2014

A Provenance Methodology And Architecture For Scientific Projects Containing Automated And Manual Processes

Nicholas J. Car

Matthew Paul Stenson

Michael Hartcher

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

A PROVENANCE METHODOLOGY AND ARCHITECTURE FOR SCIENCTIFIC PROJECTS CONTAINING AUTOMATED AND MANUAL PROCESSES

N.J. CAR (1), M.P. STENSON (1), M. HARTCHER (1)
(1): CSIRO Land and Water, Dutton Park, Brisbane, QLD, Australia

The management of provenance metadata is a pressing issue for high profile, complex, science projects needing to trace their data products' lineage in order to withstand scrutiny. To represent, capture, transfer, store and deliver provenance data from a project's processes, specialized metadata, new IT system components and the human and automated procedures are necessary. The collection of metadata, components and procedures can be termed a *provenance methodology and architecture*. Through our involvement with several large Australian science projects ([4], [5], [6], [7], [11]), we have developed a methodology that provides:

- Use Case assessments of project clients' requirements for provenance;
- team structures and project processes to facilitate provenance requirements;
- systems' behaviour to capture provenance from automated processes;
- behavioural patterns for project staff to capture provenance from manual processes;
- procedures for process compiling, storing and using provenance records.

Semantic web provenance ontologies have been created ([1], [2], [3]) that allow generic, abstracted provenance representation and we have extended the PROV ontology through our provenance data management ontology (PROMS-O) [8] in order to address provenance Use Cases required by our projects that PROV-O does not address.

Due to our project experience, we have developed a provenance architecture that specifies:

- a single provenance representation format for all project processes;
- the use of a persistent ID systems to alias other systems' URIs;
- an archival systems to store data and provide access to versions of their data via URIs;
- provenance management systems to store and provide access to provenance data;
- provenance exporters to capture and transmit provenance data from automated systems;
- provenance procedures to collect provenance data from human processes, and;
- an overarching integration architecture.

In this paper, we briefly mention our work regarding each of the points above which, together, provide a range of pointers to projects wanting to embark on provenance management.

PROVENANCE REQUIREMENTS

Requirements Analysis

Since provenance, as a distinct area of computer science / information systems / data management investigation, is a relatively new field, there is little literature regarding formal requirements analysis on the subject. Articles about provenance capture and use in large scientific project, such as [9], do not relate formal requirements analyses for provenance although, most likely, some sort of analyses was carried out. [9] defines provenance as *"the end result of ap-*

plying context-specific reasoning over a set of records that document the execution of a process, with the goal of deriving a set of properties of the data products involved in the execution" and states that provenance researchers' goal is to allow users to understand the origin of data by looking at the derived set of provenance properties for the data product under analysis.

Our experience with project stakeholders who profess a desire for provenance is that they do not yet have sufficient knowledge of provenance tasks and terminology to request specific provenance functionality. The Bioregional Assessments programme [5] clients insisted on high-level provenance goals such as "total process transparency" and "process repeatability" and well as the "long-term availability [of data]". In order to develop a nuanced response to this sort of requirement, we have undertaken tasks to understand and document specific provenance use cases in that project and, from there determine generic use cases across multiple projects. Table 1 gives some of the Use Cases we have documented in [8][1].

Table 1: Some generic provenance Use Cases

Category	Title	Question	Description
Inspection	Provenance of a data product	What is the lineage of Data Product X?	A user wants to know all about the ancestor Entities (data products) and Activities (processes) that contribute to the production of a data product.
Inspection	Provenance of a data product component	What is the lineage of Data Product component X including its relation to its parent?	A user wants to know the provenance of individual elements (database entries or individual files) within a project's data product.
Inspection	Descendants of a data product	What are the processes using, and Data Products derived from Data Product Y?	A user wants to know about the descendant Entities (data products) and Activities (processes) of a particular data product.
Inspection	Assemble a provenance graph	What's the complete provenance graph for Data Product Z?	A user wants to draw the complete provenance graph of all known processes including Data Product Z.
Reimplementation	Re-run a process	Can I re-run a project process and get a result either identical to or explainable different from the original?	A user wants to reproduce/regenerate results from a previous process. They mayn't get identical results due to changed input data but they do expect an identical process to run.
Reimplementation	Perform subsequent process runs	Can I re-run a project process using updated data inputs?	A user wants to produce new results using a previous process' methodology.

Team structures and project processes

To cater for from project clients' requirements for provenance and derived Use Cases, we suggest team structures and project processes to implement. It is clear to the authors and it has been published previously [10] that the human resource implementation of provenance tasks is perhaps the most significant obstacle that large-scale science projects need to overcome in order to implement effective provenance solutions.

Large science projects that have a significant data management component, such as [4], [5], [6] & [11], require dedicated data management staff. Project provenance duties fall within the wider remit of data management however implementing provenance systems can be an extremely technical task that may require skills not always possessed by data managers. Until such time as provenance methodologies become routinely understood and 'off the shelf' tooling is available for provenance management, it is likely that data management staff will have to rely on additional IT systems engineers who understand provenance.

From [10] and subsequent work, the authors suggest the integration of provenance tasks into project's main deliverable processes and both staff incentives and discipline measures to

[1] Specifically on the page https://wiki.csiro.au/display/PROMS/Provenance+Use+Cases

ensure they are carried out in line with other critical project processes such as financial accounting. Despite its data management successes, it is impossible to inspect datasets generated for the Murray Darling Basin Sustainable Yields project [11] in line with the Use Case questions in Table 1. Provenance processes, while specified for the project, were not given primary consideration by management. The Bioregional Assessments programme [5], however, has included the delivery of data product provenance in its requirements for datasets' public release. This will ensure provenance tasks are carried out by project staff.

Automated systems provenance capture

Most large science project can be expected to implement many automated processes using workflow tools or similar. Provenance capture form those processes can be especially efficient given that all the resources and logic associated with running them are present at execution time. However, long-term data management and the storage of workflow executables and configuration, in addition to their representation in provenance mark-up, need to be considered if Use cases from Table 1 in the Reimplementation category are desired. We propose a generic scientific process model in [8][2] that categorises inputs to scientific processes according to their *role* with examples being *data*, *configuration* and *algorithm* as depicted in Figure 1. Thinking of inputs to automated processes in these terms prompts project data management staff and the process owners (project scientists) to widen the range of digital artifacts they store for provenance reasons. There is a temptation for automated process owners to believe that, since their process is automated, it will naturally be implementable without specific process and data curation work. We have found this not to be the case by showing that the reimplementation of Microsoft Trident[3] workflows on systems other than the workflow designers' required significant effort to make data available and configuration settings known to new implementations.

Human process provenance capture

For science projects with heterogeneous process, many will not be automated and consist of manual actions. Detailed provenance capture for such processes is very hard however all-of-process provenance – what data went into and what data came out of a process – can be recorded reasonably easily. Project staff responsible for manual processes can record this level of provenance in accordance with the same conceptual model used for automated processes shown in Figure 1. In the next section we describe tooling that can help with this task.

PROVENANCE MANAGEMENT ARCHITECTURE

Provenance Representation

Since there is now, as of 2013, an international standard for provenance with an ontological expression (PROV-O) [1], we have chosen to implement it across all our projects. Where it is insufficient for our Use Cases we have specialized the ontology with the creation of our own Provenance Management System Ontology (PROMS-O) [8][4]. This ontology enables data access for Recreation Use Cases whereas PROV-O is primarily about provenance representation. Where this specialized ontology is used, it can be generalized to pure PROV-O ensuring that were someone inspecting provenance traces for a process represented in PROMS-O, they would be able to extract PROV-O details only if required.

PROMS-O allows for provenance metadata capture at a range of granularities including the data product level (the regular outputs of scientific projects), the sub-data product level

[2] Specifically on https://wiki.csiro.au/display/PROMS/PROMS+Scientific+Process+Modelling
[3] Trident project homepage: http://tridentworkflow.codeplex.com
[4] Specifically on the page https://wiki.csiro.au/display/PROMS/PROMS+Ontology

(items within data products such as individual files, individual database elements) and service-delivered data product and their elements.

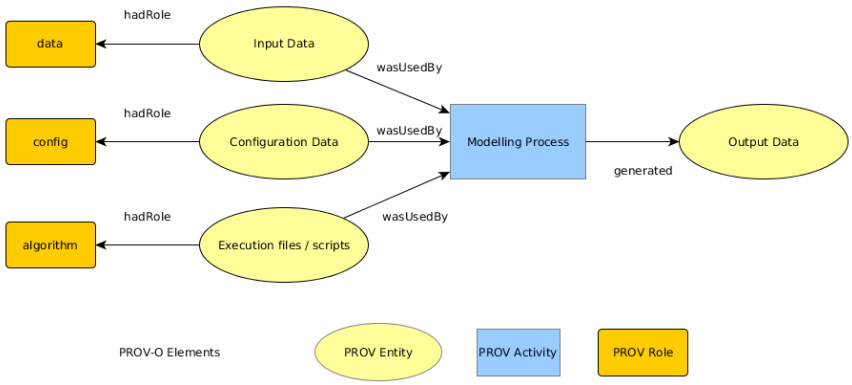


Figure 1: Role base classification of inputs to scientific processes

Persistent identification

The persistent identification of data products, data elements and other process inputs as per Figure 1 is critically important for Reimplementation Use Cases. Without this, a reference to a data item or execution code will not give access to a version of it or perhaps further metadata about it. To ensure long-term identify persistence we have implemented dedicated ID services for projects that allow a mapping between published identifiers and stored copies of the things, or metadata bout the things, that they represent.

For their universality of use and zero cost, we use HTTP URI[5]-based identifiers for all objects within project processes that need to be recorded for provenance. Compared with using externally managed identity systems such as DOI[6], we are able to mint new identifiers at will (as project processes need them) however the burden of their continuing resolution is also ours. HTTP URIs can be used directly in PROV-O/PROMS-O provenance documents to refer to objects and processes since the Semantic Web format they use, RDF[7], uses URIs for this task.

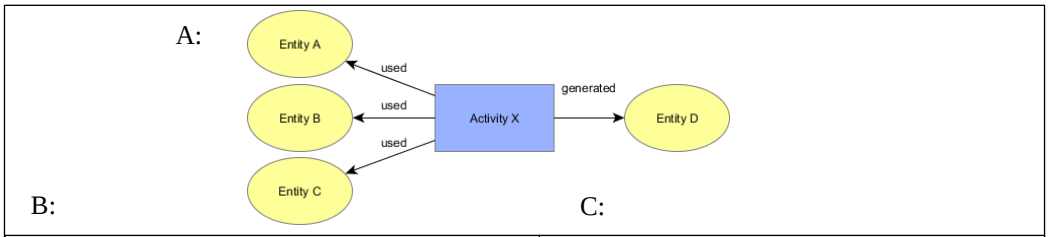
The primary tool we implement to manage URI-based identity is the PID Service[8] which acts as a much advanced version of the Apache web server's *mod_rewrite*[9]. This tool allows direct 1-to-1 URI mappings, pattern-based mappings and lookup table functionality. Its use, therefore, allows the storage and management of digital items to be managed independently of their published identity. This is crucial for science projects that require their resources to be available long term as storage systems and even the institutions in which items are stored, change over time and yet their identity must persist to must grant access to them.

Data storage

While PID Services provide a mechanism to abstract item identity from storage, storage is nevertheless an important component of provenance systems. "Semantically Enabled" storage systems are those that implement digital data curation with mechanisms allowing for metadata about their holdings, to be used in Semantic Web [12] applications. Digital repositories such as Fedora Commons and those conforming to the Open Archives Initiative[10] provide IDs for their data holdings as well as RDF-based metadata which allows provenance graphs to be construct-

[5] Uniform Resource Identifier: http://en.wikipedia.org/wiki/Uniform_resource_identifier
[6] The Digital Object Identifier system: http://www.doi.org
[7] RDF: http://en.wikipedia.org/wiki/Resource_Description_Framework
[8] Persistent Identifier Service: https://www.seegrid.csiro.au/wiki/Siss/PIDService
[9] Apache mod_rewrite homepage: http://httpd.apache.org/docs/current/mod/mod_rewrite.html
[10] http://www.fedora-commons.org & http://www.openarchives.org

ed that derive information about data items, such as their titles and descriptions, from the data store, rather than having them stored within the provenance document. This greatly reduces the burden on provenance production and storage systems. Figure 2 shows a snippet from an example PROV-O provenance record in graphical and *turtle*[11] formats: subplot A shows the graphical representation of an example process using PROV-O notation, subplot B contains a turtle representation of subplot A storing data product metadata within the provenance trace and subplot C leaves metadata storage to a Semantically Enabled data store.




```
:Activity_X
  a                 prov:Activity;
  dcterms:title     "Activity X"^^xsd:string;
  prov:startedAtTime "2013-01-01T12:30:01+11:00";
  prov:wasStartedBy  :eg_agent;
  prov:used          prov:Entity_b;
  prov:used          prov:Entity_d;
  prov:used          prov:Entity_d;
  prov:generated     prov:Entity_D;
.
:Entity_A
  a                 prov:Entity;
  dcterms:title     "Entity A"^^xsd:string;
  dcterms:creator   :eg_agent;
  proms:data        "http://ba.gov.au/dataset/abc123"
.
...
:Entity_D
  a                 prov:Entity;
  dcterms:title     "Entity D"^^xsd:string;
  dcterms:creator    :eg_agent;
  proms:value         "http://ba.gov.au/dataset/def456"
.

:eg_agent
  a                 prov:Agent;
  foaf:name            "Nicholas Car"^^xsd:string;
  foaf:mbox            <mailto:nicholas.car@csiro.au>
```

```
:Activity_X
  a                 prov:Activity;
  dcterms:title     "Activity X"^^xsd:string;
  prov:startedAtTime "2013-01-01T12:30:01+11:00";
  prov:wasStartedBy  <http://ba.gov.au/ppl/car587>;
  prov:used          <http://ba.gov.au/ds/abc123>;
  prov:used          <http://ba.gov.au/ds/123abc>;
  prov:used          <http://ba.gov.au/ds/abcadc>;
  prov:generated     <http://ba.gov.au/ds/def456>
.
```

Figure 2: A, graphical B & C, machine readable representations of a provenance graph

Provenance management systems

Provenance data, when generated, needs to be stored in a metadata system. While there are a range of off-the-shelf tools for storing general metadata (metadata about the spatial location, ownership and thematic context of data items), such as GeoNetwork[12] which implement international metadata retrieval standards[13], and while there is the PROV-O provenance metadata format, there are no widely used provenance storage system implementations.

The authors have been working on a document-driven database system with a web API[14], known as the Provenance Management System (PROMS) ([8] and [13]) that can store PROV-O and PROMS-O documents for processes carried out on project data products and other items of interest. It can deliver the contents of those documents in a number of ways including via graph visualization, machine readable provenance encodings (RDF and turtle) and via query endpoints. This allows PROMS to be used in much the same way as a metadata catalogue.

Due to Semantic Web methods of delivery, PROMS can also be viewed as a triplestore[15] allowing clients capable of semantic reasoning[16] to do so over its contents. Used with Semanti-

[11] Turtle - Terse RDF Triple Language: http://www.w3.org/TeamSubmission/turtle
[12] http://geonetwork-opensource.org
[13] The Catalogue Service: http://www.opengeospatial.org/standards/cat
[14] Application programming interface: http://en.wikipedia.org/wiki/API
[15] http://en.wikipedia.org/wiki/Triplestore

cally Enabled storage systems, PROMS can deliver provenance representations of past actions for scientific processes and also access to data products the processes used and generated.

Provenance exporters from automate processes

"Scientific" workflow engines store information about runs according to formal data models – effectively provenance information. The Microsoft Trident workflow engine stores a superset of the data required to generate a PROV-O representation of a process. When used with Semantically Enabled data stores, Trident can generate PROMS-O representations of its runs.

For the WIRADA Geofabric project [7] some of these authors were involved with building workflow exporter elements for Trident that deliver PROMS-O representations of any workflow that includes them. Figure 3 shows the Report Provenance element which can be placed at the end of a workflow. Note there are few input fields (left side) thus this workflow component generates PROV-O/PROMS-O outputs with little first use configuration and no manual run-time input. Used in conjunction with workflow elements that interact with Semantically Enabled data stores, users can report provenance to PROMS with very little per-run effort.

Other workflow systems, such as CSIRO's own Workspace tool[17] have recently also added the ability to report provenance in a method similar to above. Any workflow tool with a data model commensurate with the PROV-O/PROMS-O can be able to do so.

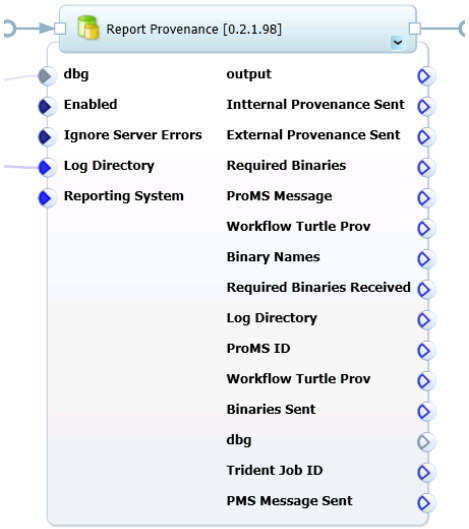


Figure 3: The provenance reporting component of MS Trident's workflow engine

Provenance procedures for manual processes

Reports of provenance from manual processes are more difficult to make and less detailed than those from automated processes. Reporting standardized all-of-process provenance, which we term an *External* view of provenance according to the PROMS-O, can be reported for manual processes and is a major advance on non-standardized reporting. Data products that conceptually form a product chain can be linked through External provenance reports when they are all stored in a PROMS instance or equivalent system. This means that reports generated from manual and automated processes can be used together to describe entire project processes.

In order to capture External provenance from non-automated processes in accordance with PROV-O/PROMS-O, we have created a series of project-specific web forms as well as a generic one that can be used to receive user input and deliver a standards-compliant report to a designated PROMS instance. Figure 4 shows the partial screenshot of a generic web form re-

[16] http://en.wikipedia.org/wiki/Semantic_reasoner
[17] http://www.csiro.au/Outcomes/ICT-and-Services/Workspace.aspx

porter. Using such a form requires the user to have already stored his input and output data products in a Semantically Enabled data store and to have their URI-based IDs ready to hand. Such web forms can be incorporated into other project processes such as general metadata reporting. This is the approach taken by the Bioregional Assessments programme where the fields in Figure 4 are seamlessly integrated into GeoNetwork-style metadata entry forms.

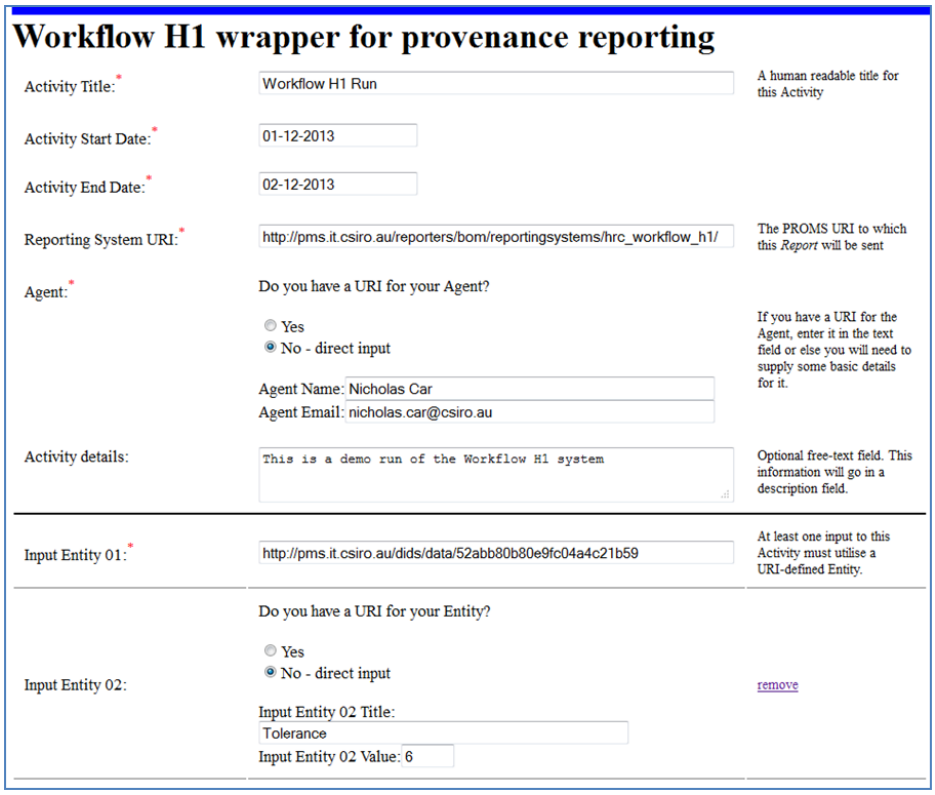


Figure 4: Partial screenshot of a manual process provenance reporting form

Provenance architecture

A provenance architecture using the elements described in this paper is shown in Figure 5. With the use of standards between them, components in the architecture can be replaced with functional equivalents. This ensures system robustness for long-term data and metadata access.

DISCUSSION & CONCLUSION

Analyzing Use Cases for software projects is standard practice and needs to be applied to provenance tasks. Such a move will allow standard provenance tasks to be discerned which can then be used to guide scientific project clients' provenance choices. Experience with Use Cases similar to those in Table 1, such as metadata capture, indicate that project processes and staffing structures need to be considered as well as the implementation of provenance data models and infrastructure components in order to reach satisfactory levels of reporting.

Provenance reporting and representation has an international standard to use however the standard doesn't cater for all of the provenance Use Cases the authors have derived from scientific project client's wishes. It is expected that a second version of the PROV standard will eventually be developed and it is hoped that Use Cases not catered for now then will be.

While standardized provenance reporting from a range of both manual and automated processes is proved, using certain kinds of data stores reduces the reporting effort. With the Semantic Web as the interoperability layer between data stores, reporting components and provenance stores, semantic web clients can infer relationships between these components and build provenance graphs containing useful information sourced from all of them. The standardized

layer also allows for components of the architecture to be replaced if need be. This removes dependencies on particular technology stacks and builds in the ability to develop new components or extended components for new provenance Use Cases over time.

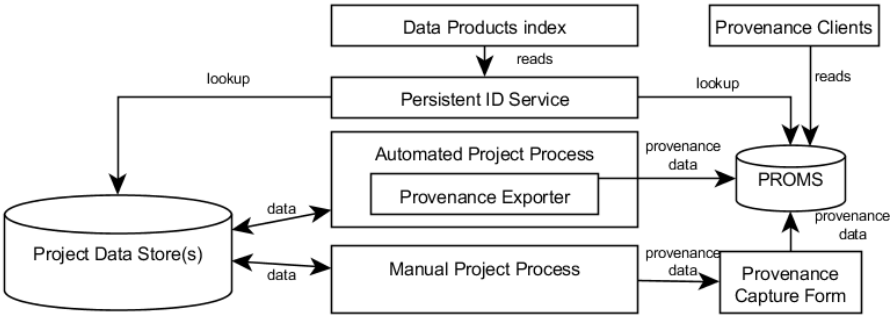


Figure 5: A complete provenance architecture

REFERENCES

[1] 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 2014-03-23.

[2] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, E. Plale, Y. Simmhan, E. Stephan, and J. Van den Bussche (2010), The open provenance model core specification (v1.1). Future Generation Computer Systems, July 2010. doi: 10.1016/j.future.2010.07.005.

[3] D.L. McGuinness, L. Ding, P. Pinheiro da Silva, C. Chang. (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.

[4] CSIRO (2012), Australian Water Resources Assessment (AWRA). Project homepage: http://eos.csiro.au/awra, accessed at 2014-03-23.

[5] CSIRO (2014), Bioregional Assessments. Project homepage: http://www.csiro.au/Organisation-Structure/Flagships/Water-for-a-Healthy-Country-Flagship/Water-Resource-Assessment/Bioregional-Assessments.aspx, access 2014-03-23.

[6] eReefs Project Management (2014), eReefs project homepage: http://ereefs.org.au, accessed 2014-03-23.

[7] Bureau of Meteorology (2014), The Australian Hydrological Geospatial Fabric. Project homepage: http://www.bom.gov.au/water/geofabric, accessed at 2014-03-23.

[8] Car, N.J. (2014) Provenance Management System (PROMS). CSIRO Land & Water, online wiki page, https://wiki.csiro.au/display/PROMS, accessed 2014-03-23.

[9] M. Branco & L. Moreau (2006). L. Moreau & I. Foster (eds.) Enabling Provenance on Large Scale e-Science Applications, chapter in Provenance and Annotation of Data, Lecture Notes in Computer Science pp 55-63. Springer Berlin, ISBN 978-3-540-46302-3.

[10] Car, N.J., Hartcher, M.G. and Stenson, M.P., 2013. Driving data management cultural change via automated provenance management systems. MODSIM2013. MSSANZ, Dec 2013, pp. 2173 - 2179. ISBN: 978-0-9872143-3-1. (PDF).

[11] CSIRO (2013), The Murray-Darling Basin Sustainable Yields Project. Project homepage: http://www.csiro.au/Organisation-Structure/Flagships/Water-for-a-Healthy-Country-Flagship/Sustainable-Yields-Projects/MDBSY.aspx, accessed at 2014-03-23.

[12] Berners-Lee, T. (2001). The Semantic Web. Scientific American. Online at http://www.scientificamerican.com/article/the-semantic-web, accessed 2014-03-01.

[13] Car, N.J. (2013). A method and example system for managing provenance information in a heterogeneous process environment a provenance architecture containing the Provenance Management System. MODSIM2013, 20th International Congress on Modelling and Simulation. MSSANZ, Dec 2013, pp. 824 - 830. ISBN: 978-0-9872143-3-1. (PDF).