A Stacked-RuleSet Methodology for Provenance Management

M. Ayre, S. Woodman, C. Wise, N.J. Car
CSIRO
Email: Melanie.Ayre@csiro.au

Abstract: According to the W3C, provenance is "information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness". For modelling and simulation results to be accepted as a key part of decision-making processes, this isn't enough - the assumptions, data, and modelling processes need to be available and repeatable. Finally, the system needs to be useable and accessible. Correct data management practices need reliable tools with simple human interactions.

Here, we consider the tooling and design choices for a provenance information management architecture supporting NeCTAR virtual laboratories. Each such lab that wishes to manage provenance in the larger sense generates provenance reports by incorporating one of a number of toolkits we provide into its systems. This toolkit is used to generate reports which are sent to a single, central repository established for all the labs to use. Reports are logged with the lab identifier and any job identifier the lab allocates for lab/repository linking. Such reports rapidly become complex graphs with challenging validation, all of which is performed seamlessly by the tools to minimise the need to understand provenance which would be required to implement the methodology.

To ensure that each lab achieves the provenance standard it aims for, it is necessary to validate many layers - from the validity of the provenance report, such as ensuring each activity in the process ends after it starts, to the actual availability and integrity of the datasets - in a tailorable manner. The tools' validation ensures, at a minimum, that reports adhere to the W3C's PROV data model. Our stacked-rulesets approach allows any architecture user such as a lab to specify additional validation criteria. Reports on the status of validation are returned to the lab process for logging, and can be used during software design to ensure that valid reports are generated; however in operation, errors in providence reporting are not flagged to the user, but are stored alongside the report.

With the repository guaranteeing to provide access to validated, standardised, reports for each job run by each lab, sophisticated provenance data mining may take place to establish the required trustworthiness. Further, it enables holistic investigation into the scope of lab activities to assess coverage in experimentation and explore variation between operations. This approach minimises the inconvenience of data management to the user; while maximising the potential for a process to be truly repeatable.

The various provenance reporting toolkits, the repository system and the validation rulesets form a collection of Provenance Management System (PROMS) tools which can be used for systems other than the virtual laboratories. We conclude this paper by mentioning some in-development usage scenarios.

Keywords: *Provenance, virtual laboratories, reporting systems*