21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

Metadata in Research Data Australia and the Open Provenance Model: A Proposed Mapping

Mingfang Wu and **Andrew Treloar**

Australian National Data Service
Email: mingfang.wu@ands.org.au, Andrew.treloar@ands.org.au

Abstract: Research Data Australia (RDA) is a flagship data discovery service provided by Australian National Data Service (ANDS). RDA consists of two major components: a collection registry component that harvests metadata from a range of data providers and a data discovery component that makes metadata visible and searchable over the web. Metadata maintained at RDA are encoded according to the RIF-CS (Registry Interchange Format - Collections and Services) schema. RIF-CS is based on an international standard information model ISO2146:2010 that contains descriptive and administrative metadata for collections and related services, parties (people and organisations) and activities, and also supports the expression of relationships between those entities.

In the past, metadata ingested to the RDA registry have had a focus on metadata describing data collections once created, and less on the data provenance describing how a data collection came to be. As data provenance has been gaining increasing importance in areas such as data intensive research and policy making, more and more data providers intend to capture and publish data provenance information in order to enhance data's trustfulness and reproducibility. ANDS partners are at different maturity stages in implementing systems to capture, represent and publish data provenance information. In terms of provenance representation, some data providers may conform to the W3C recommended provenance data model (PROV-DM), some may use a discipline specific data model which may or may not conform to a community endorsed standard, and others may use just free text.

This paper compares the RIF-CS data model with PROV-DM and suggests a mapping between RIF-CS schema and the W3C recommended provenance ontology (PROV-O). The paper also discusses how one might go about linking provenance to RIF-CS records. This work will help to derive PROV data from contextual rich RIF-CS records and thus promote awareness and adoption of data provenance by ANDS partners.

Keywords: *Provenance, metadata, RIF-CS, PROV-O, Research Data Australia (RDA)*

1. INTRODUCTION

Provenance, also known as “lineage” or “pedigree”, is a term that has traditionally been used in the context of art history to document the history of an artwork or in the context of digital libraries, to document a digital object’s life cycle. In a similar way, data provenance is used to document where a piece of data comes from and the process and methodology by which it is produced. 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 et al., 2005).

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 back in time, but also with the relationships between a data product and other entities that enable the creation of the data as well.

According to the above definition, provenance is a kind of metadata. In this paper, we will differentiate it from descriptive metadata that describe properties and attributes of a resource, examples of descriptive metadata include the domain-generic Dublin Core, and the domain-specific ISO19115 (for describing Geographic information and services) and FITS (Flexible Image Transport System for encoding astronomical data). Although descriptive metadata (we will simply refer it as metadata in this paper) and provenance may have overlaps, for example when a metadata record states creation date of a data collection or describes a deriving relationship, they are intended for different uses in general. Descriptive metadata focuses on data and data properties, where its primary purpose is to show to a user what the data is about. While provenance is a record that describes entities and processes involved, it means not just the history back in time of a collection, but the relationships between that collection and other entities that have influenced its history. Provenance answers questions about where data originated, how data are produced, and who has been involved in producing them.

Metadata and provenance can complement each other. When a user discovers a seemingly relevant data collection from its metadata, the user may want to see its provenance to build confidence of using the data. This leads to a requirement to couple metadata and provenance. In this paper, we compares the Registry Interchange Format – Collections and services (RIF-CS) data model with PROV-DM and suggests a mapping between RIF-CS schema and the W3C recommended provenance ontology (PROV-O). The paper also discusses how one might go about linking provenance to RIF-CS records. This work will help to derive PROV data from contextual rich RIF-CS records and thus promote awareness and adoption of data provenance by ANDS partners.

This paper is structured as follows. Section 2 describes a data discovery portal that is run by Australian National Data Service (ANDS) and why ANDS cares about provenance. Section 3 first gives a short description of metadata schema RIF-CS used by ANDS data portal and provenance standards, and then presents a mapping between the two. Section 4 provides use cases of linking provenance to RIF-CS and a proposed simple solution. Section 5 discusses limitations and future work.

2. ANDS AND PROVENANCE

The Australian Commonwealth Government funded ANDS in January 2009, with the aim to transform Australia’s research data environment to enable Australian researchers to reuse research data more often. To achieve this goal, 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 richer context by interlinking data collection to researchers, their projects and software to make data well collected and add value to data, 3) provide Research Data Australia (RDA) data registry for ANDS partners to register their data and the RDA data portal to publish and make data discoverable. And all the above are for the purpose that the data that leads to a scientific finding or publication can be trusted and verified, data that is used once can be re-purposed and reused trustfully in more research. For that data provenance plays a crucial rule.

ANDS has also partnered with other national eResearch initiatives such as National eResearch Collaboration Tools and Resources (NeCTAR) and Research Data Storage (RDS) to fund a number of projects that provide better data services and tools to data users and to ensure better connections between data-focused compute services, data storages and data descriptions made available through RDA. As Figure 1 shows, researchers

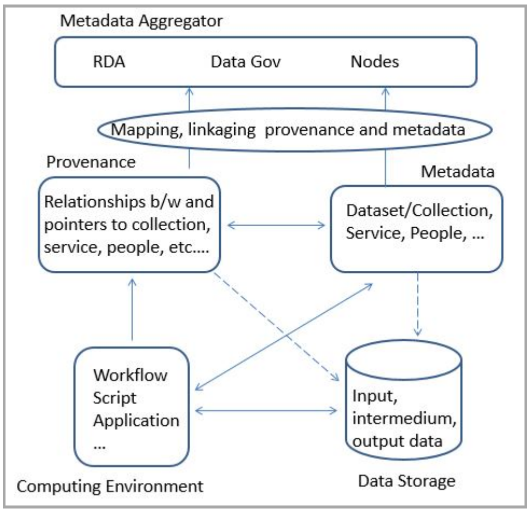

Metadata Aggregator
RDA
Data Gov
Nodes
Mapping, linkaging provenance and metadata
Provenance
Metadata
Relationships b/w and pointers to collection, service, people, etc....
Dataset/Collection, Service, People, ...
Workflow
Script
Application
...
Input, intermedium, output data
Computing Environment
Data Storage

Figure 1. Capture and publish provenance

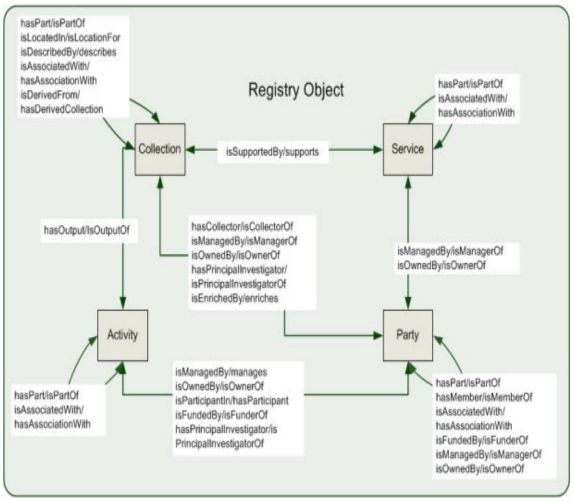


Figure 2. RIF-CS data/information model

can run a workflow for a specific experiment from a discipline specific virtual laboratory (VL) provided by a national computing service, experimental data (input and output) are pulled from and stored to a data storage. When a researcher runs a workflow, provenance and some metadata are captured (either automatically or manually). The resulting data collections, the process by which the data were generated and their relationship to used data resources are aggregated and published to RDA and other data portals for discovery.

3. RIF-CS AND PROVENANCE

ANDS has developed and maintains a web-based data discovery service RDA[1]. RDA stores and publishes metadata that are harvested from ANDS partners, with data stored as per arrangement by the partners.

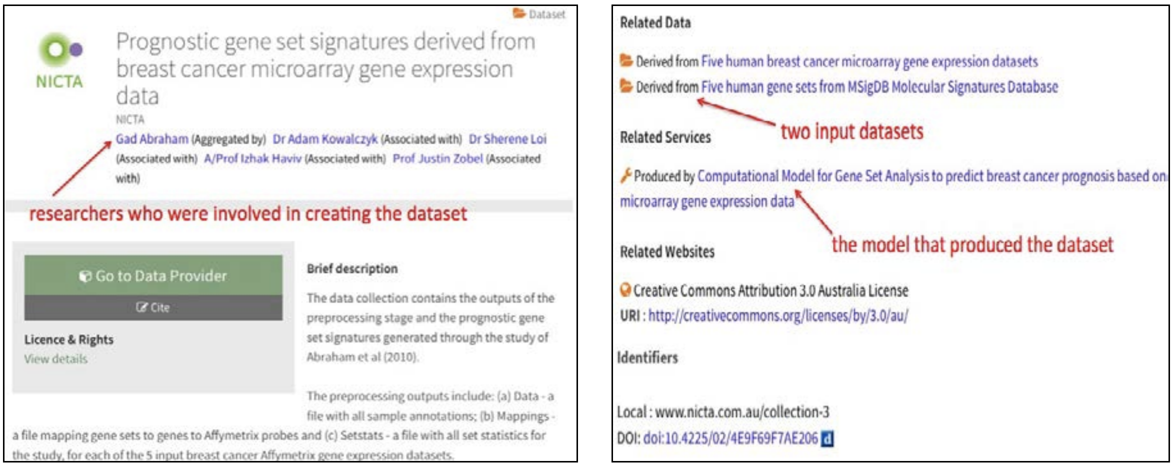
3.1. RIF-CS Schema

Metadata records registered and published in RDA follow the RIF-CS schema which is based on the ISO 2146:2010 Registry Services for Libraries and Related Organizations Standard (ISO 2146, 2010). The ISO 2146 information model has support for a federated registry service that contains descriptive and administrative metadata for the following four classes:

- Collection: an aggregation of physical or digital objects;
- Party: a person or group;
- Activity: something occurring over time that generates one or more outputs; and
- Service: Services support the creation or use of collections.

Each class is further subclassed. For example, Collection has subclasses: Collection, Dataset and Repository; and Party: Person and Group. The RIF-CS information model also supports the expression of relationships between those classes. As Figure 2 shows: Each class can be related to any number of other classs (zero, one or many). For example, a collection record may be related to several party records that describe researchers and/or institutions. The RIF-CS relation vocabulary has 42 values. 20 pairs of them are reciprocal relationships, for example, isProducedBy versus Produces. The Relation vocabulary can be expanded if there is a request and convincing use case from the user community. ANDS advises data providers to treat Collection as the first class object, i.e. to publish metadata of the other three classes only if they are linked to a Collection so that they provide context how the collection is created (through a relation to a Service) and who are involved in the data creation process (through a relation to a Party). ANDS is relaxing this data-central requirement, as more and more data providers want to describe their data tools and services in RDA.

[1] https://researchdata.ands.org.au/



Prognostic gene set signatures derived from breast cancer microarray gene expression data

NICTA

Gad Abraham (Aggregated by) Dr Adam Kowalczyk (Associated with) Dr Sherene Loi (Associated with) A/Prof Izhak Haviv (Associated with) Prof Justin Zobel (Associated with)

researchers who were involved in creating the dataset

Licence & Rights
View details

Brief description

The data collection contains the outputs of the preprocessing stage and the prognostic gene set signatures generated through the study of Abraham et al (2010).

The preprocessing outputs include: (a) Data - a file with all sample annotations; (b) Mappings - a file mapping gene sets to genes to Affymetrix probes and (c) Setstats - a file with all set statistics for the study, for each of the 5 input breast cancer Affymetrix gene expression datasets.

Related Data

Derived from Five human breast cancer microarray gene expression datasets
Derived from Five human gene sets from MSigDB Molecular Signatures Database

two input datasets

Related Services

Produced by Computational Model for Gene Set Analysis to predict breast cancer prognosis based on microarray gene expression data

the model that produced the dataset

Related Websites

Creative Commons Attribution 3.0 Australia License
URI : http://creativecommons.org/licenses/by/3.0/au/

Identifiers

Local : www.nicta.com.au/collection-3
DOI: doi:10.4225/02/4E9F69F7AE206

Figure 3. An example of linking resources involved in creating the dataset Prognostic Gene Set Signatures[2]

Figure 3 shows an example Collection record in RDA. The dataset "Prognostic gene set signatures derived from breast cancer microarray gene expression data" resulted from running a "Computational Model for Gene Set Analysis …" (Service), which took two input datasets (Collection>Dataset) "Five human breast cancer microarray gene expression datasets" and "Five human gene sets from MSigDB Molecular Signatures Database". Five researchers (Party) were involved in creating the dataset - one of them is the data collector. DOI link at the bottom of the record leads to a landing page that provides more metadata and provenance information than the record in RDA. For example, the landing page of the service record of the computational model describes the model in general, the steps that were taken and the software codes associated with each step.

3.2. Provenance Standards

ANDS partners are at different maturity levels of provenance management (Taylor et al., 2015). In terms of provenance representation, this may range from providing just a text file, or discipline Metadata such as Geographic Information Metadata ISO 19115-LE (ISO 19115, 2009), through to a highly semantic enabled provenance ontology (PROV-O). Regardless of different standards and recording granularity etc. for data provenance, there should exist a high level provenance information model and standard that combines vocabularies of different provenance standards and provides interoperability for provenance vocabularies in the linked-data sphere, in the way that generic Dublin Core does for other disciplined oriented metadata.

For that purpose, the W3C Provenance Working Group 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 et al., 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). DCMI Metadata Provenance Task Group also worked on and made available a mapping from generic metadata Dublin Core to PROV-O (W3C, 2013).

Based on PROV-DM - a generic data model and expressed in OWL2 Web Ontology Language, PROV-O provides a set of classes, properties, and restrictions that can be used to represent and interchange provenance information generated in different systems and under different contexts (Groth et al., 2013). The basic (or starting point) model of PROV-O has 3 classes (prov:Entity, prov:Agent and prov:Activity) and 9 properties. The basic model is expanded to include additional 7 elements and 14 properties. The extended model is not structurally different from the basic model, but adding some specification of subclasses and sub-properties of those in the basic model. For example, some extra classes Organization, Person and SoftwareAgent are subclasses of Agent; some extra properties describe versioning, influencing, invalidation or creation of entities, etc..

[2] https://researchdata.ands.org.au/prognostic-gene-set-gene-expression/11572/?refer_q=rows=15/sort=score%20desc/class=collection/p=1/q=Prognostic%20gene%20set%20signatures%20derived%20from%20breast%20cancer%20microarray%20gene%20expression%20data/

Table 1. Mapping between PROV-O and RIFCS

RIF-CS Class>Subclass	PROV-O Class>Subclass	RIF-CS Vocabulary	PROV-O Properties
Collection>Collection Collection>Dataset Collection>Registry Collection>CatalogueOrIndex Collection>Repository (Collection>SourceCode)	Entity>Collection Entity>Plan Entity>Bundle	**Collection - Collection** isDerivedFrom hasDerivedCollection isPartOf **Collection - Party** hasCollector isOwnedBy hasPricipalInvestigator isEnrichedBy **Collection - Service** isOperatedOnBy isProducedBy	**Entity - Entity** wasDerivedFrom hasPrimarySource hasMember **Entity - Party** wasAttributedTo **Entity - Activity** wasGeneratedBy
Service>Generate Service>Report Service>Annotate Service>Transform Service>Assemble Service>Create …	Activity	**Service - Service** isPartOf **Service - Collection** operatesOn **Service - Party** hasAssociationWith *dc:dateFrom* *dc:dateTo*	**Activity - Activity** wasInformedBy **Activity - Entity** used **Activity - Agent** wasAssociatedWith *startedAtTime* *endedAtTime*
Party>Person Party>Group	Agent>Person (foaf:Person) Agent>Organization (foaf:Organization) Agent>SoftwareAgent	**Party - Party** isPartOf isMemberOf	(org:memberOf) **Agent - Agent** actedOnBehalfOf

3.3. RIF-CS and PROV-O

RIF-CS schema as a high level metadata registry standard focuses on what a collection (or resource) is, as its primarily use is for human users to discover desired data collections. Yet, many RIF-CS components are actually provenance metadata; for example, the relation vocabulary answers who and how a collection is created. In addition, its Dates element adopts dates vocabulary from Dublin Core, which can record when a collection is created (dc:created), made available (dc:available), and published (dc:issued) etc..

On the other hand, PROV-O and data models focuses on expressing actions and resource states in a provenance chain rather than on resources and resources' attribute, as its primary use is for machine consumption. A software routine can be instrumented to capture actions and resource stages in a scientific computing context, for example, Galaxy and Kepler all support provenance recording and tracking.

Yet, the RIF-CS information model and PROV data model as well as their vocabularies have a substantial overlap; it is possible for data providers to generate a RIF-CS record from provenance chain and source metadata, or vice versa, to construct a provenance chain from RIF-CS records if all relationships are properly described. For example, the example presented in Figure 3 can be expressed in both models as shown in Figure 4.

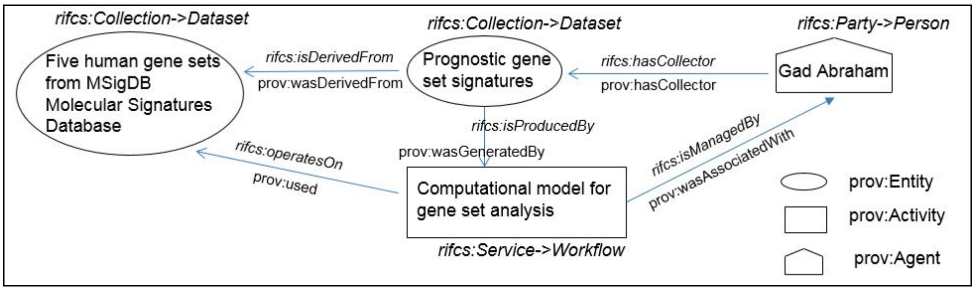


Figure 4. A mapping example of RIF-CS to/from PROV-O

Table 1 shows a mapping between RIF-CS and PROV-O. The table includes only those RIF-CS classes and vocabulary that can be directly mapped to corresponding PROV-O classes and properties. The RIF-CS classes and relation vocabulary offer finer granularity than their corresponding RROV-O classes and properties, thus leading to many to 1 mapping. For example the RIF-CS Collection class has 5 subclasses (Collection, Dataset, Registry, Repository and catalogueOrIndex), all of them can be mapped to prov:Entity or Entity subclass prov:Collection. In cases that a RIFCS class (Activity) or some vocabularies (e.g. a party isMemberOf another party) can not be properly mapped to PROV-O, they can be better mapped to linked open vocabularies such as FOAF[3] (e.g. rifcs:Activity>Project to foaf:Project) and ORG[4] (e.g. rifcs:isMemberOf to org:memberOf).

4. LINKING PROVENANCE TO RIF-CS

Although we can have a mapping between RIF-CS and PROV-O at a high level, each of them still has strengths for their original purpose as discussed above. The RIF-CS data model and schema are not designed for tracking provenance, and have the following limitations:

- Some data providers may not be able to publish metadata of all data collections and associated resources in a provenance chain to RDA. For example, a data provider in Australia owns a source collection and records provenance information about the source collection, a researcher in US uses the source data, combines it with other data and produces a derived dataset. The researcher also pings back to the data provider's provenance service and provides forward-links in the provenance chain. This use information would enhance the value and the trustworthiness of the source collection. In this case, the data provider can only be able to publish metadata of the source collection in RDA but not those forward-linked collections.
- Some data providers have set-up a Provenance Access and Query service (PROV-AQ), which enables construction of a whole picture of provenance chain that involves a collection (or a service etc); some data providers may go further to present this information in a way that suits the intended user's need whether via a visualisation or a textual report.
- Usually up-to-date provenance information is kept in the data provider's provenance management system.

To address the above limitations, we propose to link provenance information to a resource description (collection or service) in RDA. Specifically, we propose to expand the Related Information type vocabulary to include "provenance". Related information is primarily intended for resources located outside the ANDS Collections Registry and is described using a (preferably persistent) identifier. Its latest vocabulary type includes "publication", "website" and other objects (e.g. collections and services that data providers can't publish to RDA). Here we don't distinguish whether a URI is a provenance record or a result of a provenance query because they both return provenance information from a user's point of view, no matter how a URI is resolved.

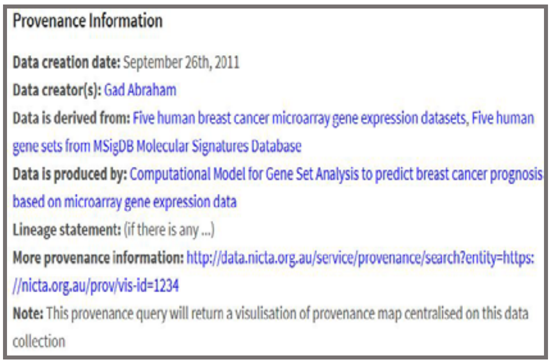
Data providers can specify the type of URIs in the Notes element if they think that is necessary. Figure 5 shows a snippet of this proposed RIF-CS element. Finally, all provenance information as described in a RIF-CS record could be gathered together and presented to a user as shown in Figure 6: from this page, a user gets a local view of provenance information of the collection, the user can follow the provenance link if the user would like to see the whole and up-to-date provenance trail of the collection.

[3] http://xmlns.com/foaf/spec/

[4] http://www.w3.org/TR/vocab-org/

```
<relatedInfo type="provenance">
   <title>Nicta Provenance Access and Query Service</title>
   <identifier type="uri">http://data.nicta.org.au/service/provenance/search?entity=https://nicta.org.au/prov/vis-id=XXX</identifier>
    <note>This provenance query results in a trail of provenance information of the collection.</note>
</relatedInfo>
```

Figure 5. An example of having relatedInfo of the type *provenance*



Provenance Information

Data creation date: September 26th, 2011

Data creator(s): Gad Abraham

Data is derived from: Five human breast cancer microarray gene expression datasets, Five human gene sets from MSigDB Molecular Signatures Database

Data is produced by: Computational Model for Gene Set Analysis to predict breast cancer prognosis based on microarray gene expression data

Lineage statement: (if there is any ...)

More provenance information: http://data.nicta.org.au/service/provenance/search?entity=https://nicta.org.au/prov/vis-id=1234

Note: This provenance query will return a visulisation of provenance map centralised on this data collection

Figure 6. A possible view of displaying provenance information

5. DISCUSSION

In this paper, we provide a mapping between RIF-CS and PROV-O. We also propose to link provenance from the RIF-CS by adding "provenance" to the "related information" type vocabulary. In our future work, we may need to further breakdown the type "provenance" into sub or specific types when such a requirement arises from our data providers. For now, most provenance access and query services record and provide provenance information in XML or other formats that are machine actionable but not (human) user friendly. As RDA primarily targets researchers to assist them with their data finding task, we need to discuss with our data providers how best to present provenance information should a user follow the provenance link. Future work may involve creating another presentation layer to post-process a provenance query to generate a human readable provenance report.

ACKNOWLEDGMENTS

The authors would like to thank the ANDS services director Adrain Burton and the Australian Research Data Provenance Interest Group[5] for helpful discussions.

REFERENCES

Feng, C. C. (2013). Mapping Geospatial metadata to open provenance model. IEEE Transactions on Geoscience and Remote Sensing. Vol.51(11). 5073-5081.

Groth P. and Moreau L. (2013). PROV-Overview: An overview of the PROV family of documents. [online] Available: http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

ISO 19115 (2009), ISO 19115-2: Geographic information - metadata - part 2: Extensions for imagery and gridded data. ISO19115-2 Standard.

ISO2146 (2010), ISO 2146:2010-Information and documentation --Registry services for libraries and related organizations.

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

Taylor, K. Woodcock, R., Cuddy, S., Thew P., and Lemon, D. (2015). A provenance maturity model. Environmental Software Systems. Infrastructures, Services and Applications - IFIP Advances in Information and Communication Technology. Vol.448. 1-18.

USGIN. Use of ISO Metadata Specifications to describe geoscience information resources. (2015, July) [online]. Available: http://lab.usgin.org/sites/default/files/profile/file/u4/USGIN_ISO_Metadata_1.1.4.pdf

W3C (2010, Dec.). Provenance XG Final Report (2010) [online]. Available: http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/

W3C (2013, April). W3C Working Group Note: Dublin Core to PROV Mapping. [online]. Available: http://www.w3.org/TR/2013/NOTE-prov-dc-20130430/

[5] http://www.ands.org.au/partner/provenance_interest_group.html