged as a powerful computing paradigm to support large-scale experiments in medical and other scientific domains. It is defined as the "flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources" [1]. The Grid and latterly the Cloud [2] have provided the infrastructures and platforms to address the research challenges in medical informatics (as examples see [3], [4], and [5]). Emphasis has now shifted from the development of such infrastructures, to the provision of services through which medical informaticians can access data and algorithms to facilitate their programmes of research. For example, consider computational neuroimaging tools which require huge computing resources and the increasing availability of large image datasets will further enhance this need. Many efforts have been directed at creating brain image repositories including the recent US Alzheimer Disease Neuroimaging Initiative (ADNI) [6]. Geographically distributed computing infrastructures have been launched with the goal of fostering shared resources and intensive data analysis to advance knowledge of neurodegenerative diseases. Numerous projects, such as NeuroLOG [7] and

Neurogrid [8], have been undertaken to focus on providing Grid infrastructures that support neuroimaging applications.

neuGRID [9] was an effort which targeted the limitations of existing neuroimaging based Grid infrastructures. It aimed to provide an infrastructure and a set of platform services that were designed to support and enhance research, necessary for the analysis of neuro-degenerative diseases. neuGRID was an EC-funded scheme, arising from the needs of the Alzheimer's disease imaging community, which allowed the collection and archiving of large amounts of imaging data paired with services, Grid-based algorithms and sufficiently powered computational resources. It is being followed up by the N4U (neuGRID for You [10]) project which will provide user-facing services, including provenance and querying services, to enable neuroimaging analyses to be performed using the data stored in the neuGRID infrastructure. The intended benefit of these projects is to enable, ultimately, the faster discovery of new disease markers that will be valuable for earlier diagnosis and development of innovative drugs. The traceability (or provenance) of data is essential in carrying out meaningful analyses of large scale neuroimaging datasets. Provenance typically means the history, ownership and usage of data in some domain of interest. In both the neuGRID and N4U projects the vital need for data provenance has been identified by the end-user research community. We have addressed this through the provision of a Provenance Service whose description is the main contribution of this article.

In this paper we firstly outline the infrastructures that support service-based neuroimaging analysis. Suitable architectures for supporting neuroscience analysis are considered and the infrastructure adopted in the neuGRID and N4U projects is described. We then investigate the needs for data traceability that emerges in the specification and execution of (stages in) neuroimaging analysis pipelines and in the definition and refinement of data samples used in studies of Alzheimer's disease; this section also introduces the neuGRID/N4U Provenance Service. The next section describes the use of a system called CRISTAL, as the basis of the provenance service. The use of CRISTAL is evaluated as a practical use case in the penultimate

section of the paper and we draw lessons on its use. The paper concludes with discussion of future research including how a provenance repository could act as a knowledge resource for providing guided assistance to clinical researchers in biomedical analyses.

2. Infrastructure and Architecture for Neuroscience Analyses

The design philosophy that underpins both neuGRID and N4U is one that embodies the principles of reuse, flexibility and expandability. A layered and service-oriented approach has been followed in order to deliver against this philosophy. This approach enables a separation of concerns between the details of the application services (brain imaging) and the Grid deployment infrastructure. Different services have been delivered to satisfy the specific requirements of neuroscientists but have been designed and implemented to be flexible in nature and reusable in application. As shown in Figure 1, a set of generic services ‘glues’ a wide range of user applications to available Grid platforms, resulting in a system that addresses specific applications but retains a large degree of underlying generality, thus being able to cope with the still rapidly changing Grid environment.

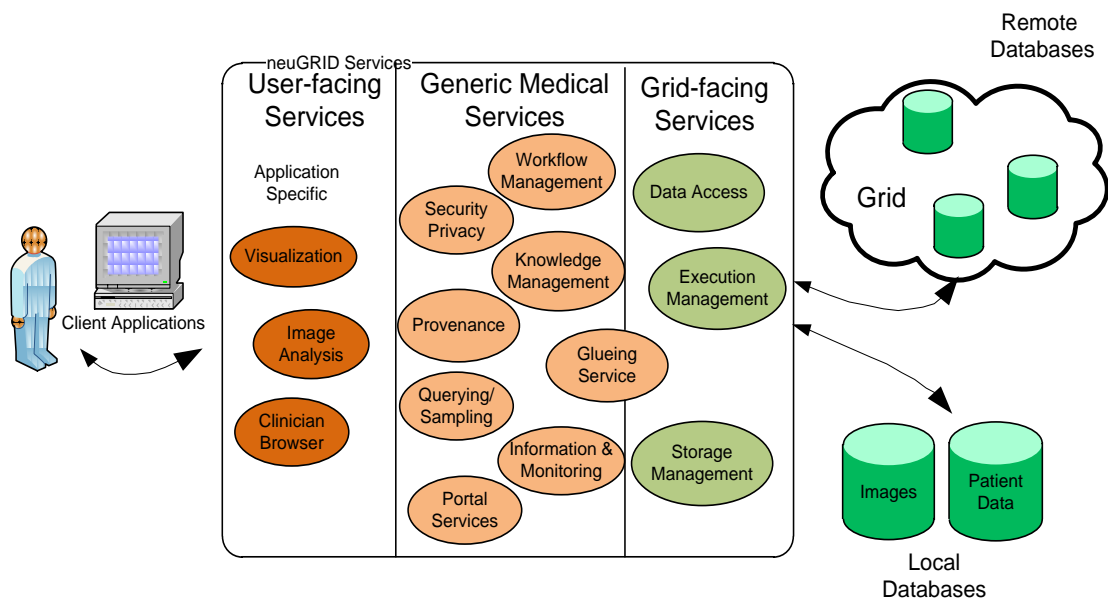


Figure 1: Layered Services Architecture in neuGRID

There is general consensus that Service Oriented Architectures [11] provide a suitable basis for supporting Grid or Cloud-based services required by the community of researchers. In computing, a service-oriented architecture (SOA) represents a method in which a loose coupling of functions between operating systems, programming languages and applications can be achieved; a SOA separates functions into distinct units, called services [12]. The MammoGrid project proposed perhaps the first healthcare example of a Grid-based service-oriented architecture [13]. The MammoGrid philosophy was to provide a set of generic services that were independent both of the front-end clinician-facing software and of the back-end Grid-facing software. In other words, a three-layer service architecture was proposed and implemented where the clinician was protected and isolated from the choice of Grid technologies.

The aim of the neuGRID project was to provide a user-friendly e-infrastructure and generalised services to enable the European neuroscience community in carrying out research that is necessary for the study of degenerative brain diseases. neuGRID is an example of an infrastructure that is designed to provide researchers with a shared set of facilities through which they can carry out their research. At the heart of the platform is a distributed computation environment that is designed to efficiently handle the running of complex image processing workflows such as the CIVET algorithm [14] (that enables longitudinal measurement of the thinning of the brain cortex). It is not enough on its own however, to provide support for neuroimaging analysis, since users require more than simple 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 will work. A set of common image processing workflows is also necessary within the infrastructure for users to work with. A significant proportion of clinical research involves the development of new workflows and image analysis techniques. The ability to edit existing, and to construct new workflows using established tools is therefore important to researchers. Researchers need to be able to examine each stage in the processing of an analysis workflow in

order to confirm that it is accurate. neuGRID has followed the MammoGrid philosophy in developing a platform based on a SOA [15] in order to enhance its flexibility and interoperability and to promote re-usability, potentially across other medical applications. This architecture (see Figure 1) allows neuGRID, and now N4U, to be implemented in such a way that its users do not require any advanced Grid know-how. This will be a great benefit since it has been shown that users often find it difficult to cope with the inherent complexities of Grid infrastructures. These include setting, configuring and maintaining the infrastructure. Performing these tasks requires in-depth technical know-how that most neuroscientists usually do not possess.

By abstracting Grid middleware specific considerations and customizations from clinical research applications, the neuGRID generic services provide functionality aimed at medical applications. Lower-level services hide the peculiarities of any specific Grid technology from the upper service layers thereby providing application independence and enabling the selection of 'fit-for-purpose' infrastructures. The generic services glue a wide-range of user applications to the available Grid platforms creating a foundation of cross-community and cross-platform services. The generic medical services are not tied to a particular application or Grid middleware; they could be used in any application domain and can be deployed potentially on any Grid infrastructure. These services are designed in such a way that a variety of applications and Grid middleware can be supported.

3. Data Traceability and the neuGRID Provenance Service

In clinical research environments analyses or tasks can be expressed in the form of workflows (elsewhere called pipelines). A scientific workflow is a step-wise formal specification of a scientific process; an example of a workflow is one which represents, streamlines, and automates the steps from dataset selection and integration, computation and analysis, to final data product presentation and visualization. A workflow management system supports the

specification, execution, re-run, and monitoring of scientific processes. As a consequence of the complexity of scientific (or biomedical) workflows, researchers require a means of tracking the execution of specified workflows to ensure that important analyses are accurately (and reproducibly) followed. Currently this is carried out manually and can often be error-prone. Errors may include, amongst others, incorrect workflow specifications, inappropriate links between pipeline components or execution failures because of the dynamic nature of the resources. A real challenge in this scenario is tracking the faults as and when they occur, due to the absence of a data and information tracking mechanism during the workflow specification, distribution and execution. This may in turn lead to a loss of user control or repetition of errors during subsequent analyses. Users may be prevented from being able to:

- Reconstruct a past workflow or parts of it to view the errors at the time of specification.
- Validate a workflow specification against a reference specification.
- View the intermediary results produced in the execution of a workflow to determine that those results are valid.
- Validate overall workflow execution results against a reference dataset.
- Query information of interest from past analyses.
- Compare different analyses.
- Search annotations associated with a pipeline or its components for future reference.

The benefit of managing specified workflows over time is that they can be refined and evolved by users collaboratively and can ultimately reach a level of maturity and dependability (the so-called 'gold standard'). Users need to collect information about (versions of) workflow specifications that may have been gathered from multiple users together with whatever results or outcomes were generated and then to use this so-called 'provenance data' as the drivers for improved decision making. This provenance data may, thereby over time, become a valued source of acquired knowledge as the nature of the analyses evolve; the provenance data store will essentially become a knowledge base for researchers. In future users must be able to invoke services to automatically monitor and analyze provenance databases to return statistical results

that match some criteria as set by the end user. This should provide efficient and dependable problem solving functionality and decision support system for the researchers.

In neuGRID a Provenance Service has been designed and implemented that is primarily intended to capture the information needed to populate a project-wide provenance database. As described below the service will support and enable the refinement of the workflows in the neuGRID project by capturing (see Figure 2):

1. Workflow specifications.
2. Data or inputs supplied to each workflow component.
3. Annotations added to the workflow and individual workflow components.
4. Links and dependencies between workflow components.
5. Execution errors generated during analysis.
6. Output produced by the workflow and each workflow component.

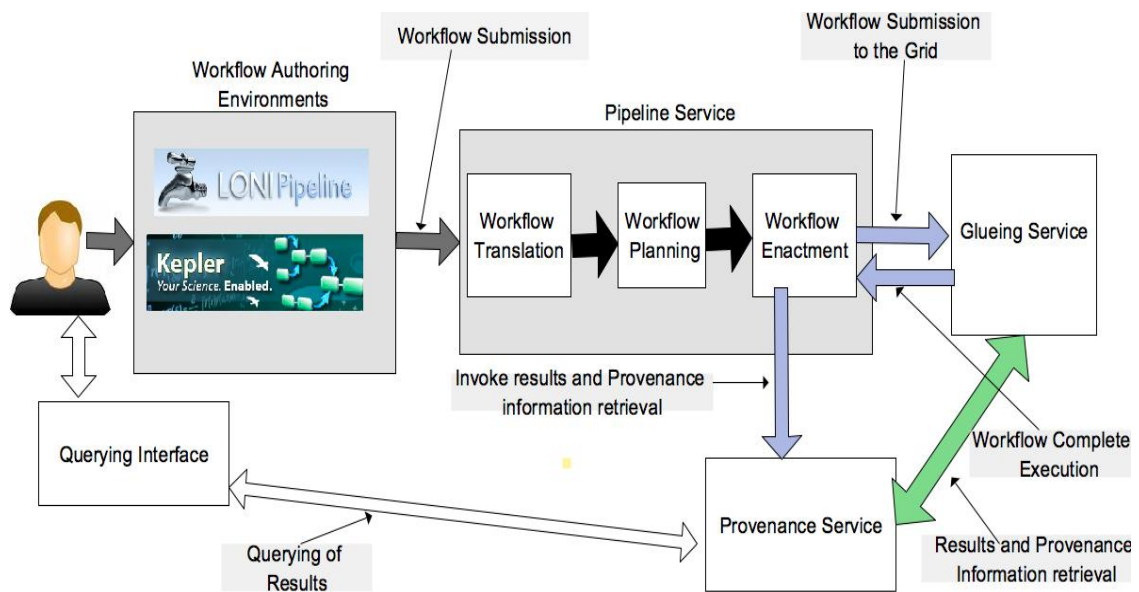


Figure 2: Provenance Service and neuroimaging analysis

The neuGRID infrastructure supports interoperable, reusable, extensible and scalable services.

The Provenance Service is one such service that traces, stores and provides access to analysis data for improved decision making. It is primarily intended to capture workflow specifications,

final and intermediate datasets and meta-data produced during an analysis and information pertaining to the environment where the data was produced (compute node, architecture etc.)

The neuGRID Provenance Service allows users to query analysis information, to regenerate analysis workflows, to detect errors and unusual behaviour in past analyses and to validate analyses. It assists users by providing them with access to past executions or histories of their analyses. This may allow them to learn lessons regarding a particular analysis which can be applied before a new study is initiated. It also gives users a means of capturing and maintaining workflow specifications and execution information in its provenance database. After the execution of a workflow all the information that was initially provided and that which was generated during an analysis is stored in the provenance database. The service thereby supports the continuous fine-tuning and refinement of pipelines in the neuGRID project. The provenance database can be queried by the user to verify results or improve and fine-tune pipelines and acts as a rich knowledge base of accumulated analysis steps and outcomes for users to consult.

The neuGRID Provenance Service has adapted a system called CRISTAL [16] that has been developed by the authors to manage the construction of large-scale High-Energy Physics (HEP) detectors for the Large Hadron Collider (LHC) at CERN, Geneva, for the purposes of tracking neurological analyses of Alzheimer's disease. The next section outlines how CRISTAL has been adapted to provide provenance tracking functionality for medical analysis.

4. CRISTAL as the Basis of a Provenance Service in neuGRID and N4U

Designing a system to provide data management, workflow tracking and change management to an agreed set of user requirements is a challenging task and there have been many previous projects aimed at this. CRISTAL is a distributed data and workflow management system which uses a generic and extendable storage repository, a multi-layered architecture for its component abstraction and dynamic object modelling for the design of its objects and components. These techniques are critical in handling the complexity of data and workflow traceability in distributed systems and to provide the flexibility to adapt to the changing analysis scenarios

typical of any research production system. CRISTAL has been based on a so-called description-driven approach in which all logic and data structures are “described” by meta-data, which can be modified and versioned online as the description of the object, component, item or an application changes. A Description-Driven System (DDS) architecture, as advocated previously in [17] is an example of a reflective meta-layer architecture (see Figure 3).

DDSs make use of meta-objects to store diverse domain-specific system descriptions (such as items, processes, lifecycles, goals, agents and outcomes) which control and manage the lifecycles of instances or domain objects that is the essential objects of a particular domain. In neuroimaging these objects might be, for example, raw image datasets, pipelines, derived datasets or analysis outcomes etc. As objects, reified system descriptions of DDSs can be organized into libraries conforming with frameworks for modelling of languages in general, and to their adaptation for specific domains.

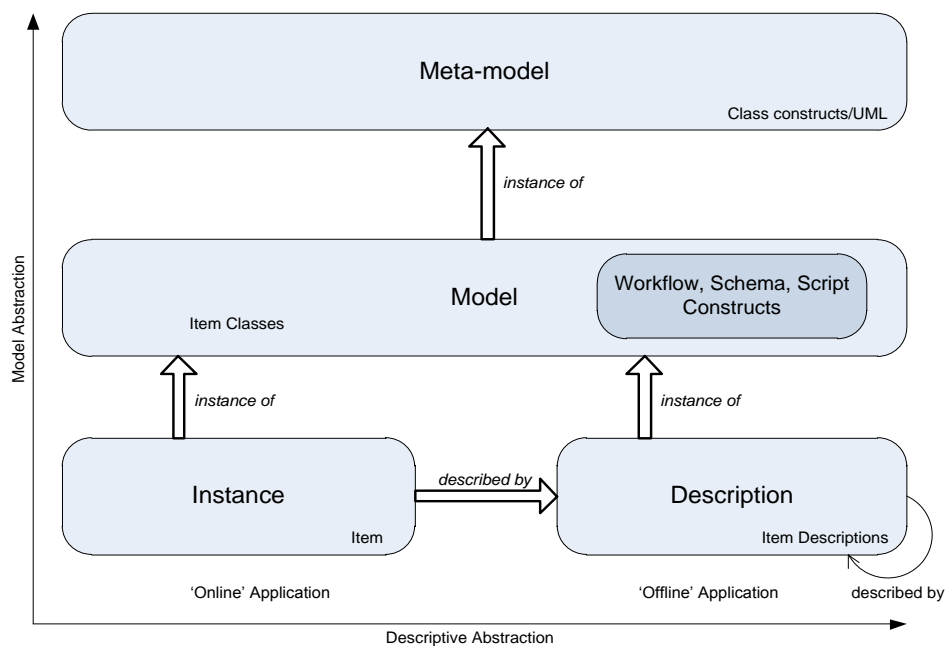


Figure 3: Model vs. Description in CRISTAL

The meta-data along with the instantiated elements of data are stored in the database and the evolution of the design is tracked by versioning the changes in the meta-data over time. Thus

DDSs make use of meta-objects to store domain-specific system descriptions that control and manage the lifecycles of domain objects. The separation of descriptions from their instances allows specification and management to evolve independently and asynchronously. This separation is essential in handling the complexity issues facing many web-based computing applications and facilitates interoperability, reusability and system evolution. Separating descriptions from their instantiation allows new versions of defined objects (and in turn their descriptions) to coexist with older versions. Neuroimaging is constantly developing new algorithms and workflows which may require variations to the provenance data that is collected. At the same time provenance data to be useful needs to remain consistent over time, to be traceable, queryable and easily accessible and scientists' analyses need to be conducted on those data. CRISTAL handles all of this. The reader is directed to previous publications on DDS for further background ([17], [18]). CRISTAL [19] is essentially a provenance tracking system which has previously been used to track the construction of large-scale experiments e.g. the CMS project [20] at CERN. It is both a process modelling and provenance capture tool which addresses the harmonisation of processes so that multiple potentially heterogeneous processes can be integrated and have their workflows tracked in the CRISTAL provenance database.

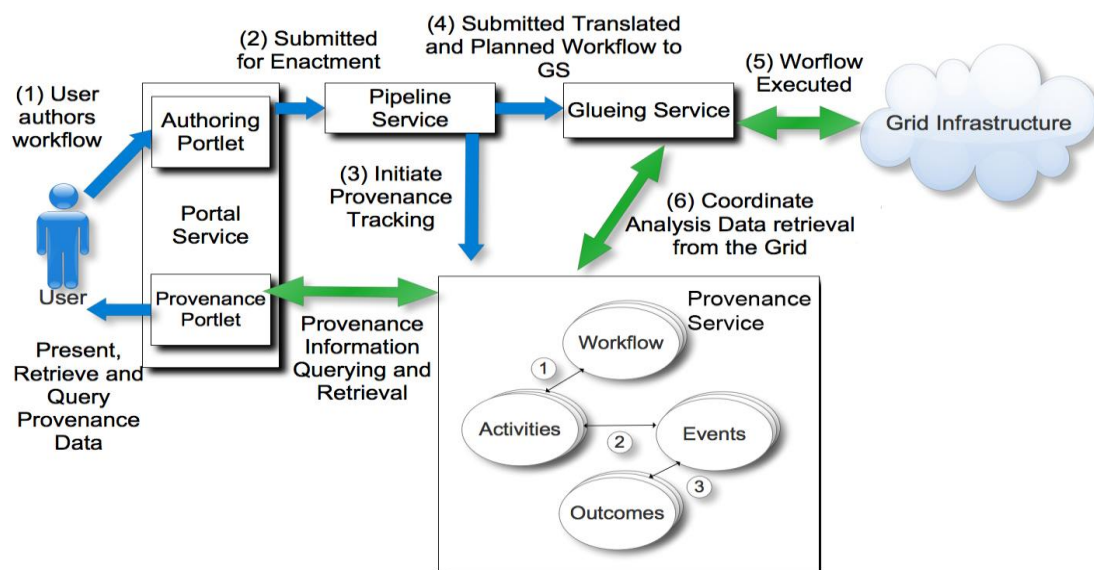


Figure 4: The Provenance Service in neuGRID

Using its facilities for description and dynamic modification in a generic and reusable manner, CRISTAL is able to support modifiable and reconfigurable workflows. It uses the description-driven nature of the CRISTAL models [17] to act dynamically on process instances already running and can thus intervene in the actual process instances during execution. Figure 4 shows how the CRISTAL system captures and tracks workflow provenance. User interaction starts with the authoring of a pipeline, which the user wants to execute on datasets held on the Grid. The workflow specification is enriched by including provenance actors for provenance collection. In neuGRID the so-called Pipeline Service translates the workflow specification into a standard format and plans the workflow. Another service, called the Glueing Service, facilitates the linking (and importantly the isolation) of services such as the Pipeline Service and Provenance Service to/from the underlying enabling (Grid) infrastructure. The reader should consult [21] for more information on these services.

The planned workflow, as depicted in Figure 4, 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. This schema has sufficient information to track the workflow during subsequent phases of the workflow execution. The workflow activity is represented as a directed acyclic graph; all associated dependencies, parameters, and environment details are represented in this graph. The schema also provides support to track the workflow evolution and the descriptions of derived workflows and its constituent parts are related to the original workflow activity.

Once a workflow has been enacted in the Grid, the Provenance Service coordinates the retrieval of all final data outcomes as well as intermediate data that was produced during the lifetime of the workflow. CRISTAL populates the appropriate structures to enable provenance tracking. In the CRISTAL model, a workflow consists of a number of activities, these activities being associated to multiple events and each event is associated with an outcome. An activity event may denote that the activity has failed and its outcome may include associated error logs, while the outcome of a successful activity may be the data produced during the runtime of the job. The

adopted model enables the pervasive tracking of the entire life-cycle of a neuroimaging workflow, from the pre-planned workflow to the final data outcomes.

5. The neuGRID Provenance Service in Practice

A prototype of the Provenance Service was developed in the neuGRID project and evaluated using a functional testing scheme that was based on a user defined case study. It has been further refined in the N4U project to ensure that it satisfies the tracking as detailed in the neuroscience research community's requirements [22]. The Provenance Service was deployed in the final stages of neuGRID and evaluated against the users' requirements. The results of the evaluation will be considered later on in this section. It satisfied the need for independent analysis traceability via CRISTAL. This provided the ability for users to define, run and share the output of user analysis between neuroimaging researchers for the first time. The Provenance Service consisted of two layers; an API layer and the CRISTAL layer, as shown in Figure 5.

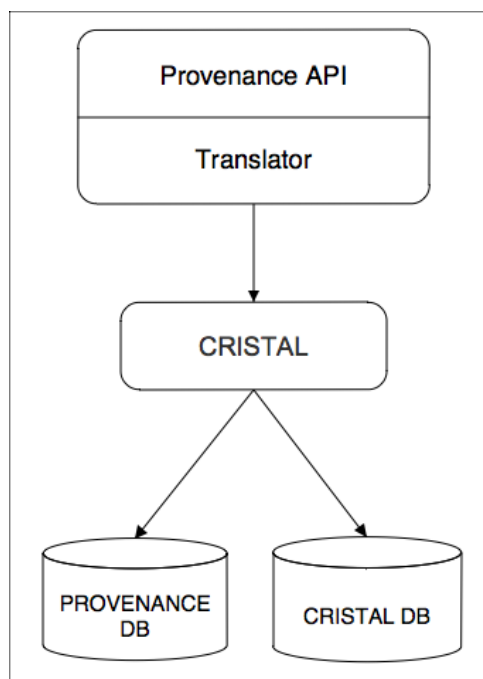


Figure 5: Provenance Service architecture

The API layer implements the Web Service API that serves as entry point into the Provenance Service. It was implemented using Apache Axis 1.4 (<http://ws.apache.org/axis/>) and Tomcat

6.2.20 (<http://tomcat.apache.org/>). This layer also consists of a Translator component. The API allows clients to store workflow templates, create workflow instances and update workflow instance status.

The Translator is responsible for converting the workflow passed to the Provenance Service in a standard format into CRISTAL's internal format. It employs a two-pass translation mechanism. In the first pass, the workflow is mined for information about 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 CRISTAL layer is the main tracking element in the Provenance Service. Once a workflow execution starts in the neuGRID infrastructure, a parallel workflow simulation is created within CRISTAL. This allows clients (typically the Pipeline Service) 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. As shown in Figure 5, CRISTAL uses two databases; the "CRISTAL DB" stores only the CRISTAL internal model and "Provenance DB" stores workflows' provenance. Other neuGRID services (e.g. querying service) directly interact with the Provenance DB for the execution of end-users' queries.

6. Evaluation of the neuGRID Provenance Service

In order to evaluate the neuGRID Provenance Service practically, an end-to-end use case scenario was developed in conjunction with end users. The community of end users included medically trained clinical researchers, statisticians, PhD students and imaging experts both from within the neuGRID project and externally. Figure 6 shows the use case which spans a complete analysis cycle in neuGRID from initial data collection, through analysis workflow execution to collaborative data analysis. By carrying out test cases which were based on this use case, the neuGRID platform has been thoroughly tested and evaluated by end users. The early stages of

the use case initialise the platform prior to an analysis being carried out. The first action in the use case, shown on the left in Figure 6 (the progression in time is from left to right), was to register images in the neuGRID store that have been collected from the hospital data acquisition systems or have been imported from other research projects.

Existing data was put through a process that enforces quality control, formatting and ethical compliance (plus anonymisation if required). The data was then integrated with the neuGRID standard data model, which enabled other researchers to access it and carry out their research. As new data sets were acquired they were put through a local quality control step before passing through the standard system-wide quality control, formatting, ethical compliance and data model integration processes. The registered data was tracked by the Provenance Service which recorded its creation and stored the definitions of the datasets for subsequent tracking of their usage in analyses.

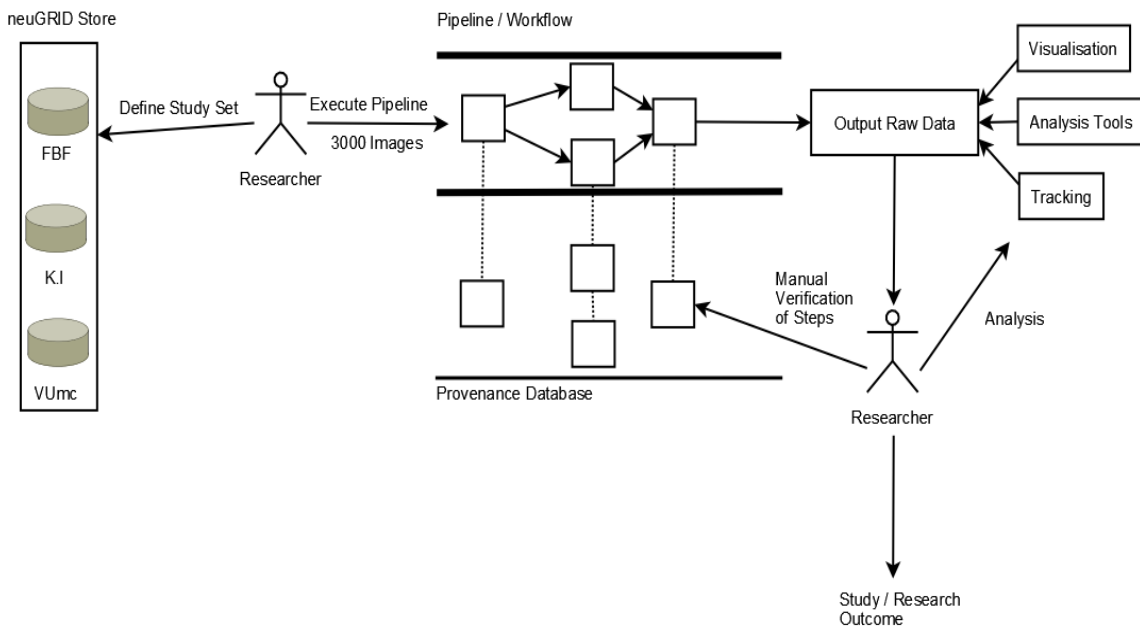


Figure 6: An end-to-end example of the neuGRID analysis environment

When the data had been registered, the next step in the analysis process was to make the data browsable through automated querying tools. The user interacts with the system using the neuGRID data store, to search for and to identify an appropriately large set of images from a

group of hospitals that match the required criteria. When the data has been imported into the neuGRID system and users are able to access and query these data, they can carry out studies with their corresponding data analyses to find results of interest (see Figure 6). The Provenance Service is made aware of the creation of analysis workflows so that their subsequent usage may be tracked. Users create workflows and then execute them on distributed resources provided by the Grid. For example, a researcher may wish to run a comparative analysis using a study set of MRI scans stored in geographically distributed medical centres. The user would typically interact with the system to choose a study set of perhaps as many as 3000 images (again this study is recorded in the provenance database), may select the CIVET [14] pipeline through which the analysis will take place and would start the analytical process. Users are not limited to using previously specified workflows and study samples; they can also construct new workflows. These workflows are also captured in the provenance database for later execution and monitoring on the Grid.

It is important that results, as and when required, should be reproduced and reconstructed using their full set of Provenance information. It may also be necessary to validate and, if required, to view the original workflow that has been used to obtain the results. For example, a user may create a new workflow and run it on a test data set. At each stage in the execution of the workflow, the intermediary images or data would be stored by the Provenance Service and a full audit trail of the data will be kept. After results have been produced, the user can examine the provenance database to check that each stage of the analysis has been completed correctly. The raw results can then be exported into the user's preferred analysis tool such as SPSS [23] and the whole process can be added to the researcher's history for future reference (Figure 6). Without the mechanism to validate workflows, it would be possible, but difficult, to correct the process and generate accurate results. Once a workflow has been developed and verified, a user is able to share it with other researchers.

Reflecting on the reasoning behind why certain user requirements were captured and described in the previous use case, can guide us in further evaluating the completed neuGRID

infrastructure and help us to assess the degree to which CRISTAL addressed the needs of researchers in terms of traceability. Clinical researchers develop research hypotheses which are then evaluated by teams of statisticians, computer scientists and associated experts. The collective effort that is necessary makes traceability a key feature of the research process. There is always a fear that a small error or even misunderstanding in one part of a highly complex analysis can render the final results useless, thereby wasting a large amount of time, money and potentially embarrassing the research team if they need to retract published work. In order to avoid such problems researchers need to follow established methodologies.

The standard scientific method requires that research results should be repeatable. This means that an independent team should be able to take an experiment that has been carried out elsewhere and repeat it in order to verify the results that were originally produced. In many scientific disciplines this is a relatively simple task, provided that methods are described with an appropriate level of detail. In the medical imaging domain however repeating experiments becomes much more complicated. Researchers would need to know details such as which images were used, their source, which pipelines were applied, which algorithms, algorithm versions, settings that were given and much more in order to be able to reproduce the analysis. The level of the traceability that the underlying CRISTAL provenance system provides in neuGRID can therefore be determined by comparing such requirements against what is actually captured by CRISTAL in terms of provenance data.

While designing the integration of CRISTAL as the basis of the neuGRID Provenance Service with the Pipeline and Glueing Services, it was decided that CRISTAL would not be used to orchestrate each and every step in the execution of each workflow. This was because it was believed that the round-trip times to receive the result of each step executed in the workflow from the Pipeline Service, to extract the next step from the stored workflow specification and to delegate it to the Pipeline Service would increase the already high workflow execution times. In addition the gLite V2.1 framework [24] used to provide the Grid platform already supported 'DAG jobs' (Directed Acyclic Graph based jobs) that facilitated workflow execution on the

Grid computing element itself. Furthermore it was believed that this will reduce the dependency of the low level services (such as workflow execution) on CRISTAL, allowing those services to run workflows without the Provenance Service if a particular deployment did not require it. In this instance CRISTAL simply tracked the workflow, replicating its state when it received notifications from the Pipeline Service about remote workflow events.

During the experimentation and evaluation of neuGRID, it became clear that CRISTAL should in fact be used as the orchestrator for each and every workflow step. CRISTAL could then control the execution of the workflow by acting on any changes in state during the execution as notified by the Pipeline Service and could instruct it as to the next workflow step. In the absence of such adequate pushing of workflow state changes to the Pipeline Service, events could be missed and provenance lost. In practice, without such controlled orchestration, it proved impossible to replay the correct state changes on the CRISTAL workflow to keep it in line with the workflow execution on the Grid. Moreover when the support for DAG jobs was removed from gLite (in its upgrade from V2.1 to V2.2) the ability of the Pipeline Service to run independent workflows on the Grid became unsupported. It would have required a major rewrite of the Pipeline Service itself in order to replace the functionality that was removed. If control of workflow orchestration had resided within CRISTAL, manifested through the Pipeline and Provenance services, then neither of these flaws would have made any difference to their operation. Unfortunately, once these drawbacks became apparent, it was too late in the project to make the dramatic changes to the architecture required to take control of orchestration and the project was forced to retain gLite V2.1 as its Grid middleware.

The resulting lesson learned from neuGRID was that provenance and execution control need to be tightly coupled to avoid any problems of infrastructure evolution later in the project lifecycle. Consequently, for a provenance system to be able to provide the level of traceability and control required to enable researchers to have full analysis tracking, the system must have all elements of analysis provenance data pushed to it. In other words the provenance system must be

informed of all information generated through the execution of the analysis workflows. If fragments of the workflow need to be farmed out remotely for execution, then a robust mechanism of delivering the execution progress must be in place to make sure that the relevant provenance data is returned.

The neuGRID project concentrated on the management, orchestration and infrastructure aspects of provenance tracking rather on the user-facing services to facilitate usability. Consequently users found the interface to the Provenance Service non-intuitive and with only infrequent access, difficult to use. In the N4U project we address this by the provision of a so-called Analysis Service in which users can define their personalised analyses visualised through a so-called Virtual Laboratory (VL) interface. The VL offers scientists access to a wide range of datasets, algorithm applications, access to computational resources, services, and provenance support (see [10]). The VL comprises a 'Dashboard' to present the underlying system, a set of integrated data resources, a set of services enabling access to the neuGRID infrastructure, and user analysis workbench to define/configure pipelines and inspect analysis output.

CRISTAL has been adapted in the Provenance Service to track the provenance of neuroimaging analysis in neuGRID. The important question that immediately comes to mind is how well the CRISTAL model can cope with tracking detailed provenance in the medical domain in general. In the neuGRID context, pipelines (workflows) provide a means of capturing processes and their associated metadata, which is a relatively common format to specify and build analyses in almost all types of medical applications. The underlying CRISTAL structure is flexible, this flexibility allows CRISTAL to store all of the information that users require in terms of traceability. CRISTAL can therefore capture and link all of the provenance data that was requested by users during the requirements analysis. This was validated through the feedback of users during infrastructure testing where the provenance capture mechanism was demonstrated to them. All of the processes outlined in the neuroimaging example from neuGRID and N4U are common across the domain of (bio) medical informatics. The need to capture analysis specifications, both in terms of the data (e.g. images) and the workflows (or pipelines) to be run

on those data are not specific to neuroimaging. Furthermore the requirement to collect provenance data as workflows are executed and to compile a history or audit of the analysis processes is also common practice especially with the need for repeatability and independent verification. Having worked with mammographers and paediatricians in the research domain it is clear that a general purpose Provenance Service, potentially delivered using CRISTAL, would be widely applicable and desirable to the community of (bio) medical informatics. This hypothesis will be tested in future applications of the neuGRID/N4U Provenance Service.

7. Future Directions and Conclusions

In this paper we have outlined the approach to provenance management that is being developed in the neuGRID and N4U projects. Such mechanisms are developed for the purpose of capturing and preserving the data that emerges in the specification and execution of (stages in) analysis pipelines and in the definition and refinement of data samples used in studies of Alzheimer's disease. neuGRID has built the foundations for exploitation of Grids in the neuroscience domain through the construction of an adaptable and extensible platform providing customisable, generalised services. The major benefit should be earlier diagnosis and faster development of new drugs, which will improve the quality of life of elderly people. The set of services has been designed and developed using the service oriented architecture paradigm. These services can be reusable both across Grid-based neurological data and later for wider medical analyses.

The neuGRID Provenance Service captures the workflow information needed to populate a project-wide provenance database. It tracks the origins of the data and their evolution between different stages of research analyses allowing users to query analysis information, to regenerate analysis workflows, to detect errors and unusual behaviour in past analyses and, finally, to validate analyses. The use of CRISTAL as the Provenance Service database and engine for neuGRID has enabled neuroscientists to support their complex image analyses over time and to collaborate together in teams with fully versioned workflows and datasets.

In the context of provenance data management, the CRISTAL data model is currently being adapted for compliance with the emerging standard Open Provenance Model [25] in the so-called Virtual Laboratory environment of the follow-up project to neuGRID called N4U (neuGRID for You see [10]). In N4U we are researching and developing an Analysis Service using CRISTAL. The analysis service will develop and transform neuGRID into a Virtual Laboratory (N4U) capable of meeting the requirements of the vast majority of scientists working in the field of imaging of neurodegenerative diseases, white matter diseases, and psychiatric diseases. The N4U Virtual Laboratory will provide the environment for users to conduct their analyses on sets of images and associated clinical data derived from a collection of project-specific data. This will provide facilities for users to interact with the underlying set of N4U services. This Dashboard like service presents the underlying system with a standard user-defined look and feel and which can handle role-specific (novice, regular, advanced user) access to the services. This enables users to prototype their analyses by refining data selections and pipelines, by trying out simple tests and ultimately larger experiments and to visualize the results of those test and experiments. As the Dashboard is the first point of entry for the user to the underlying N4U services, the work area is the first point of experimentation that empowers the user to use the N4U services in an infrastructure. In this way we believe that CRISTAL could become an essential building block for future projects requiring data and analysis tracking with provenance management both in execution and design.

Acknowledgements

The authors wish to thank their institutes and the European Commission for their support and to acknowledge the contribution of the following neuGRID and N4U project members: Dr Giovanni Frisoni and Alberto Redolfi (Fatebenefratelli, Brescia), David Manset and Jerome Revillard (Maat GKnowledge, France), Professor Frederick Barkov and his team at VU Medical Centre (Amsterdam), members of HealthGrid (Clermont, France), Dr Lars-Olaf Wahlund and

Eva Orndahl (Karolinska Institute, Stockholm) and Alex Zijdenbos of ProdeMa Medical (Braunshofen, Switzerland).

References

1. Foster I, Kesselman C, Tuecke S. 2001, The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International J. Supercomputer Applications*, Vol 15 No 3.
2. Foster I, Zhao Y, Raicu I, Lu SD, 2008, Cloud Computing and Grid Computing 360-Degree Compared, in *Grid Computing Environments Workshop*, 2008. GCE '08, pp. 1-10.
3. Ellisman M. et al., 2005, BIRN: Biomedical Informatics Research Network *Stud Health Technol Inform*;112: 100-109 IOS Press ISBN ISBN 1-58603-510-x, ISSN 0926-9630.
4. CBRAIN Project, <http://cbrain.mcgill.ca/>. Last accessed 9th March 2012.
5. Benkner S et al., 2008, GEMSS – Grid Infrastructure for Medical Service Provision. In *Methods of Information In Medicine*, Vol 44 No.2, ISSN 0026-1270, Schattauer GMBH publishers.
6. Mueller SG, Weiner MW, Thal LJ et al., 2005, Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement*. 1(1), 55–66.
7. Montagnat J, Gaignard A, Lingrand D, Rojas Balderrama J, Collet P, Lahire P., 2008. NeuroLOG: a community-driven middleware design. *Proc of HealthGrid'08*, IOS Press, Chicago.
8. Joseph S, 2002, NeuroGrid: Semantically Routing Queries in Peer-to-Peer Networks, In *Proceedings of the International Workshop on Peer-to-Peer Computing (co-located with Networking 2002)*, Pisa, Italy,03;95:194-199 IOS Press ISBN 1-58603-347-6
9. Redolfi A et al., 2009, Grid infrastructures for computational neuroscience: the neuGRID example, *Future Neurology*, Volume 4, pp. 703-722(20). See: www.neugrid.eu. Last accessed 9th March 2012.
10. neuGRID for You (N4U). See: neugrid4you.eu. Last accessed 9th March 2012.
11. Erl T., 2005, *Service-oriented Architecture: Concepts, Technology, and Design*. Prentice Hall PTR.
12. Newcomer E, Lomow G., 2005, *Understanding SOA with Web Services*. Addison-Wesley Professional.
13. Amendolia SR, Estrella F, Hassan W, et al., 2004, MammoGrid: A Service Oriented Architecture based Medical Grid Application. *Grid and Cooperative Computing* 3251, 939-942.
14. Zijdenbos AP, Forghani R, Evans AC, 2002, Automatic 'pipeline' analysis of 3D MRI data for clinical trials: Application to multiple sclerosis, *IEEE Trans Med Imag* 21(10): 1280–1291
15. Hauer T, McClatchey R, et al., 2007, Experiences of engineering Grid-based medical software. *International Journal of Medical Informatics*. 76(8), 621-632.
16. McClatchey R, Le Goff J-M, Baker N, Harris W, Kovacs Z, 1998, A Distributed Workflow & Product Data Management Application for the Construction of Large Scale Scientific Apparatus NATO ASI Series F : Computer & Systems Sciences Vol 164 pp 18-34. ISBN 3-540-64411-3 Springer Verlag.
17. Estrella F et al., 2003, Pattern Reification as the Basis for Description-Driven Systems *Journal of Software and System Modelling* Vol 2 No 2, pp 108-119 Springer-Verlag, ISSN: 1619-1366.

18. Estrella F et al., 2001, Meta-Data Objects as the Basis for System Evolution. LNCS Vol 2118, pp 390-399 ISBN 3-540-42298-6 Springer-Verlag.
19. Branson A et al., 2012, Evolving Requirements: Model-Driven Design for Change. Under final review at Information Systems, Elsevier publishers.
20. Chatrchyan S et al., 2008, The CMS Experiment at the CERN LHC. The Journal of Instrumentation Vol: 3 Article No: S08004, Institute of Physics Publishers.
21. Anjum A et al., 2009, Reusable Services from the neuGRID Project for Grid-Based Health Applications. Studies in Health Technology & Informatics Vol 147, pp 283-288 ISBN 978-1-60750-027-8 IOS Press.
22. Anjum A et al., 2010, Research Traceability using Provenance Services for Biomedical Analysis. Studies in Health Technology & Informatics Vol 159, pp 88-99 ISBN 978-1-60750-582-2 IOS Press.
23. SPSS software. See: <http://www-01.ibm.com/software/uk/analytics/spss/products/statistics/> Last accessed 9th March 2012.
24. Kretsis A, Kokkinos P, and Varvarigos E 2009, Developing Scheduling Policies in gLite Middleware. Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, 2009, pp. 20-27
25. Open Provenance Model (OPM), <http://openprovenance.org/>. Last accessed 9th March 2012.