Jun 16th, 2:00 PM - 3:20 PM

Leveraging quality assurance and quality control processes to deliver provenance as a first-order scientific output in large-scale environmental assessments

William Francis
CSIRO Land and Water, will.francis@csiro.au

Nicholas J. Car
CSIRO Land and Water, nicholas.car@csiro.au

Rebecca K. Schmidt
CSIRO Land and Water, becky.schmidt@csiro.au

Simon Gallant
CSIRO Land and Water, simon.gallant@csiro.au

International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA,
Daniel P. Ames, Nigel W.T. Quinn and Andrea E. (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Leveraging quality assurance and quality control processes to deliver provenance as a first-order scientific output in large-scale environmental assessments

William Francis[1], Nicholas J Car[2], Rebecca K Schmidt[1], Simon Gallant[1]
[1]CSIRO Land and Water, Black Mountain, ACT, Australia
[2]CSIRO Land and Water, Dutton Park, QLD, Australia
{will.francis, nicholas.car, becky.schmidt, simon.gallant}@csiro.au

Abstract: In large-scale environmental assessments, such as the Australian Government Bioregional Assessment Programme, the science is integrated when products (scientific reports) are assembled. Product assembly involves synthesising material written by multiple authors from multiple disciplines and producing standard tables, maps and charts. High-quality products require that quality assurance and quality control (QA/QC) procedures are built into the scientific and publication processes.

Further, funders and stakeholders are increasingly demanding that scientists provide sufficient information to explain and justify the evidence provided, even to the extent that an independent group can repeat the science. Recording an effective lineage of processes and data – known as *provenance* – requires standardised reporting that relies on potentially complex methodologies for representation. Recording of provenance is enabled by process modelling, integration of scientific processes, and automation.

These requirements, to embed QA/QC procedures and cater for provenance, have significant areas of overlap. Whilst embedding QA/QC procedures is a more pragmatic, implementation-driven activity, and provenance annotation is more theoretical, these two areas are compatible and can be complementary.

This paper covers the implementation of a provenance system in the context of the Bioregional Assessment Programme; how existing QA/QC procedures were leveraged to deliver provenance; and, in turn, how the delivery of provenance resulted in improvements in the QA/QC procedures. The future evolution of QA/QC procedures in the Programme will be discussed and general principles synthesised for other similar large-scale environmental assessments.

***Keywords**:* provenance; quality assurance; scientific process integration; interdisciplinary; environmental assessment

1. BACKGOUND

1.1 Quality assurance and quality control in interdisciplinary large-scale environmental assessments

Since 2007, the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and its partners have undertaken a number of large-scale environmental assessments for the Australian Government. These comprehensive scientific assessments of current and future water availability in major water systems across Australia were designed to provide a consistent framework for future water policy decisions. The first large-scale assessment was the Murray-Darling Basin Sustainable Yields (MDBSY) Project, followed by similar assessments for Tasmania, south-west Western Australia, northern Australia and the Great Artesian Basin (CSIRO, 2014c). Similar approaches were used in the Flinders and Gilbert Agricultural Resource Assessment (CSIRO, 2014a) and the Pilbara Water Resource Assessment (CSIRO, 2014b).
The QA/QC procedures applied in the MDBSY Project are discussed in detail in Merrin and Cuddy (2009). An important part of their process involved the management of data *elements*. Data elements are the tables, maps and charts used in a product (a scientific report). They were managed through MS Excel spreadsheets which acted as registers for all the elements in a given product or section of product.

The QA/QC procedure involved the following steps:

1. generate data element from a data source in the project shared directory
2. create a new page in the MS Excel spreadsheet with associated identifier
3. copy the element onto the page and populate the element metadata, including (importantly) the link to the source data
4. copy the element into the product (MS Word document)
5. update the log in the product that links elements to the appropriate workbook via their unique element identifier.

Schmidt and Ahmad (2013) describe their experience since 2009 in implementing this non-automated workflow. They identified further development of data element and provenance processes as being required to improve QA/QC procedures for products.

Hartcher and Lemon (2009) discussed the role of data audit trails in the MDBSY Project. In their process, data element registers were managed in a metadata catalogue. This catalogue provided metadata for all datasets following the ANZLIC standard (ANZLIC, 2007) including data lineage in the form of written statements and links between datasets. The catalogue catered for data lineage only up to the second last step of the process to create datasets, with products (scientific reports) forming the last step. Archiving, as well as the lineage statements and links, ensures that datasets can be retrieved. These processes were not designed to support the storage and querying of the provenance of specific data elements in products.

1.2 The Bioregional Assessment Programme

A bioregional assessment is a scientific analysis of the ecology, hydrology, geology and hydrogeology of a particular geographic area, with explicit assessment of the potential direct, indirect and cumulative impacts of coal seam gas and large coal mining development on water resources (DoE, 2013; Barrett et al., 2013). The information generated by the Bioregional Assessment Programme is potentially useful for a range of stakeholders including state government regulators, coal seam gas and large coal mine proponents, and interested community members. The outputs are a suite of 13 distinct products for each of the 13 geographic areas currently being studied. All unencumbered datasets will also be made available.

The Programme team is both multi-disciplinary and multi-agency with four main agencies: CSIRO, Geoscience Australia, the Bureau of Meteorology and the federal Australian Government Department of the Environment. Nearly 200 people are working together to deliver over 169 products over the course of three years. In addition to scientists who specialise in the relevant disciplines, the Programme also includes a products team that undertakes the QA/QC procedures with respect to con-

tent, format and delivery. These editors, map-makers and technologists specialise in integration in interdisciplinary projects and have broad domain knowledge.

A key goal of the Programme is to provide transparent and discoverable information. Thus provenance information needs to be adequately captured at all stages of the development of products, including data processing and human decision making. The reporting of science outcomes in products is an important area requiring provenance, because readers of products often want answers to questions such as 'how was this result reached?'.

1.3 Provenance in scientific projects

Car et al. (2013) describe issues around implementing provenance reporting and storage systems in scientific projects with heterogeneous processes and also the use of automation as a driving force for cultural change through reducing the effort required for reporting. Workflow tools are posited as a key ingredient to success in the implementation of repeatable, integrative science projects. The automation of processes and the capture of provenance in environmental domains are presented by Cuddy and Fitch (2010).

The implementation of a provenance system for the Bioregional Assessment Programme – as discussed in Section 3 – follows aspects of these automation approaches to embed provenance capture in existing processes.

2. PROVENANCE FOR PRODUCTS IN THE BIOREGIONAL ASSESSMENT PROGRAMME

2.1 Class and process models

In a bioregional assessment, scientists analyse data and create *Element*s such as a table, map or chart. *Element* are used in *Sections* which are aggregated into *Products*, as illustrated in the workflow in Figure 1.

The data model for the objects referenced in Figure 1 is presented in Unified Modelling Language (UML) notation in Figure 2.

Having a formal data model for products and their sections allows the automated checking of products for adherence to Programme standards. Formal data models for datasets – and for other items relevant to products – are required for automated testing of products throughout the entire process.

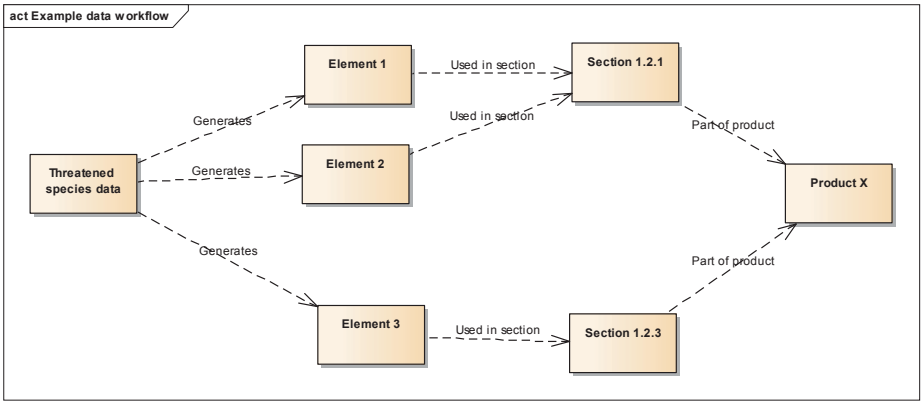

Figure 1 Workflow for analysing data and creating *Elements* that are incorporated into *Sections* and *Products*

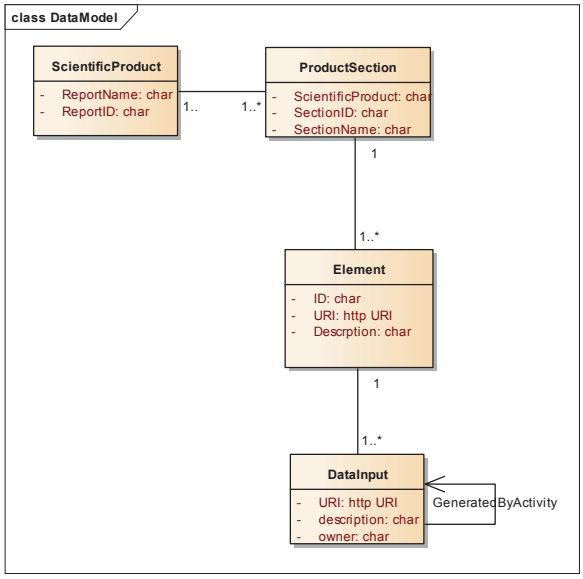

Figure 2 Data model for objects in reporting workflow

2.2 Automation of provenance reporting

The Provenance Reporting Tool is a software application written in Python using the third-party packages *arcpy* (Esri, 2012) and *xlrd* (Simplistix, 2012). The application reads element registers and delivers (i) provenance information for each element, (ii) validation of dataset use, (iii) quality statistics and (iv) an issue log. Using the arcpy function *mapping*, *ListLayers* and the *name* and *dataSource* properties of layer objects, documents that contain maps are interrogated and a list of the datasets underlying each map is created. The steps performed by the script are detailed in Figure 3.

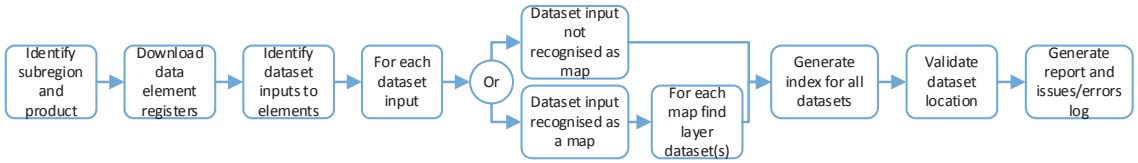

Figure 3 The Provenance Reporting Tool: processing steps for a product

While the Provenance Reporting Tool is specifically tailored for particular element types (maps), the process to verify the data source also applies to all other element types (tables and charts) and thus all products.

Before the Provenance Reporting Tool was implemented, data managers needed to perform the tedious manual task of checking to see whether each element type in a report really was represented in a data element register and linked to data in the repository. In contrast, data managers can now run the Provenance Reporting Tool repeatedly and with little effort. This allows errors in reporting to be identified quickly (overnight, when tests are automatically run) which prevents a large number of issues building up and having to be addressed in bulk which used to frequently happen just before a product was due for final release.

2.3 Provenance representation

The Provenance Reporting Tool takes its provenance reporting data model from the PROV standard (Lebo et al., 2013) and its outputs can be delivered in formats compliant with PROV Ontology (PROV-O). Figure 4 presents an example of the Provenance Reporting Tool's provenance output in simplified form for ease of reading. PROV-O is high-level and generic and other processes within the Programme report dataset provenance using PROV-O-compliant forms, meaning that output from them and the Provenance Reporting Tool is commensurate. With the use of a general-purpose provenance standard, provenance reports from this tool and the others can be used together allowing for their

automated traversal. This allows complete provenance graphs from report elements to reports to datasets and eventually source data to be constructed using semantic web queries. Without this general-purpose standard use, manual linkages between elements and datasets provenance would be required to answer questions about the full history of the development of the product.

```
:Dataset_5
   a      Entity;
   value"\\wron\project\BA\NIC\All\Maps\Templates\Shapefiles\State_Borders_10million.shp";
   title  State_Borders_10million;
.

:Activity_1
   a              Activity;
   used           Dataset_8;
   used           Dataset_2;
   used           Dataset_5;
   used           Dataset_10;
   used           Dataset_11;
   used           Dataset_14;
   used           Dataset_26;
   used           Dataset_16;
   used           Dataset_17;
   generated      Element_CEN-112-009;
.

:Activity_11
   a              Activity;
   used           Element_CEN-111-001;
   ...
  used            Element_CEN-112-009;
  generated       CEN_Product_1-1;
.
```

Figure 4 A snapshot the output from the Provenance Reporting Tool in pseudo PROV-O format (simplified for ease of reading)

Figure 4 models a dataset used to construct a map layer as a PROV-O *Entity* and the act of assembling a map made of multiple layers as a PROV-O *Activity*. Outputs *generated* by one *Activity,* also modelled as *Entities*, can be *used* by further, downstream *Activities.* In Figure 4, the *Entity* Element_CEN-112-009 is *generated* by Activity_1 and used by a later *Activity*, Activity_11. Standard PROV-O properties cater for most of the information fields required to make sense of the processes used to assemble the product.

The automation of the generation of provenance information at the product assembly stage has enabled easier validation of reported data. The integration of provenance reporting throughout a product's history – including the product assembly stage – has added traceability to every product's elements and allows the question 'how was this result reached?' to be answered for every element with minimal effort.

3. QUALITY IMPROVEMENTS REALISED THROUGH THE IMPLEMENTATION OF PROVENANCE

These developments in provenance reporting have had important follow-on effects to the quality assurance functions of the Programme's products team. Automation was used to increase both transparency and quality. This automation made reporting the status easier and more frequent, which – in turn – allowed Programme management to gain better insight into quality and to measure quality improvements. Flow-on effects to broader aspects of the Programme include encouraging the increased use of workflows and automated provenance reporting.

3.1 Comprehensive and efficient validation

In providing regularly scheduled, detailed provenance reports, this automated process has allowed data validation processes to leverage provenance. Data reports now provide details of whether datasets are stored in accordance with the Programme's data management protocols. If they are not

stored appropriately – for example, the referencing is incorrect, versions are out of date, or data source information is missing – then this issue is quickly revealed and can be corrected. Further, automated provenance reporting provides the means to link data elements into the Metadata Catalogue and test whether or not metadata has been associated with the data sources for elements.

3.2 Increased transparency in products

The increased frequency of provenance reporting – and associated data reports – increases the capacity of Programme management to assure quality at a higher level. Improved tools for reporting statistics on quality provide them with benchmarks and checkpoints for the approval of products. This quantitative approach ensures that approvals for products can justly be regarded as indicating their high quality.

4. DISCUSSION AND CONCLUSIONS

The use of existing QA/QC procedures to deliver provenance has provided multiple benefits to the Programme. The automation of tools to report detailed provenance at the product assembly stage has strengthened the existing QA/QC. This more detailed, and more quickly executable, reporting provides – at the same time as products are delivered – transparent information on which datasets were used in products.

The demonstration of the utility of automated QA/QC reporting has, in turn, demonstrated the utility of establishing data models for Programme data entities. This has raised confidence in proposals to establish Programme-wide data models for all Programme data entities.

The use of an international standard for provenance reporting has given focus to the forms of provenance data delivery that particular Programme tools and processes can implement. Use of a published standard also ensures long-term understanding of the provenance data, given that there is an international community of practitioners skilled in its interpretation and significant written resources to help future analysts. PROV is a recent standard and updated versions of it are expected. The work undertaken here contributes to the pool of experiences that will be used to inform future versions of it. It will also contribute to the structure of reporting for future projects of the authors: this work has established a report or a report generation data model which can be used for many scientific projects.

Automating provenance reports supports data management and the Programme's product approval process. In doing so, this activity has successfully demonstrated how technical tools can lower the barriers to implementing useful QA/QC processes. Hopefully this will drive Programme cultural change towards greater acceptance of completing the still non-automated data management tasks, such as metadata entry. If these tasks are not completed acceptably, this issue can be ascertained easily and followed up quickly.

ACKNOWLEDGMENTS

This research is part of the Bioregional Assessment Programme, which is funded by the Australian Government Department of the Environment. The Bioregional Assessment Programme is a transparent and accessible programme of baseline assessments that increase the available science for decision making associated with the impacts of coal seam gas and coal mining development on water resources. Bioregional assessments are being undertaken in a collaboration between the Department of the Environment, the Bureau of Meteorology, CSIRO Water for a Healthy Country Flagship and Geoscience Australia. For more information: www.bioregionalassessments.gov.au

REFERENCES

ANZLIC (2007), Metadata Profile: An Australian/New Zealand Profile of AS/NZS ISO 19115:2005, Geographic information — Metadata. ISBN: 978-0-646-46940-9.

Barrett DJ, Couch CA, Metcalfe DJ, Lytton L, Adhikary DP and Schmidt RK (2013) Methodology for bioregional assessments of the impacts of coal seam gas and coal mining development on water re-

sources. A report prepared for the Independent Expert Scientific Committee on Coal Seam Gas and Large Coal Mining Development through the Department of the Environment. Department of the Environment, Australia. Viewed 1 November 2013, http://www.environment.gov.au/coal-seam-gas-mining/pubs/methodology-bioregional-assessments.pdf.

Car NJ, Hartcher MG and Stenson MP (2013), Driving Data Management cultural change via automated provenance management systems 20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013.

CSIRO (2014a) Flinders and Gilbert Agricultural Resource Assessment. Website: http://www.csiro.au/fgara. Accessed 13/05/2014.

CSIRO (2014b) Pilbara Water Resource Assessment. Website: http://www.csiro.au/Organisation-Structure/Flagships/Water-for-a-Healthy-Country-Flagship/Water-Resource-Assessment/Pilbara-Water-Resource-Assessment.aspx. Accessed 13/05/2014.

CSIRO (2014c) Sustainable Yields Projects. Website: http://www.csiro.au/Organisation-Structure/Flagships/Water-for-a-Healthy-Country-Flagship/Sustainable-Yields-Projects/MDBSY.aspx. Accessed 22/03/2014.

Cuddy SM and Fitch P (2010) Hydrologists Workbench – a hydrological domain workflow toolkit. IEMSS 2010

Department of the Environment (2013) Overview of the Bioregional Assessment Programme. http://www.environment.gov.au/coal-seam-gas-mining/pubs/overview-bioregional-assessment-programme.pdf.

Esri 2012. ArcGIS 10.1 SP1 for desktop. Esri Inc, Redlands, California.

Hartcher, MG 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.

Lebo T, Sahoo S and McGuinness D (eds.) (2013) PROV-O: The PROV Ontology. W3C Recommendation 30 April 2013. http://www.w3.org/TR/prov-o/.

Merrin LE and Cuddy SM (2009). Implementation of a reporting workflow to maintain data lineage for major water resource modelling projects. In Proc. of 18th World IMACS Congress and MODSIM09. MSSANZ and International Association for Mathematics and Computers in Simulation, July 2009. https://mssanz.org.au/modsim09/J4/merrin.pdf

Schmidt RK and Ahmad ME (2013) Editorial workflows. IPED Conference Perth, 2013. http://marisa.com.au/conference/wp-content/uploads/2013/08/Schmidt_Ahmad-1-Editorial-workflows1.pdf.

Simplistix 2012. xlrd 0.7.2. Simplistix, Birmingham. www.python-excel.org.