Contribution ID: **37** Type: **not specified**

Data Provenance Tracking as the Basis for a Biomedical Virtual Research Environment

Tuesday, 7 March 2017 16:00 (30 minutes)

In complex data analyses it is increasingly important to capture information about the usage of data sets in addition to their preservation over time in order to ensure reproducibility of results, to verify the work of others and to ensure appropriate conditions data have been used for specific analyses. Scientific workflow based studies are beginning to realize the benefit of capturing this provenance [1] of their data and the activities used to process, transform and carry out studies on that data. This is especially true in biomedicine where the collection of data through experiment is costly and/or difficult to reproduce and where that data needs to be preserved over time. There is a clear requirement for systems to handle data over extended timescales with an emphasis on preserving the analysis procedures themselves and the environment in which the analyses were conducted alongside the processed data.

One way to support the development of workflows and their use in (collaborative) biomedical analyses is through the use of a Virtual Research Environment. However, the dynamic and geographically distributed nature of Grid/Cloud computing makes the capturing and processing of provenance information a major research challenge. In addition most workflow provenance management services are designed only for data-flow oriented workflows but researchers are now realising that tracking data alone is insufficient to support the scientific process [2]. What is required for collaborative research is traceable and reproducible provenance support in a Virtual Research Environment (VRE) that enables researchers to define their analyses in terms of the datasets and processes used, to monitor and visualize the outcome of their analyses and to log their results so that others users can call upon that acquired knowledge to support subsequent analyses.

We have extended the work carried out in the neuGRID and N4U projects in providing a Virtual laboratory [3] to provide the foundation for a generic VRE in which sets of biomedical data (images, laboratory test results, patient records, epidemiological analyses etc.) and the workflows (pipelines) used to process those data, together with their provenance data and results sets are captured in the CRISTAL software [4]. This paper outlines the functionality provided for a VRE by the Open Source CRISTAL software and examines how that can provide the foundations for a practice-based knowledge base for biomedicine and, potentially, for a wider research community.

References:
[1] Y. Simmhan et al., “A Survey of Data Provenance in e-Science”. In SIGMOD RECORD, Vol 34, P. 31-36. ACM, 2005.
[2] S. Bechhofer et al., “Why Linked Data is not Enough for Scientists”. Future Generation Computer Systems Vol 9 No. 2 pp 599-611, Elsevier Publishers, 2013.
[3] R. McClatchey et al, “Traceability and Provenance in Big Data Medical Systems”. Proc of CBMS 2015, Sao Carlos, Brazil.
[4] A. Branson et al., “CRISTAL : A Practical Study in Designing Systems to Cope with Change”. Information Systems 42, pp 139-152. Elsevier publishers.

Primary author: Prof. MCCLATCHEY, Richard (University of the West of England, Bristol UK)

Presenter: Prof. MCCLATCHEY, Richard (University of the West of England, Bristol UK)

Session Classification: VRE

Track Classification: Virtual Research Environment (including Middleware, tools, services, workflow, … etc.)