Inter-agency standardised provenance reporting in Australia

Nicholas Car[1]

[1]CSIRO Land & Water Flagship, Dutton Park, QLD, Australia, nicholas.car@csiro.au

INTRODUCTION

Scientific results must be reproducible in order for them to satisfy scientific peer review. In the early days of modern science, simple experiments could be described on paper and re-run elsewhere to verify results. Today, many scientific results are produced by long and complex chains of computer assisted data acquisition, processing and modelling. It is not feasible to record all the steps of such data lineage in scientific papers or even perhaps in the printed form at all.

Information platforms (IPs) in Australia such as eReefs [1] and the Bioregional Assessments Data Repository [2], present data from multiple Australian government agencies and other institutions for particular purposes. In eReef's case it is to ascertain water quality within the Great Barrier Reef lagoon. To ensure that the data they present is as scientifically sound, IPs need to deliver metadata for their datasets alongside the data itself them that helps users understand the data's lineage. Not only are eReefs' datasets complex but the fact that they come from multiple agencies further adds to the difficulty in implementing useful lineage models, formats and delivery systems for that dataset metadata.

The term provenance, when used in the context of scientific computing, refers to the lineage of data – what steps were taken in its production and what foundational data were input into those steps. Provenance research is relatively new but does have an international standard, PROV, [3] and emerging communities of practice both within Australia, such as Research Data Australia's Provenance Interest Group.

For eReefs and for Bioregional Assessments IPs, similar approaches to managing provenance data have been taken with the implementation of provenance stores and the use of PROV by both projects. These approaches ultimately see inter-agency standardized provenance reporting within Australia. This paper describes those approaches.

PROVENANCE SCIENTIFIC PROCESS MODELLING

In order to allow multiple, heterogeneous, systems to produce provenance metadata for the datasets they produce, a generic "scientific process provenance model" has been used. It specifies how to record the various inputs and outputs of scientific processes using the PROV standard and extensions to it [4]. Figure 1 shows a generic example of such.

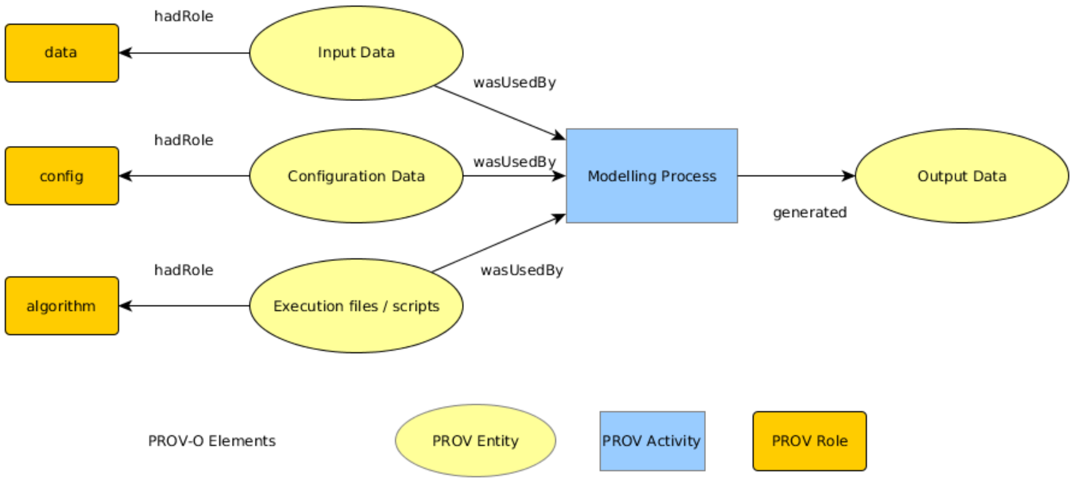

Figure 1: Characterisation of a generic scientific process using PROV

Using the characterization shown in Figure 1, an 'activity centric' view of data generation is recorded. This allows people to ask what source data and system configuration were used to generate derived data. In eReefs' case, the data provision processes characterized in this way range from CSIRO's Ocean Colour satellite data sets [5] to Queensland Government's

Catchment Modelling (see [6] for a description of previous iterations of this process). These are very different processes however they may all be represented by such a generic characterization.

The PROV standard guides what information is required for each element in Figure 1 and the relationships between them. In addition, conventions for use and data management-specific details that are not specifically part of the PROV standard, are implemented through adherence to the PROMS Ontology [7].

Using these tools, multiple process steps and multiple processes can be characterized and joined together and have their provenance recorded. Powerful querying languages exist [8] to traverse the provenance graphs then created that allow for metadata interrogation.

RECORDING AND DELIVERING PROVENANCE DATA

Data provision processes that have been characterized as described above need to have their manual and automated systems record the requisite provenance data. That data (metadata to the main data the processes are generating) then needs to be expressed in accordance with PROV/PROMS and stored.

Within eReefs, the author has worked with the owners of various eReefs data providers to implement "provenance exporter" modules that record the requisite data from the native systems and send it to a centralised provenance store. In both eReefs and the Bioregional Assessments programme, these centralised stored are implementations of the Provenance Management System server [9]. That system is a simple, massively scalable, document database able to store provenance *Reports* for each run of each process associated with its IP.

FURTHER USE

The principles and systems in use for eReefs and the Bioregional Assessments programme regarding provenance are able to be used for other IPs and non-IP systems within Australia. Since an open, international standard form of provenance is reported, longevity of access and usability is assured. The methods mentioned here work equally well with service-delivered data as with repository-housed data.

REFERENCES

1. The eReefs Data Interoperability and Visualisation Team (2014) eReefs: Information Platform. Website, online at http://data.ereefs.org.au, accessed 2014-06-02.
2. The Bioregional Assessments Data Management Team (2014) Bioregional Assessments Data Repository. Website, online at http://data.bioregionalassessments.gov.au, accessed 2014-06-02.
3. Lebo, T., Sahoo, S. and McGuinness D. (eds.) (2013) PROV-O: The PROV Ontology. W3C Recommendation April 2013. http://www.w3.org/TR/prov-o, accessed 2014-06-02.
4. Car, N.J. (2014) PROMS Scientific Process Modelling. Wiki web page, online at https://wiki.csiro.au/display/PROMS/PROMS+Scientific+Process+Modelling, accessed 2014-06-02.
5. King, Edward; Brando, Vittorio; Clementson, Lesley; Lovell, Jenny; Franklin, Heidi; Anstee, Janet; Besnard, Laurent (2013) Ocean Colour in Australia's Integrated Marine Observing System. Conference Paper – Poster for the International Ocean Colour Science Meeting 2013, Darmstadt, Germany, 6-8 May 2013. Online at https://publications.csiro.au/rpr/pub?list=ASE&pid=csiro:EP133226&expert=false&sb=RECENT&n=3&rpp=25&page=22&tr=2838&dr=all&csiro.affiliation=B3800, accessed 2014-06-02.
6. Queensland Government (2010) Great Barrier Reef Second Report Card 2010 Reef Water Quality Protection Plan Catchment pollutant loads methods. Online as a web page at http://www.reefplan.qld.gov.au/measuring-success/methods/catchment-pollutant-loads.aspx, accessed 2014-06-02.
7. Car, N.J. (2014) Provenance Management System Ontology. Web page and semantic web ontology online at http://promsns.org/def/proms, accessed 2014-06-02.
8. SPARQL, Wikipedia web page, online at http://en.wikipedia.org/wiki/SPARQL, accessed 2014-06-02.
9. Car, N.J. (2014) The Provenance Management System. Wiki web page, online at https://wiki.csiro.au/display/PROMS, accessed 2014-06-02.

ABOUT THE AUTHOR

Nicholas Car

Is an informatics researcher in CSIRO Land & Water's Environmental Information Systems team. His particular interests are in scientific data provenance and distributed data management systems. He has worked on provenance aspects for a number of high-profile scientific data management systems including the Bioregional Assessments program's data repository, eReefs and the Bureau of Meteorology's Geofabric hydrological spatial dataset. He is working towards widespread implementation of provenance standards and methodologies within Australia via Research Data Australia's Provenance Interest Group and internationally via the Belmont Forum and Research Data Alliance.