Driving Data Management cultural change via automated provenance management systems

Nicholas J. Car[a], Michael G. Hartcher[a] & Matthew P. Stenson[a]

[a] CSIRO Land & Water, Environmental Information Systems
Email: nicholas.car@csiro.au

Abstract: Large multi-disciplinary scientific projects that inform government policy and have a high public profile are often exposed to high levels of scrutiny. Such projects rely on a range of input datasets and modelling software packages and generate high volumes of output data, which are presented as summarised results in published reports. Defending the scientific integrity of project reporting requires that all project results have demonstrable integrity with clear evidence of the workflows and processes used to generate them, i.e. they must implement structured data management including provenance capture and storage.

Provenance data capture forms part of effective data management. The reporting of data provenance needs to occur in all workflows within a project and crucially needs support from project management, and adoption by project staff so that provenance chains are unbroken at every step, thus providing demonstrable integrity. Even when project funds and milestones are allocated to provenance tasks, such as ensuring staff store project datasets in managed locations and generate standardised dataset metadata records, data provenance capture has often been poor. This indicates that the barrier to the adoption of useful data provenance tasks is still significant. The development and application of automated systems, which capture and report provenance without additional user effort, are therefore of critical importance in helping to lower this barrier thus easing cultural change in data management.

Even if a project or organisation has motivation, has made the case, established a vision, and developed plans to implement provenance management, buy-in from all project staff is still required for success. This is because provenance chains containing information about data lifecycles need to be unbroken for all results, thus requiring involvement from all project staff. Some, perhaps the majority, of project processes cannot be automated, thus they will require significant manual effort in order to be included in provenance management.

This paper outlines previous best-practice regarding CSIRO's data management approach as demonstrated by the Murray Darling Basin Sustainable Yields project, and reflects on their shortcomings, such as the lack of adequate provenance capture, with improvements suggested. It then describes several automated provenance management tools that employ semantic web technologies and preserve the identity of provenance reports and datasets; which may be used to help with bottom-up practice adoption. The automated provenance management tools can provide well-defined, automated processes, which may help to lower the barriers preventing cultural change for data management at the project and organisational level.

It is hoped that the improved data management practices and the automated tools discussed here can inform current and new high-profile projects, such as the Bioregional Assessments program, to attain a higher quality of demonstrable data integrity through more robust provenance management.

Keywords: *Provenance, data management, cultural change, semantic web, metadata*

1. INTRODUCTION

In this introductory section we present a case study in data management best-practice, i.e. the Murray-Darling Basin Sustainable Yields (MDBSY) project, which is an example of a large multi-disciplinary scientific project that has informed government policy. We then detail initial and subsequent approaches taken by the Water Informatics Research and Development Alliance (WIRADA – between CSIRO and the Australian Bureau of Meteorology) regarding its automated workflows, which have allowed the testing of ideas regarding automated data management infrastructure and served as a test bed for some of the precursor tools and techniques introduced in this paper. In Section 2 we describe new technical approaches and in Section 3 new organisational and cultural approaches to provenance management stemming from our MDBSY and WIRADA experiences. In Section 4 we discuss the potential impact of the new approaches outlined in Sections 2, and how developments in automated provenance capture along with incentives, as outlined in section 3, can help drive further cultural change. Finally, in Section 5 we briefly conclude and mention our hopes for future projects.

1.1. Murray-Darling Basin Sustainable Yields

The Murray-Darling Basin Sustainable Yields (MDBSY) project commenced at CSIRO in 2007 and undertook a complete assessment of the Murray-Darling Basin's water availability. This large and complex project contained a great volume and diversity of data, models and reports, and data management (DM) was undertaken as a separate project component.

For each dataset, model output and report document, it was determined to be essential that knowledge of how it was produced and where it came from was preserved in order to provide a complete audit trail (Hartcher and Lemon, 2009). Protocols were established to cover data storage, archiving, data exchange with external agencies, shared project documents etc., which was overseen by a Data Management team consisting of a team leader, data manager, individual project data coordinators, and systems engineers developing a metadata catalogue tool.

A single project repository was created for all model outputs, base data and software versions and reports and a relational database was used to store and manage the metadata statements. This repository was structured with directories for each geographical region considered by the project, and some additional directories for project-wide work. Directory and file naming conventions were used to identify where versions of datasets were stored. Checks were applied to ensure certain standards were met on specific data types, such as spatial data layers, before they were migrated from working space into the appropriate project archive directories (Hartcher and Lemon, 2009).

The final DM initiative in the MDBSY project was the employment of a metadata catalogue into which any dataset that was used as an input to another dataset or model, needed to catalogue a metadata statement. Automatic generation of some metadata elements and delivery to the catalogue was facilitated by a software tool that was able to discover new data delivery to the repository given that the directories and files used naming conventions. Random checks on the existence and quality of manually generated metadata statements were carried out by DM staff. After metadata statements were loaded into the catalogue, queries were able to be run against it to identify audit trails links for all derived results (tables and charts) appearing in the final report publication.

A reporting database was designed to maintain a provenance trail for results and to assist project teams in delivering those results. Development of the reporting database was only partially completed during the project with some teams using a single data model for their work and generating early results (Hartcher and Lemon, 2009). The reporting database proved useful where it was applied but due to it not being completed within the project timeframes there wasn't a comprehensive test of its utility.

The MDBSY approach to data provenance management improved on previous CSIRO efforts regarding project data management, particularly with respect to:

- dedicating staff to the effort;
- establishing project protocols and processes to facilitate metadata capture, audit trail development, and reporting;
- building tools to help with metadata capture and reporting;
- establishing check procedures to ensure protocols were followed and systems used.

After the project's conclusion, the MDBSY data managers suggested that future projects build on methods they had developed with formal statements of data management responsibility for projects; mandatory project management processes; and a 'blue print' model for data archiving and audit trail development.

1.2. WIRADA Provenance Management – the Central Provenance Store

The Australian Hydrological Geospatial Fabric (Geofabric) and Australian Water Resources Assessments system (AWRA; Stenson et al. 2012) projects at the Australian Bureau of Meteorology (the BoM) implemented automated workflow processes that required DM. The goals were firstly to store provenance data identifying workflow input data and workflow structure so that one could handle queries about how particular workflow outputs were generated and secondly to be able to know enough about a workflow's configuration to recreate it. A Central Provenance Store (CPS) was built to hold provenance data from both projects' systems and a harvester was built to extract representations of processing execution from the Geofabric and AWRA workflow systems (Trident[1] and Delft-FEWS[2] respectively). The harvester used a series of system-specific queries to get the data it needed from the Geofabric's Trident databases and AWRA's FEWS log files and then reported them to a single database for long-term storage (Kloppers et. al., 2012). Search & visualization tools were built to assist with provenance trace analysis (Lee & Box, 2012).

The CPS stored provenance data in the Proof Markup Language (PML) format (da Silva et. al., 2004), which was created as an exchange language for use as part of the *semantic web*. PML has its origins in proof theory but has shown to be of use in exchanging provenance data (McGuinness *et. al.*, 2007).

In contrast to the MDBSY case above, only two systems used this approach, and both of them consisted of fully automated processes. Complex management of staff was not required and long data processing chains could be described precisely using automated reporting.

This system demonstrated two purpose-built harvesters, which extracted information from the internals of automated systems, and the use of a single provenance markup format, PML, which could be used regardless of the systems reporting provenance data. It also demonstrated the use of a central, purpose-built, provenance data repository, as opposed to the more common metadata repositories and showed some visualizations (deemed inappropriate for final use) of the stored provenance data. Fundamentally, this work showed that provenance data from multiple, heterogonous, automated, systems could be extracted, stored and accessed with little or no operator effort, i.e. that the provenance processes could be automated.

As far as the authors are aware, no validation tests were carried out on stored provenance data to ensure that workflows identical to the original could be recreated and/or rerun using the provenance reports.

1.3. WIRADA Provenance Management #2 – the Provenance Management System

From late 2012, the authors redesigned the CPS described in Section 1.2 in the WIRADA Geofabric 12/13 project (Bureau of Meteorology, 2013) for many reasons, some of which were:

1. A large part of the international provenance community had recently coalesced work around the PROV family candidate standards (Groth and Moreau, 2013) thus we wished to implement PROV, rather than the original WIRADA implementation of PML. This was due to the expected larger user and developer community around PROV;
2. It was thought that systems should be responsible for generating and reporting their provenance, rather than a single harvester component that had to access multiple systems. This would allow a simple provenance store design and place the onus for use on new systems wishing to report provenance, rather than the provenance system development team;
3. Experience with the Geofabric project had shown that it's hard to generate provenance data at the correct *granularity* even if all of the requisite data are available, thus an architecture should accept provenance data at multiple levels of granularity;
4. Some elements of provenance reports were doomed not to maintain value over time, most obviously file names for data inputs (and outputs) in PML. A better method of indicating which files were used, and what their contents had been, was required;

[1] *Project Trident: A Scientific Workflow Workbench* software built by Microsoft Research. See http://tridentworkflow.codeplex.com/.

[2] *Flood Early Warning System* software by Deltares. See https://publicwiki.deltares.nl/display/FEWSDOC/Home.

5. The CPS was not a generic, simple, scalable provenance management solution. It was complex to understand, complex to use and questions about its ability to extend to storing large volumes of data needed to be addressed.

The Provenance Management System (PROMS)[3] was designed in 2012/2013 to address these and other issues. Key design features of PROMS are the:

1. Ability to use free text, or one or many of several data model schemas in PROV, for provenance data storage. This allows a single system to capture provenance data over a range of *granularities* ;
2. Storage of files to capture input and output data and the optional storage of files to capture the executable parts of workflows as well as their referencing in provenance reports. This allows workflows to be rebuilt using identical copies of the original file, and means, importantly, that workflow artefacts can be trusted to have the same identity as those used in the original workflow;
3. Use of a simple, RESTful[4], Application Programming Interface (API) to access the PROMS system itself which displays both human and machine readable forms (HTML web pages and Resource Description Framework (RDF)[5] pages respectively). This allows people to initially understand what is required of provenance reporting by inspecting other reporting systems' reports;
4. Use of Uniform Resource Identifiers (URIs)[6] to universally and uniquely identify input & output data items and workflow instances, as opposed to file locations, timestamps and names;
5. Ability to use a range of file storage mechanisms (repositories & version control systems) thus allowing workflow builders to use tools already employed by projects for this purpose;
6. Ability to use any provenance reporting agent that can communicate via the well-known and widely supported HTTP protocol thus allowing as universal access as possible.

PROMS has been tested with a Trident Geofabric workflow using a provenance exporter component within the workflow (see Figure 1), as opposed to a harvester that runs externally, to report Trident workflow provenance. It has also been tested with other, human, workflows also used in the Geofabric project that report their provenance using a custom data entry web page.

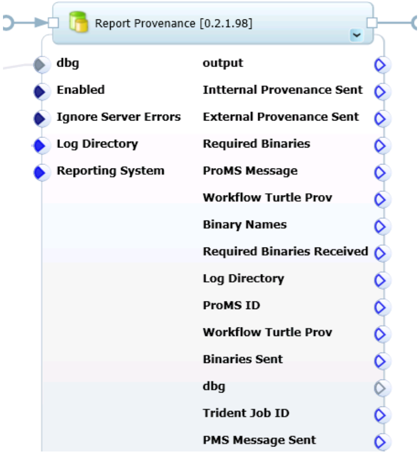

Figure 1: The generic *Report Provenance* workflow component that can be added to any Trident workflow that sends Reports to a specified PROMS installation each time the workflow is run. Note there are few configuration input fields (left side) and a large number of output fields (right side) indicating that this workflow component generates the majority of its outputs without much user input.

1.4. Other past data management projects

Various Sustainable Yields-type projects employed, and built on, the DM model developed within the MDBSY project, including the Northern Australia Sustainable Yields (NASY), South-West Western Australia Sustainable Yields (SWSY), Tasmania Sustainable Yields (TasSY), and Great Artesian Basin Water Resource Assessment (GABWRA). Although the MDBSY DM approaches were refined through these projects, the common weakness of DM's reliance on human processes for capturing metadata and provenance information remained, ensuring that a high level of data integrity was not attained.

2. NEW APPROACHES TO DM AND PROVENANCE

2.1. Learning from previous projects

To build on the MDBSY project's experiences with DM, we address criticisms of it from two sources: the authors themselves (who are also authors of this paper) and authors of a provenance scoping report for the Bioregional Assessments project. The first wished for:

1. Presenting immediate benefit to project staff in following metadata recording protocols – an immediate carrot, rather than a procedural stick to prompt action;
2. Better automation of metadata recording which would reduce staff effort;

[3] See the project homepage, https://wiki.csiro.au/display/proms/, for a full description of PROMS.
[4] See http://en.wikipedia.org/wiki/Representational_state_transfer for a discussion of the RESTful concept.
[5] http://en.wikipedia.org/wiki/Resource_Description_Framework.
[6] https://en.wikipedia.org/wiki/Uniform_resource_identifier

3. Better quality control of metadata records.

The second source, taken from Taylor, *et. al.*, (2013) lists that MDBSY DM:

4. Did not capture decisions by human experts;
5. Did not capture the actual processing steps;
6. Lacked consistency in metadata quality due to the manual processes required.

Issues 2, 3, 5 & 6 - implied that further automation of metadata recording could reduce staff effort while simultaneously capturing processing steps and ensuring consistent, high quality metadata reporting – these are addressed in Section 2.2. Point 4., relating to the previously mentioned issues of not being able to automate all (or even most) project processes, is addressed in Section 2.3 and Point 1. is addressed in Section 2.4.

2.2. Establishment of a flexible Provenance Architecture

The PROMS system described in Section 1.3 allows different project processes to use any file storage mechanism they choose, as long as it presents access to those files via URIs. Similarly, any provenance reporting agent can be used as long as it can communicate via HTTP and its reports conform to the PROMS data model. This allows a provenance reporting *approach* to be used via an architecture of replaceable components, rather than a specified provenance reporting *method*. A *method* was used by the MDBSY project where specific tools and storage locations had to be used. The more flexible *approach* will improve the uptake of provenance reporting by allowing the reporting and file storage components to be incorporated into existing processes.

Since the URIs are used for file and report storage addresses, URI redirection tools should be used to preserve access to them by preventing dead references when underlying systems change. This grants the flexibility to system implementers to use changing infrastructure for projects as they develop and also for long-term, post project, lifetime archiving.

```
import requests

def post_new_item(base_uri, title_text):
    resp = requests.post(base_uri, data=title_text)
    return resp.content

base_uri = "http://dids.example.com"
title_text = "Title for item posted from Python"
data_item_id= post_new_item(base_uri,title_text)

# DIDS_URI="http://dids.example.com"
# curl --request POST $DIDS_URI -H "Content-type:
text/plain" -d "Title for item posted from Shell"
```

Figure 2: Different types of code to register a file in the data storage component of PROMS thus making it available for a provenance report. Python (top) and Linux Shell Script (bottom).

2.3. Handling automated processes

Where project processes, such as model ensemble runs, and data manipulation are automated, DM becomes easy to institute by incorporating additional process steps that capture provenance each time a process is executed. Automation of project processes can be achieved through a number of means including dedicated workflow engines and simple computer scripting. Figure 1 showed a provenance reporting element from the dedicated workflow engine Trident and Figure 2 shows a reporting action to PROMS carried out using two different scripting languages. PROMS defines when and how file storage and provenance reporting must occur. As more projects come to build custom components to store files and report provenance, they may share those components with other users, thus further reducing the effort required to generate them for new projects[7].

2.4. Handling non-automatable processes

The general approach to managing processes that cannot be automated ("decisions by human experts" as they are referenced by Taylor, *et. al.*, (2013)) within the context of the provenance architecture is to:

1. **Identify the process** and record its existence in PROMS. Processes, automated or human, can be reported upon in the same way at the highest level (who conducted it, when, how long it took etc.);
2. **Store the data inputs** to that process in one of the applicable file storage systems. All processes have inputs and recording them should not differ whether human or automated. This allows chains of multiple processes to be linked (the outputs of one to the inputs of another);

[7] The authors maintain a list of such components, which will be demonstrated at this paper's delivery in MODSIM2013.

3. **Create a *wrapper*** agent that can be run every time the process is conducted. The wrapper will likely present as a web page with fields requiring input. The wrapper's back end can use the same data and provenance reporting tools used by automated workflow tools.

The above process has been tested in the Geofabric project (Car, N.J., 2013) resulting in provenance records for an expert process with inputs, outputs and occurrence of the process recorded but not the internal processing steps. This is commensurate with the PROMS data model's External level of reporting. One cannot inspect the inner workings of processes reported in this way but one can incorporate them into multi-process provenance chains.

3. CULTURAL CHANGE

3.1. What has changed

There have already been some changes in the data management culture within multi-disciplinary projects including the acknowledgement that they need to have a formal approach to data. There is also now a widespread agreement that it is necessary to focus on organisational needs beyond projects in order to allow the re-discovery of data for future use. There is also an acceptance of having common directory structures for project data archiving and the need for specific roles and responsibilities to manage data cataloguing across projects teams. Perhaps most critically, there is an agreement of the importance of recording audit trails that demonstrate the integrity of data allowing scientific results to be reliably defended.

3.2. What still needs to change

There is still a cultural barrier preventing scientists, modellers and technical support staff from following defined data provenance processes. It is still common for scientists to develop their processes without using standardised workflow engines meaning key decisions, analysis steps, and coding parameters are not being captured. This presents a risk to the defensibility of scientific information and remains the 'Achilles heel' of meaningful audit trails. It is therefore necessary to focus attempts at cultural change in this area by applying the new approaches to data provenance capture outlined in Section 2, along with appropriate resources and incentives, to migrate scientists and modellers into workflow engines and other automated frameworks.

3.3. Staff incentives for data management

Publication citations have long been a measure of scientific performance and achievement so perhaps the use of data citations may well provide a similar measure of data provenance management. CSIRO is currently considering widening scientific success metrics for staff performance from paper publications to include standards development and dataset generation. A metric for staff engaged within projects requiring DM could perhaps also be developed. In addition, rewards have been proposed to formally recognise DM achievement within projects and across CSIRO generally.

4. DISCUSSION

By providing an architectural approach to provenance reporting, rather than a specific method, we maximise the number of automated processes that can report provenance. By automating as many project processes as possible we reduce the effort required by staff and improve reporting quality. By providing a methodology to include non-automatable processes in the provenance reporting approach we ensure the single approach can cover a realistic range of project processes. These three points lower the barrier to a significant data management cultural change meaning the incentives required for staff to make the final leap over those barriers are also lowered. In addition, the quality of provenance information and the integrity of audit trails are significantly increased with most or all key decisions, processing steps and parameters being automatically captured. The provision of staff incentives through the recognition and implementation of data citation metrics and rewards will also help projects breach the final barrier to having a complete provenance capture for all reported scientific results.

A large and complex series of cross-disciplinary and cross organisation projects with strong audit trail requirements are about to be undertaken. The Bioregional Assessments program (IESC, 2013) contains six major projects and is being conducted over a 3 year period utilising numerous software models in a range of disciplines involving well over one hundred staff across 5 key organisations. It has a strong requirement to provide maximal data product lifecycle transparency, as well as dataset production repeatability, which can hopefully be achieved through data audit trails and provenance chains. It will need to implement processes to those undertaken by the MDBSY project, develop and apply appropriate workflow tools as developed for WIRADA, and then add the improvements suggested in Section 2 to avoid the shortcomings of the pure MDBSY approach outlined in Section 2.1. It will also present additional cultural and technical DM challenges due to its cross-organisational nature.

5. CONCLUSION

The developments in the data management culture and provenance capture technologies within the MDBSY project and WIRADA form a foundation for a new generation of DM that is hopefully more robust, coherent, responsive, and reliable. If this new DM can be applied to the Bioregional Assessments program with its unique DM requirements, that program may well prove to be a watershed event in the history of CSIRO's DM and perhaps serve as best practice within the Australian scientific community.

REFERENCES

Bureau of Meteorology (2013), Australian Hydrological Geospatial Framework (Geofabric). Web page by Bureau of Meteorology. Online at http://www.bom.gov.au/water/geofabric. Accessed 26/06/2013.

Car, N.J. (2013), Deliverable 2.3 + 2.4 Report - Demonstration of Geofabric product quality verification via provenance tracing indicating inputs & processes used in production. Unpublished project report from the Geofabric 12/13 Project within the Water Information Research and Development Alliance (WIRADA). http://www.csiro.au/en/Organisation-Structure/Flagships/Water-for-a-Healthy-Country-Flagship/WIRADA_WFHC_ResearchProfile.aspx.

Hartcher, M.G. and Lemon, D. (2009), Developing data audit trails for the CSIRO Sustainable Yields projects. In Proc. of 18th World IMACS Congress and MODSIM09. MSSANZ and International Association for Mathematics and Computers in Simulation, July 2009, pp. 2377-2383. ISBN: 978-0-9758400-7-8. http://www.mssanz.org.au/modsim09/J4/hartcher.pdf.

Groth, P., Moreau, L. eds. (2013). PROV-Overview. Web page retrieved on 2013-06-27 from http://www.w3.org/TR/prov-overview.

Independent Expert Scientific Committee on Coal Seam Gas and Large Coal Mining Development (IESC) (2013), Bioregional assessments. Web page retrieved on 2013-05-18 from http://www.environment.gov.au/coal-seam-gas-mining/bioregional-assessments/index.html.

Kloppers C, Liu Q, Taylor K, Walker G. 2012. WDTS provenance. Internal Project Report. CSIRO Water for a Healthy Country Flagship. 30 pp.

Lee, B. and Box, P. (2012), Generation of data product provenance information from HWB. CSIRO Water for a Healthy Country Flagship, Australia.

McGuinness, D., Ding, L., da Silva, P.P., Chang, C. (2007) PML2: A Modular Explanation Interlingua. In: Proceedings of the AAAI 2007 Workshop on Explanation-aware Computing, Vancouver, British Columbia, Canada, July 22-23, 2007, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.186.8633&rep=rep1&type=pdf

da Silva, P.P., McGuinness, D.L. and Fikes, R. (2004), A Proof Markup Language for Semantic Web Services, Technical Report KSL-0401, Knowledge Systems Laboratory, Stanford University.

Stenson, M.P., Fitch, P., Vleeshouwer, J., Frost, A., Bai, Q., Lerat, J., Leighton, B., Knapp, S., Warren, G., Van Dijk, A., Bacon, D., Pena Arancibia, J., Manser, P. and Shoesmith, J (2012). Operationalising the AWRA system. In WIRADA: Science Symposium Proceedings, Melbourne, Australia, 1–5 August 2011. CSIRO: Water for a Healthy Country National Research Flagship, pp. 36-45. http://www.csiro.au/~/media/CSIROau/Flagships/Water%20for%20a%20Healthy%20Country%20Flagship/WIRADA_Science_Symposium_Proceedings.pdf.

Taylor, K., Woodcock, R., Cuddy, S.M., Thew, P. and Lemon, D. (2013), Provenance for Bioregional Assessments. A commercial-in-confidence report to the Office of Water Science from CSIRO's Water for a Healthy Country Flagship, CSIRO Australia.