IJDC | *Practice Paper*

Provenance in Support of the ANDS Four Transformations

Andrew Treloar
Australian National Data Service

Mingfang Wu
Australian National Data Service

Abstract

This article introduces the provenance activities that are being carried out at the Australia National Data Services (ANDS). Since its beginning, ANDS has been promoting four data transformations so that Australia's research data become more valuable and reusable by researchers. Among many other activities that enable the four transformations, ANDS has been encouraging ANDS partners to capture and describe rich context at the time when a data collection is created. In 2015, ANDS funded a number of external projects that had provenance components. In addition, ANDS is working on the interoperability between the schema that is used by the ANDS research data registration and discovery service – Research Data Australia (RDA) – and the W3C recommended provenance standard, Provenance Ontology (PROV-O), and investigating how to enrich the schema to access provenance information. The article concludes by discussing the lessons we learnt and our future planned activity.

Received 20 October 2015 ~ *Accepted* 24 February 2016

Correspondence should be addressed to Mingfang Wu, Building F, 900 Dandenong Rd, Malvern East 3145 Australia. Email: mingfang.wu@ands.org.au

An earlier version of this paper was presented at 11th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v11i1.416

ANDS and the Four Transformations

The Australian National Data Service[1] was first funded by the Australian Commonwealth Government in January 2009, with the aim of transforming Australia's research data environment. Since that time, ANDS has been working on the following four transformations:

- from data that are unmanaged to managed, structured collections;
- from data that are disconnected to well-connected collections;
- from data that are invisible to researchers other than its creators to collections easily findable by other researchers;
- from data that are single-use to reusable collections.

To enable the four transformations, ANDS has been working with ANDS partners to: 1) set up data management policy and procedure to make data well managed; 2) encourage ANDS partners to describe not only data collections, but also rich context by linking data collection to researchers, their projects and software to make data well collected and to add value; 3) provide the Research Data Australia (RDA)[2] data registry for ANDS partners to register their data and the RDA data portal to publish and make data discoverable. All the above are intended to ensure that the data that leads to a scientific finding or publication can be trusted and verified, and that data can be re-purposed and reused trustfully in more research. For that, data provenance plays a crucial rule.

Relevance of Provenance for the ANDS Four Transformations

The W3C Provenance Incubator Group defines provenance of a resource as:

> 'a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance' (W3C, 2010).

This definition states that provenance is associated with not just a data product's history, but also with the relationships between a data product and other entities that enabled the creation of the data. Provenance answers questions about where data originated, how data are produced, and who has been involved in producing them. Data provenance is becoming increasingly important, especially in the eScience community where research is data intensive and often involves complex data transformations and procedures (Simmhan, Plale and Ganno, 2005).

1 ANDS: http://www.ands.org.au
2 Research Data Australia: https://researchdata.ands.org.au/

The W3C Provenance Working Group[3] recommends six high level specifications including: PROV Primer, PROV Ontology (PROV-O), PROV Data Model (PROV-DM), PROV Notation (PROV-N), PROV Constraints, and PROV Access and Query (PROV-AQ) (Groth and Moreau, 2013). A discipline may implement a provenance management system in its own way and adopt a discipline oriented schema to capture and describe provenance information, but may map it to a high level provenance description such as PROV-O (or other serializations of PROV-DM) in order to enable interoperable interchange of provenance information (Feng, 2013). For example, Feng (2013) and Di, Shao and Kang (2013) did a mapping from the lineage section of a discipline specific metadata ISO 19115 to PROV-O, and the DCMI Metadata Provenance Task Group worked on and made available a mapping from generic metadata Dublin Core to PROV-O (W3C, 2013).

Although ANDS didn't explicitly express the need for provenance in the four transformations, each transformation and a range of ANDS programs have entailed provenance activity.

Management

For the Management transformation, provenance information ideally needs to be captured as closely as possible to the data capture or generation event. Once captured, it can then be used in support of ongoing management of the data. For instance, if a particular instrument has been found to need recalibration, knowing which data was captured from that instrument can be used to also recalibrate the existing data. From 2010 to 2013, ANDS ran a data capture program[4] (about 73 projects) with the aim to simplify and automate the process of researchers' routinely capturing data and rich metadata as close as possible to the point of creation, and depositing these data and metadata into well-managed stores.

Connection

For the Connection transformation, provenance information provides many possible connection points: from the instrument that generated/captured the data, to the samples being analysed, to the software or workflow that processed it, and so on. Each of these connections add more context to the data. Three activities are worth noting here. Firstly, ANDS has encouraged ANDS partners through all ANDS engagements to capture and describe rich context that is related to a data collection. Secondly, the Research Data Australia, a data discovery portal run by ANDS, has adopted the Registry Interchange Format – Collections and Services (RIF-CS) scheme[5] to model the rich relationship between a data collection and other resources involved in creating the data collection. The RIF-CS data model is based on ISO2146:2010 (ISO2146, 2010) and has four high-level entities: Collection, Service, Group and Activity; and a rich relation vocabulary to describe the relationship between any of two objects. This relationship, together with other RIF-CS elements such as dates that a collection was created or modified, provides provenance information on what has been used to produce a data collection, who has been involved, who was responsible, when the data were collected, and for what purpose a data collection was created. Thirdly, ANDS has run an eResearch

3 W3C Provenance Working Group: http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
4 ANDS project registry – data capture projects: https://projects.ands.org.au/getAllProjects.php?start=dc
5 About RIF-CS: http://guides.ands.org.au/rda-cpg/rifcs

Infrastructure Connectivity (eRIC) Program (with nine projects)[6] to make better connections between data storage, data-focused compute services, and metadata. The eRIC projects are described below.

Discovery

For the Discovery transformation, each of the provenance connections provide possible alternative discovery paths. Figure 1 shows an example of a data collection record in RDA. The data collection 'Prognostic gene set signatures derived from breast cancer microarray gene expression data' is connected to its computation model 'Computational Model for Gene Set Analysis…' (Service), two input Collections ('Five human breast cancer microarray gene expression datasets' and 'Five human gene sets from MSigDB Molecular Signatures Database'), five researchers (Party) who were involved in creating the dataset (one of them the data collector), and publications with scientific findings drawn from analysis of the collection. A RDA user can follow the links to discover what other output collections have been generated from the computation model; and if a user goes to the landing page for the computational model, they will find a description of the model in general, the steps taken and the software codes associated with each step.

Reuse

For the Reuse transformation, information about the processing paths applied can be used to help with re-analysis. ANDS has run an applications program from 2012 to 2014 (with 24 projects)[7] that demonstrated value of connecting and providing services over data (Wu, Kethers and Treloar, 2013). Provenance information also adds to the levels of trust that the data re-user can assign to a data collection, meaning they are more likely to want to reuse it.

Provenance Approaches

In general, approaches to capture and represent provenance can be described on a number of dimensions:

- internal or external,
- described using generic or discipline specific schema,
- represented in machine readable message or human readable report.

Examples of the internal approach are workflow systems such as Kepler[8], Galaxy (Goecks, et al., 2010) or Taverna[9] that capture provenance trails inside their own infrastructure during their processing activity. The provenance information is typically

6 ANDS Project Registry – eRIC projects: https://projects.ands.org.au/getAllProjects.php?start=eric
7 ANDS Project Registry – applications projects: https://projects.ands.org.au/getAllProjects.php?start=app
8 Getting Started with Kepler Provenance 2.5: https://code.kepler-project.org/code/kepler/trunk/modules/provenance/docs/provenance.pdf
9 Taverna Provenance Management: http://www.taverna.org.uk/documentation/taverna-2-x/provenance/

only available to other users of the same system. The external alternative would be to export provenance to a separate provenance store.

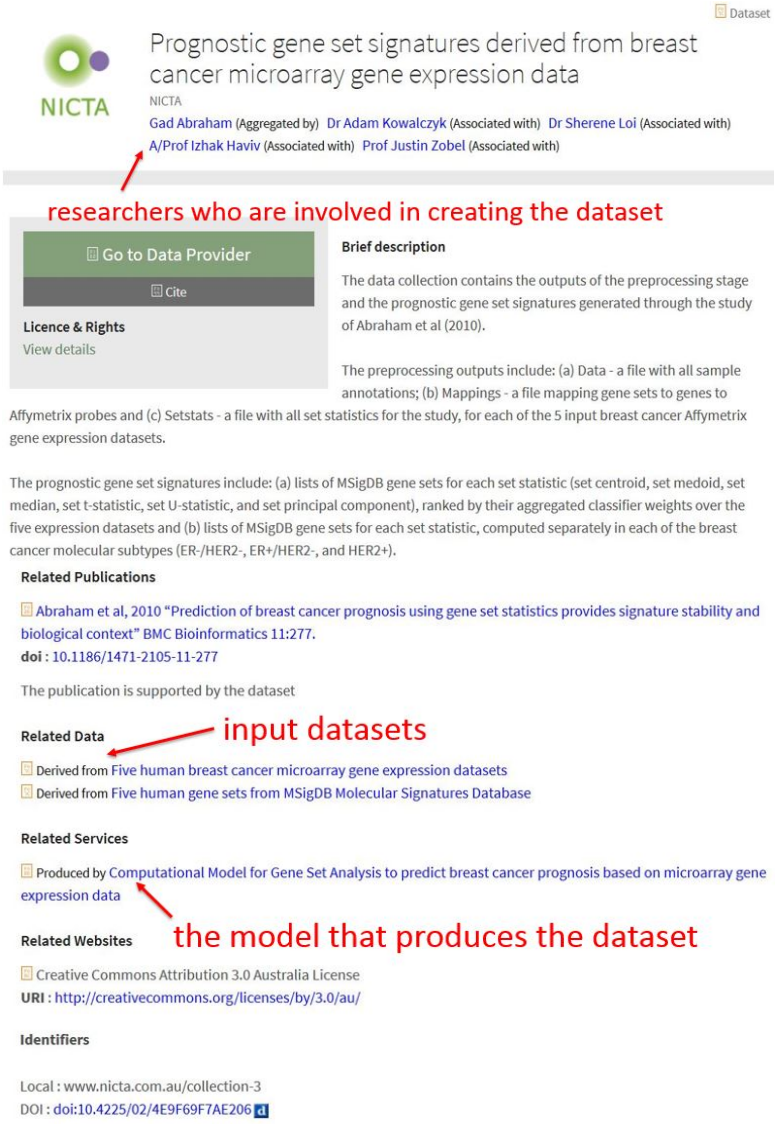

Figure 1. An example collection record in RDA.

Systems that adopt the internal approach tend to capture provenance in proprietary ways (because there is no reason not to do so). Systems that adopt an external approach often do so in a standards-based way, such as W3C PROV-O (Lebo, Sahoo and McGuinness, 2013); the external provenance stores use a standard because they need to interact with many different kinds of systems.

Provenance information can be described in a range of ways, from very generic metadata standards, such as Dublin Core metadata[10], disciplinary metadata standards, such as the linkage model sections in ISO 19115-2 (ISO19115, 2009), through to customised schemas that are developed to fit certain use cases. The last two approaches offer greater discipline richness. Alternatively, provenance information can be described directly in the W3C Provenance Data Model and (PROV-DM) Provenance Ontology

10 Dublin Core Metadata Initiative: http://dublincore.org/

(PROV-O) (W3C, 2010). Provenance information captured in Dublin Core and domain specific metadata can be mapped to PROV-O representation (W3C, 2013; Feng, 2013), so that provenance can be talked about at domain specific level and more abstract PROV-O level.

Finally, provenance information can be captured in a way that supports machine-machine interactions (for instance, to allow resource identification and location, and to allow workflows to be rerun) and/or at a higher level that allows for human users (without computing and semantics background) to more easily read the provenance trail of a data product or a data processing workflow. In some cases, this might just be a textual description, but might also involve a visualisation of the machine-readable representation, such as VisTrails[11].

ANDS partners have been exploring all of these dimensions in differing ways.

ANDS' Provenance-Related Activity So Far

ANDS Funded External Projects

The focus of provenance activity so far has been within a set of externally funded projects that mostly concluded in mid 2015. These were under the eResearch Infrastructure Connectivity (eRIC) program of work. This involved creating or strengthening connections between data storage (typically collections held on state-based storage nodes or other storage solutions), data-focussed compute services (such as ones being developed through the National eResearch Collaboration Tools and Resources (NeCTAR[12]) Virtual Laboratories or other providers), and descriptions made available through the ANDS Research Data Australia national data discovery service. The desired result was that there would be connectivity between the data and the services, and that the data would be maintained and visible. A number of the eRIC projects chose to engage actively with a provenance-based approach.

eRIC03 – Solid Earth

The Solid Earth project implemented a provenance service based on the provenance management system (PROMS)[13] (Car, 2013) and integrated it with the Virtual Hazard Impact and Risk Laboratory[14] (VHIRL) to capture workflow provenance, thus establishing a method of transparency for data and modelling tools used in the natural hazards domain. The resulting provenance service enables VHIRL to capture all aspects of any workflow complemented with its environment, produces provenance in compliance with the PROV-O standard and publishes this information to a provenance store that is external to the VHIRL workflow environment (Wise et al., 2015). The PROMS also includes a querying and reporting component (via a RESTful API and a SQARQL endpoint) for accessing provenance data and reporting back largely in machine readable message.

11 VisTrails: http://www.vistrails.org/index.php/Main_Page
12 NeCTAR: http://nectar.org.au/
13 PROMS: https://confluence.csiro.au/display/PROMS/The+Provenance+Management+System;jsessionid=61893034C307CA705EE8A4A10F5624FE
14 VHIRL: http://www.vhirl.net/

eRIC01 – Improving Access to Shelf and Coastal Data and Models through the Marine Virtual Laboratory (MARVL)

MARVL[15] is an online eResearch environment that simplifies the workflow of oceanographic model simulation studies by increasing the efficiency of data collection, the assembly of initial and boundary conditions, and by providing different options for simulation. MARVL uses GeoNetwork to store and manage metadata. The project planned to adopt the approach as taken by the eRIC03 project. After discussions with the eRIC03 project team, this project identified suitable provenance metadata from MARVL for delivery to PROMS. The project team is also considering whether to include the provenance metadata in GeoNetwork as an addition/alternative to PROMS.

eRIC05 – Data Management/Climate and Weather

The National Computational Infrastructure facility installed the Provenance Management System (PROMS), a provenance PROV-O repository and reporter at NCI and are making it available for further researchers to use.

eRIC02 – Terrestrial Systems

The Terrestrial Systems project developed the CoESRA system (Collaborative Environment for Ecosystem Science Research and Analysis) using the Kepler workflow system. CoESRA configured the Kepler provenance add-on module to capture provenance metadata, which is stored in a relational schema. CoESRA users can view and review each of their workflow executions via SQL queries within the CoESRA platform. At this stage, the provenance metadata has not been made available for discovery purposes. The team is planning to explore extraction of the provenance from MySQL and map it to PROV-O.

eRIC04 – All Sky Virtual Observatory (ASVO) Enhancement

The Theoretical Astrophysical Observatory[16] (TAO), a part of ASVO, houses a growing ensemble of theory datasets and galaxy formation models, with tools to map simulated data onto an observer's viewpoint and apply custom telescope simulators. The project adopted an internal approach and a discipline specific schema to describe provenance. The project deployed a module within TAO to share and publish data product and workflow. The project team developed a metadata schema specifically for their use cases to capture metadata and provenance information for reproducing a workflow. Provenance information captured includes the parameter setting of each data processing module, the relationship between related modules in a workflow, and the linkage between data and module. They also support module versioning and schema versioning. Within this environment, a researcher can rerun a simulation such as the generation of simulated galaxies.

eRIC06 – Climate Change Adaptation and Diversity

The Biodiversity and Climate Change Virtual Laboratory (BCCVL) is a comprehensive platform that provides access to numerous species distribution and trait modelling tools; a large and growing collection of biological, climate, and other environmental datasets; and a variety of experiment types to conduct research into the impact of climate change on biodiversity (Hallgren et al., 2016). The provenance information is captured through the construction of an experiment by recording input datasets, transformation tools and output datasets; the provenance is described in the

15 An Introduction to MARVL: http://www.marvl.org.au

16 Theoretical Astrophysical Observatory: https://tao.asvo.org.au/tao/

 doi:10.2218/ijdc.v11i1.416

W3C's PROV-O with some Dublin Core and a few custom properties. This implementation enables the ability to determine the full provenance chain of experiments including originating datasets, versions of algorithms used, software utilised and configuration options. This information allows reproducibility of results from the Virtual Laboratory and provides the ability to publish provenance information.

eRIC07 – Data-intensive Collaboration on the Genomics Virtual Laboratory

The Genomics Virtual Laboratory[17] (GVL) aimed to facilitate best practice genomics in the Australian research community by providing a highly accessible, functional and reproducible cloud-based genomics analysis environment. For the provenance part, the GVL utilities the Galaxy's automatic history track and user annotation functions to record input datasets, tools used, parameter values, and output datasets. This provenance information is saved in files that can be reviewed in the Galaxy analysis workspace or viewed in Galaxy pages (Doak, Wu and Ganote, 2014). Provenance information and workflows can be shared and made open through the Galaxy public repositories (Goecks et al., 2010). The project also implemented a mechanism for publishing the Galaxy History metadata to the ANDS registry.

Provenance Activity within ANDS

ANDS has conducted two internal activities: one is to map the RIF-CS schema to PROV-O in order to enable interoperable interchange of provenance information; the other is to explore linking provenance information into RDA (Wu and Treloar, 2015).

Mapping from RIF-CS scheme to PROV-O

As discussed above, the RIF-CS schema that is used by RDA captures rich provenance information, which can be mapped to a high-level provenance model or ontology, such as PROV-O. Such a mapping provides many advantages. Firstly, it bridges the gap between the RIF-CS community (thus ANDS partners) and the PROV community. In the past, ANDS spent a lot of effort in promoting the RIF-CS scheme to ANDS partners and explaining why the data capture context should be captured, described, stored and linked to data. The mapping can provide valuable insights into different characteristics of both data models, and may help ANDS to explain PROV from a RIF-CS point of view to ANDS partners. Secondly, it can bring awareness of provenance to RDA data providers. Thirdly, it can improve interoperability between RIF-CS and PROV-O applications, especially to help developers to derive provenance metadata from collection records that are available from RDA.

Linking provenance information to RDA collection records

RDA was primarily designed for researchers to find data. We would like RDA users to get benefit from work done by the PROV community. Although a RDA user can get some provenance information through relationships and some other RIFCS terms, it has its limitations:

- Some data providers may not be able to publish metadata of all data collections and associated resources in a provenance chain to RDA.

17 Genomics Virtual Lab: http://genome.edu.au/

- Some data providers have set up a Provenance Access and Query service (PROV-AQ), which enables the construction of a whole picture of the provenance chain that involves a collection (or a service etc).
- Usually up-to-date provenance information is kept in the data provider's provenance management system.

We took a very simple approach as the first step to address the above limitations, this was to expand RIF-CS terms to include provenance. This is now a property of the relatedInfo element, that means data providers don't have to describe and publish provenance information in RDA, but can just provide a resolvable URL from a data collection record that leads to any extra provenance they would like a user to see. Of course, this assumes that the linked provenance information is human-viewable (see below) which may not be the case.

Community Coordination

Since 2014, ANDS has been facilitating an Australian Research Data Provenance Interest Group[18]. This brings together those active in provenance work to share experiences and expertise, and to plan joint future work. In November 2014, ANDS organised a provenance workshop that attracted about 27 attendees from seven national organisations and one from the RD-Alliance. Before the workshop, the attendees submitted their provenance use cases through an online tool[19] (developed by Nick Car, then at CSIRO, now at Geoscience Australia). During the workshop, the attendees discussed and classified the use cases into four topics for group discussions. The topics included provenance creation, use, publish and adoption[20].

This group also serves as way to facilitate Australian engagement with the Research Data Alliance Provenance Interest Group[21].

Next Steps

Lessons Learnt

Through the eRIC projects as described above, our partners have come to recognise that adopting provenance is more complicated than originally anticipated. The lessons learnt can be summarised as these:

- It is not possible to isolate provenance capture from data management practice; good data management planning and practice throughout a workflow results in better provenance information.
- It may be easy to install a provenance management system, to implement it and have it be useful, but more extensive work is needed across an organisation, such as evaluating existing provenance maturity (Taylor et al., 2015), adapting

18 Australian Research Data Provenance Interest Group: https://sites.google.com/site/rdpinoz/
19 Provenance Use Case: http://promsns.org/uc/
20 Notes from the Provenance Workshop at eRA 2014: https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxyZHBpbm96fGd4OjdjYjkzOTEwNjAxYzdlZmE
21 RD-Alliance Research Data Provenance Interest Group: https://rd-alliance.org/groups/research-data-provenance.html

 doi:10.2218/ijdc.v11i1.416

or designing a new workflow (for both system and human users) to capture provenance, and adopting persistent identifiers for all of the components to which the provenance will refer.

- Provenance may require social adjustment in some cases, as not all researchers embrace the importance of publishing provenance information; some were even negative about exposing any form of provenance that would reveal how their research workflow was undertaken, as it was felt that this would diminish their competitive advantage in the field or affect future research grant applications.

Provenance Target

One of the issues that became obvious over the course of this activity was varying understandings of the term 'publishing provenance.' In particular, key considerations include the audiences for the provenance information and the purpose for accessing the information.

If the audience is a machine process, the purpose is mainly for resource identification and easier reasoning, negotiation and processing. In this case, the provenance information is probably a link to a network/graph encoded in RDF or the equivalent.

If the audience is a human, the purpose is probably further exploration of the data to ascertain how trustworthy it is, or whether its lineage can provide pointers to other relevant datasets. In this case, the provenance information will need to be presented in a way that can easily be interpreted by a naïve user, and at the right granularity according to a user's need at the time (Davidson and Freire, 2008); there has been little work done on this so far by provenance researchers, and more is needed if provenance is going to contribute to users' increased trust in the data. This is something that ANDS is currently exploring in a set of Trusted Research Outputs projects, working with a number of different Virtual Laboratory projects jointly funded by NeCTAR, some of which are the projects who were part of the eRIC funding round.

Planned Future Activity

Over 2016, ANDS plans to continue its facilitation of an Australian provenance community. We will make a closer comparison of RIF-CS and other schemas that are used to capture provenance metadata by our partners with PROV-O, and identify if there are any domain specific concepts that could be used to extend PROV-O. We will also work with the community to explore what publishing provenance might mean, and in particular look out for research on provenance reporting or visualisation in a form that is readable and understandable by human users.

Acknowledgements

We would like to thank all our partners and project members for their contributions to the eResearch Infrastructure Connectivity (eRIC) program.

Reference

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 (PROMS). In Piantadosi, J., Anderssen, R. S., and Boland, J. (Eds), 20th International Congress on Modelling and Simulation (MODSIM), Adelaide, Australia, December 2013.

Davidson, S. & Freire, J. (2008). Provenance and Scientific Workflows: Challenges and Opportunities. In ACM SIGMOD, pp. 1345-1350, Vancouver, BC, Canada.

Di, L., Shao, Y. & Kang, L. (2013). Implementation of geospatial data provenance in a web service workflow environment with ISO 19115 and ISO 19115-2 lineage model. *IEEE Transactions of Geoscience and Remote Sensing, 51*(11). doi:10.1109/TGRS.2013.2248740

Doak, T., Wu, L.S., & Ganote, C. (2014). Galaxy for data provenance. Presented at the Indiana University Bioinformatics Clinic. Retrieved from http://hdl.handle.net/2022/18502

Feng, C.C. (2013). Mapping geospatial metadata to open provenance model. *IEEE Transactions on Geoscience and Remote Sensing, 51*(11). doi:10.1109/TGRS.2013.2252181

Goecks, J., Nekruternko, A., Taylor, J., & The Galaxy Team. (2010). Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life science. *Genome Biology, 11*(8). doi:10.1186/gb-2010-11-8-r86

Groth P. & Moreau L. (2013). *PROV-Overview: An overview of the PROV family of documents*. Retrieved from http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

Hallgren, W., Beaumont, L., Bowness, A., et al. (2016). The biodiversity and climate change virtual laboratory:Where ecology meets big data. *Environmental Modelling and Software, 76*. doi:10.1016/j.envsoft.2015.10.025

ISO. (2009). *ISO 19115-2: Geographic information – metadata – part 2: Extensions for imagery and gridded data.* ISO19115-2 Standard. Retrieved from http://www.iso.org/iso/catalogue_detail.htm?csnumber=39229

ISO. (2010). *ISO 2146:2010-Information and documentation – Registry services for libraries and related organizations.* Retrieved from http://www.iso.org/iso/catalogue_detail.htm?csnumber=44936

Lebo, T., Sahoo, S., & McGuinness, D. (2013). *PROV-O: The PROV Ontology.* Retrieved from http://www.w3.org/TR/prov-o/

Simmhan, Y.L., Plale, B., & Ganno, D. (2005). A survey of data provenance in e-Science. *ACM SIGMOD Record, 34*(3), pp 31-36.

Taylor, K., Woodcock, R., Cuddy, S., Thew P., & Lemon, D. (2015). A provenance maturity model. Environmental Software Systems. *Infrastructures, Services and Applications – IFIP Advances in Information and Communication Technology, 448*, pp 1-18. doi:10.1007/978-3-319-15994-2_1

W3C. (2010). *Provenance XG final report.* Report of the W3C. Retrieved from http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/

W3C. (2013). *W3C Working Group note: Dublin Core to PROV mapping.* Report of the W3C. Retrieved from http://www.w3.org/TR/2013/NOTE-prov-dc-20130430/

Wise, C., Car, N. J., Fraser, R., & Squire, G. (2015). Standard provenance reporting and scientific software management in virtual laboratories. In Proceedings of 21st International Congress on Modelling and Simulation, Gold Coast, Australia.

Wu, M., & Treloar, A. (2015). Metadata in research Australia and the open provenance model: A proposed mapping. In Proceedings of 21st International Congress on Modelling and Simulation, Gold Coast, Australia.

Wu, M., Kethers, S., & Treloar, A. (2013). Getting from managed to reused: making it easier for researchers to do something useful with data. In Proceedings of seventh eResearch Australasia Conference, Brisbane, Australiasia. Retrieved from https://eresearchau.files.wordpress.com/2013/08/eresau2013_submission_49-2.pdf