Provenance Support for Medical Research

Richard McClatchey[*], Jetendr Shamdasani, Andrew Branson and Kamran Munir

Centre for Complex Cooperative Systems, FET, Coldharbour Lane, UWE Bristol, UK
{richard.mcclatchey, jetendr.shamdasani,
andrew.branson, kamran.munir}@cern.ch

Abstract. This poster paper introduces a system known as CRISTAL [1] and the experience using it for medical research, primarily in the neuGRID [2] and neuGridforUsers (N4U) projects. These projects aim to provide detailed traceability for research analysis processes in the study of biomarkers for Alzheimer's disease. They have faced major challenges in managing data volumes and algorithm complexity leading to problems associated with information tracking, analysis reproducibility and scientific data verification. We present a working system that supports provenance data management for medical researchers.

Medical informatics has increasingly required systems that facilitate historical data capture and management in order to support researchers' analyses through workflow based algorithms. To facilitate the requirement of tracking large scale analyses, we have adopted CRISTAL [1], a workflow and 'provenance data' tracking solution. Its use has provided a rich environment for neuroscientists to track and manage the evolution of both data and workflows in neuGRID and N4U. In the N4U project in particular we have developed a so called Virtual Laboratory (VL). One major goal of the VL is to ensure the reproducibility of results and to allow sharing of analysis information between researchers. All of the workflows in N4U after design are automated, their complete history from design to orchestration being captured and stored. Another feature of the VL is its collaborative environment, allowing for 'provenance' information to be shared and used by various researchers. The N4U VL is based on services layered on top of the neuGRID infrastructure, described in detail in [2].

The VL was developed for neuroscientists involved in Alzheimer's studies but has been designed to be reusable across other medical research communities. It has been designed to provide access to infrastructure resident data and to enable the analyses required by the medical research community. This has been achieved by basing the N4U virtual laboratory on an integrated Analysis Base [3], which has been developed following the detail requirements from both neuGRID and N4U projects. This Analysis Base provides an integrated medical data analysis environment to exploit neuroscience workflows, large image datasets and algorithms for scientific analyses. Once researchers conduct their analyses information from the Analysis Base, the analysis definitions and resulting data along with the user profiles are also made available in the Base for tracking and reusability purposes in a so-called Analysis Service via a Science Gateway, Analysis Workarea and Information Services. (see Figure 1).

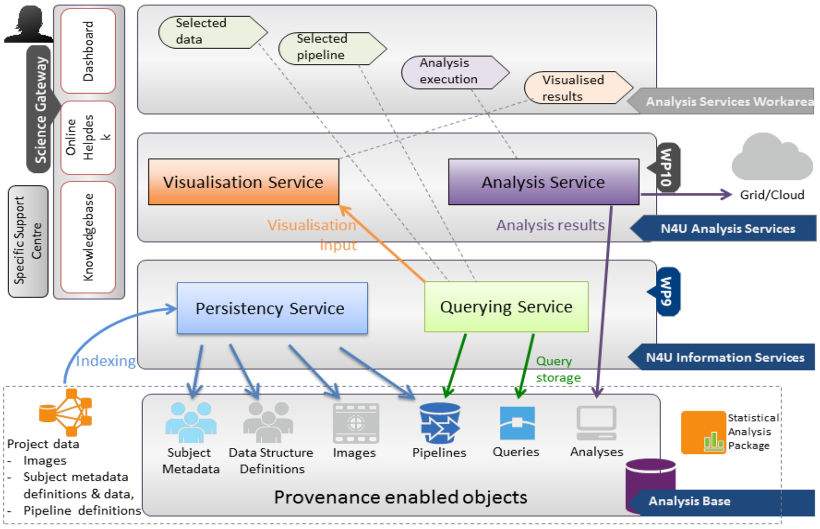

Figure 1 : The N4U Virtual Laboratory

The N4U Analysis Service provides access to tracked information (images, pipelines and analysis outcomes) for querying/browsing, visualization, pipeline authoring and execution. Its Work Area is a facility for users to define new pipelines or configure existing pipelines to be run against selected datasets and dispatch to conduct analysis. The N4U Science Gateway provides facilities that include a Dashboard, Online Help and Service interfaces for users to interact with the underlying set of N4U services. The N4U Analysis Base: (a) indexes all external clinical datasets (b) registers neuroscience pipeline definitions and/or associated algorithms (c) stores provenance and user-derived data resulting from pipeline executions on the Grid (d) provides access to all datasets stored on the infrastructure and (e) stores users' analysis definitions and linking them with the existing pipelines and datasets definitions.

CRISTAL is a data and workflow tracking system which was used to trace the construction of the CMS experiment at the CERN LHC [4]. 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 "description-driven" nature of CRISTAL models to act dynamically on process instances already running, and can intervene in the actual process instances during execution (for further detail refer to [1]). These processes can be dynamically (re)-configured based on the context of execution without compiling or stopping the process and the user can make modifications directly upon any process parameter whilst preserving all historical versions so they can run alongside the new version. In neuGRID/N4U, we have used CRISTAL to provide the provenance needed to support neuroscience analysis and to track individualized analysis definitions and usage patterns, thereby creating a practical knowledge base for neuroscience researchers.

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. CRISTAL is a system that records every change made to its objects, which are referred to as CRISTAL Items. Whenever a modification is made to any piece of data, the definition of that piece of data or application logic, the change and the metadata associated with that change (e.g. who made the change, when and for what purpose) are stored alongside that data. This makes CRISTAL applications fully traceable, and this data may be used to assemble detailed provenance information.

In N4U, CRISTAL manages data from the Analysis Service as Items, containing the full history of computing task execution; it can also provide this level of traceability for any piece of data in the system, such as the datasets, pipeline definitions and queries. Provenance querying facilities are provided by the Querying Service in neuGRID/N4U. The ability of description-driven systems to both cope with change and to provide traceability of such changes (i.e. the 'provenance' of the change) we see as one of the main contributions of the CRISTAL approach to building flexible and maintainable systems and we believe this makes a significant contribution to how enterprise systems can be implemented.

In the future we will develop a so-called User Analysis module which will enable applications to learn from their past executions and improve and optimize new studies and processes based on the previous experiences and results. Using machine learning approaches, models will be formulated that can derive the best possible optimisation strategies using 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 CRISTAL. 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. Work is also ongoing to make CRISTAL compliant with emerging provenance standards such as the Open Provenance Model, OPM [5].

References

1. Branson, A. et al., CRISTAL : A Practical Study in Designing Systems to Cope with Change. Information Systems journal. Elsevier Publishers Vol 42, pp 139–152 (2014).
2. Anjum, A. et al., Provenance Management for Neuroimaging Workflows in neuGRID, Intl Conf on P2P, Parallel, Grid, Cloud and Internet Computing. Barcelona, Spain (2011)
3. Munir, K. et al., An Integrated e-Science Analysis Base for Computational Neuroscience Experiments and Analysis. Procedia - Social & Behavioral Sciences. Vol 73 pp 85-92 (2013).
4. Chatrchyan, S et al., The Compact Muon Solenoid Experiment at the CERN LHC. The Compact Muon Solenoid Collaboration, The Journal Instrumentation Volume: 3 Article No: S08004, Institute of Physics Publishers (2008)
5. Shamdasani, J. et al., Towards Semantic Provenance in CRISTAL. Proc. of the 3rd International Workshop on the role of Semantic Web in Provenance Management (SWPM12) pp 29-36 ISBN: 978-1-4673-1328-5 Heraklion, Crete. IEEE Press May (2012).