20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

A Governance Framework for Data Audit Trail creation in large multi-disciplinary projects

M.G. Hartcher

[a] *CSIRO Land and Water, Ecosciences Precinct, Brisbane, QLD. 4001*
Email: mick.hartcher@csiro.au

Abstract: The creation of data audit trails, within large multi-disciplinary projects in the CSIRO Water for Healthy Country (WfHC) Research Flagship, have relied heavily on the development of appropriate data management tools for creating metadata and data audit trails, coupled with human and technological processes. However, having the tools and processes alone does not provide an effective data management system so we have established a Governance Framework which provides the holistic function of integrating the various teams, technologies, and tools in an organized and focused way.

The Governance Framework creates an environment where there are clear objectives, roles and responsibilities for handling data management activities, protocols, procedures, and processes for delivering the components of data, metadata, and audit trails, as well as providing a means of assigning accountability and developing incentives that ensure that audit trails are completed with a high degree of confidence. While the Governance Framework forms a holistic function it does require the application of an efficient tool for cataloguing data and constructing data audit trails with provenance information.

A metadata tool was originally developed in 2007 by CSIRO Land and Water through the WfHC Flagship. The main need for the metadata tool was for capturing a data audit trail for a basin-wide multi-disciplinary project known as the Murray Darling Basin Sustainable Yields (MDBSY) project. The tool was later modified to accommodate additional Sustainable Yields (SY) projects between 2008 and 2010 and proved a valuable asset as a data auditing platform for such high-profile projects. The modified tool still had some limitations and did not fully conform to a metadata standard. In 2011 the tool, known as the Regional Water Data Management System (RWDMS), was rebuilt to meet the ANZLIC metadata standard and with improvements in functionality.

Based on a similar architecture to the original tool, the RWDMS has various new features such as metadata statistics, the ability to add new projects via the web interface, and import and export functionality. In addition, it now also provides the ability to generate audit trail lineage diagrams to depict the genesis of data throughout a project via parent-child relationships that are defined in a lineage field during metadata entry. The RWDMS is also able to export metadata in ANZLIC XML format, which can then be imported into the CSIRO Enterprise, Data Access Portal (DAP) for wider access. The tool includes a web-based user interface, a relational database for storing the metadata, and a metadata 'robot' which scans the data storage file system to detect new datasets placed there and then create blank records in the database, ready for manual metadata attribute population via the web interface.

The RWDMS has been effectively applied to the creation of data audit trails. It relies on an underlying project data archive file structure with clearly defined directories and a naming convention which flags datasets so they can be detected by the metadata 'robot' and then catalogued by the user. The parent-child linkages are then defined for each dataset to create an audit trail which can be visualized within the RWDMS interface.

The Governance Framework has created a vital operational environment for managing project data in the CSIRO, WfHC Flagship while the RWDMS provides a highly functional toolset for cataloguing data, for monitoring archive construction and data custodian input statistics, for reporting of project metadata statistics, for viewing audit trail diagrams, which depict project data genesis through parent-child linkages, and a process for publishing data products into the CSIRO Enterprise DAP, post-project, for wider data discovery and reuse. While there are still limitations in the capture of good quality provenance information, future developments in the capture of project workflows and automation of provenance capture will further enhance this system and provide further support to the development of a coherent data management culture.

Keywords: *Data cataloguing, data audit trails, data discovery, metadata*

1. INTRODUCTION

The need to have formalized data management systems has become increasingly prevalent in recent years and the tools for addressing this need are now being developed. Automated capture of metadata is still a difficult issue, mainly due to the nature of research workflows, i.e. research workflows can be complex, often partially manual, sometimes run once only and possibly on distributed, heterogeneous systems. Large multi-disciplinary projects have the requirement of generating audit trails, which represents the lineage of reported scientific results and are developed by populating metadata records with provenance information including the creation of parent-child linkages via linking dataset entries in a metadata catalogue for all datasets. They are called audit trails simply because they are established in order to provide a chain of evidence showing the scientific robustness that is used in generating reported scientific results, in the event that the results are challenged and an audit is conducted.

However, the tools and technology that have been developed to capture audit trails do not alone satisfy the task: they need to be implemented within a coherent and structured Governance Framework which provides the holistic function of seeing project teams efficiently utilize the technology and tools and applying the protocols, procedures, and processes that define the flow of information.

2. GOVERNANCE FRAMEWORK

A Governance Framework creates an environment where there are clear objectives, roles and responsibilities for handling data management activities, protocols, procedures, and processes for delivering the components of data, metadata, and audit trails, as well as a means of assigning accountability and developing incentives that ensure audit trails are completed with a high degree of confidence.

2.1. Roles and responsibilities

A key aspect in having a viable governance framework is the allocation of roles and responsibilities to staff. The approach taken in the CSIRO Murray Darling Basin Sustainable Yields (MDBSY) project was in separating out Data Management (DM) as a distinct project focus (Hartcher and Lemon, 2008). The DM team was comprised of a Team Leader, Data Manager, and Data Coordinators for each of the disciplinary project teams, i.e. Catchment Yield, Environmental Assessment, Groundwater Modelling and River Modelling. This provided a role with responsibility for each team's data archiving and cataloguing work.

2.2. Objectives

It is important that there are clear objectives for what is to be achieved. In the case of data audit trails it is necessary to outline the requirements for fulfilling a data audit including:

- All datasets (including modeling software and configuration settings, coding elements, etc.) used to produce final reported results must be archived
- All datasets must have a completed metadata record
- All datasets must have a demonstrated lineage

The project teams also need to be given archiving milestones which align with reporting milestones so they are prompted to deliver a data catalogue in stages rather than leaving archiving work to be done at the end of the project.

2.3. Protocols, procedures, and processes

The Governance Framework protocols outline the agreed set of terms, or governing rules, by which the DM activities are conducted. The protocols also provide details of the procedures to be followed for acquiring, exchanging, storing, securing, archiving, and cataloguing data and will also include definition of the processes, both human and technical, used to implement procedures.

2.4. Accountability

As there are specific roles and responsibilities associated with DM activities there is also accountability. Accountability is not about creating a target to blame if things go wrong for example when audit trails are not completed: it is more so there to provide a mechanism whereby someone can be held responsible for the actions promised in order to avoid having things go wrong in the first place. Failed delivery means that an audit trail is not able to be generated for a dataset, which is not acceptable, therefore by having someone

accountable for a part of an audit trail, the tasks involved in its generation can be associated with that person and followed through successfully with incentives to help drive this.

2.5. Incentives

Incentives have been a contentious issue in DM for quite some time. The 'big stick' approach, i.e. threat of punishment, has proven to be extremely flawed as it is an attempt to force people into behavior that they are not happy with and usually results in a low quality delivery, if any at all. The introduction of motivating factors such as having correctly populated data archives/catalogues seen as project deliverables themselves, just as writing project reports are, creates a more professional stimulus with improved results. In addition, the introduction of data citation metrics could create a professional measure relevant to science careers associated with performance and advancement, which can motivate more effectively than a punishment. This would be similar to publication metrics. Also, the creation of rewards for performance in DM activities themselves could add an additional incentive to motivate people to deliver both quantity and quality in dataset archiving and catalogue entries.

3. DATA MANAGEMENT APPROACH

The approach to data management we have taken for large multi-disciplinary projects relies on team structures that apply the technology coupled with appropriate tools. Figure 1 illustrates how these factors combine to create an effective flow of information in the data management system.

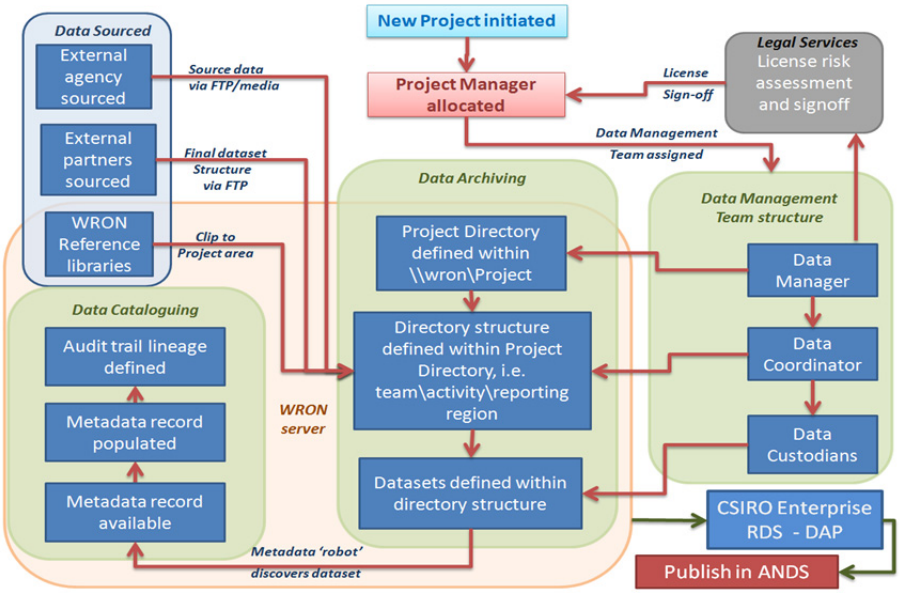

Figure 1. Data Management system

3.1. Technology

The underlying technology includes shared storage facilities, such as the Water Resource Observation Networks (WRON) server at CSIRO land and Water, Canberra, where a project data archive resides; the computer processing, often coupled to the main storage; the dataset transfer mechanisms, such as FTP gateways; the security systems and firewalls, which protect the data storage system. The underlying IT infrastructure also includes systems admin support.

3.2. Teams

The creation of a Data Management team within a project provides specific roles with assigned responsibilities to manage the creation of data archives, data catalogues, and data audit trails. A Data Manager organizes the establishment of the project archive, including security groups, facilitates the legal risk assessment process for data licensing, organizes the nomination of project team Data Coordinators, and provides protocols and procedures which govern the data management environment.

Data Coordinators within each project team organize their team's directory structure in consultation with the rest of the DM team, and manage the process of archiving and cataloguing data. Sometimes there is a single

Data Coordinator assigned across a project in which case there will also be a Data Custodian assigned within each project team. The Data Custodian role is assigned to ensure that all datasets are properly stored with the correct naming convention within the defined directory structure, and that they all have complete metadata in the catalogue as well as the audit trail linkages being defined via the lineage field in the metadata.

3.3. Tools

The software tools required are: a desktop FTP client, which provides a mechanism for data exchange with the shared storage facility to external partners; Sharepoint sites, for managing DM protocol and procedure documents; a metadata catalogue for cataloguing datasets and defining data audit trails. The metadata catalogue employed within the CSIRO Water for Healthy Country Flagship (WfHC) is the Regional Water Data Management System (RWDMS) and it has been purpose-built to generate data lineages or data audit trails, as they are now commonly known.

4. APPLYING THE RWDMS TO CREATE DATA AUDIT TRAILS

A metadata tool was originally developed in 2007 by CSIRO Land and Water through the Water for healthy Country (WfHC) Flagship. The main need for the metadata tool was for capturing a data audit trail for a basin-wide multi-disciplinary project known as the Murray -Darling Basin Sustainable Yields (MDBSY) project. The tool was later modified to accommodate additional Sustainable Yields (SY) projects between 2008 and 2010 and proved a valuable asset as a data auditing platform for such high-profile projects. The modified tool still had some limitations and did not fully conform to a metadata standard. In 2011 the metadata tool, called the Regional Water Data Management System (RWDMS), was rebuilt to meet the ANZLIC metadata standard along with some additional functionality. The RWDMS is now used to catalogue data and develop audit trails for multi-disciplinary WfHC research projects in CSIRO.

The creation of data audit trails relies upon the use of defined directory structures, file naming conventions for tagging datasets and the use of a cataloging tool. The RWDMS consists of a web-based graphical user interface (GUI) for manual metadata entry and searching; a relational database back-end, which stores the metadata content and a metadata 'robot', which is an automated script that discovers new datasets added to the project archive storage on the WRON server at CSIRO Land and Water, Black Mountain, Canberra. Users only see and interact with the web-based GUI, which can be opened in any web browser.

The RWDMS provides a catalogue of metadata records for individual datasets within projects. The metadata is tagged by project, project team and project reporting region to assist future project dataset searches. The metadata profile employed within RWDMS is based on the ANZLIC core metadata standard version 1.1 (ANZLIC, 2007) and includes all of the mandatory fields as well as some optional fields. The tool can be accessed by non-CSIRO staff via a Partner domain system so that external project partners can open the GUI in a web browser within their own organisation. The back end relational database, where the metadata content is stored, can only be accessed by a CSIRO data administrator.

The RWDMS has a range of functionality to ease and enhance the cataloguing of datasets through the automatic population of some metadata fields, the reporting of metadata statistics to support the monitoring of cataloguing activities and the generation of audit trail diagrams which graphically depict the lineage of all reported results. Also included is the ability to import and export raw metadata entries in ANZLIC-compliant XML; the ability to copy a metadata record from a pre-existing record and filtered search functions.

The RWDMS was built not only to address data archiving but also to provide a mechanism to define and visualise project data audit trails. This is carried out by defining all of the parent-child linkages between ancestor datasets within the *Lineage* field of the metadata record editing interface as illustrated in Figure 2.

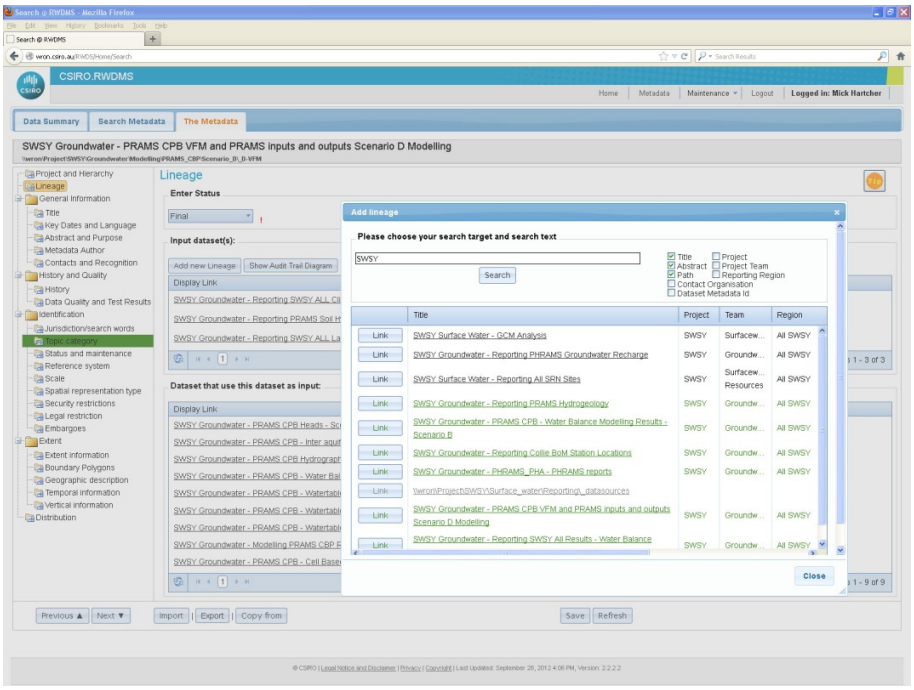

Figure 2. RWDMS lineage definition

4.1. Data archives

Defined directory structures are necessary in large multi-disciplinary projects in order to maintain order amongst the wide array of data, models, software code, and documents that are employed across numerous project teams and various reporting regions.

File and folder naming conventions are also critical within project directories in order to demarcate teams, reporting regions, models, and scenarios. In addition, the convention required by the metadata robot – the automated script which runs a scan of the project data storage volume to detect new datasets and changes to existing datasets – requires directory names begin with an underscore, e.g. *_DatasetName*, to denote that it is a dataset. Figure 3 illustrates a series of project directories for the CSIRO Sustainable Yields projects.

Figure 3. CSIRO Sustainable Yields project archives.

4.2. Defined lineage

As previously mentioned, the data audit trails are created by defining parent-child linkages in the *Lineage* field within the metadata entry for a dataset. All datasets that are used as part of the process for generating results that are published in a resulting report must have a metadata statement and must be included within the audit trail lineage via the parent-child linkages in the RWDMS. The data lineage for every report can then be visualized. This is useful to demonstrate the completeness of a particular report's audit trail, as well as a means to help monitor progress of data archiving, audit trail construction, and cataloguing activities throughout a project. Figure 4 depicts the data lineage for one published report, i.e. the Surat region report, from the Great Artesian Basin Water Resource Assessment (GABWRA) project. The GABWRA project produced 4 region reports, 4 summary reports for those regions, 7 technical reports, and 2 whole of basin reports. Each and every one of those reports has an associated data audit trail with a total of 291 datasets having been catalogued and linked within the RWDMS.

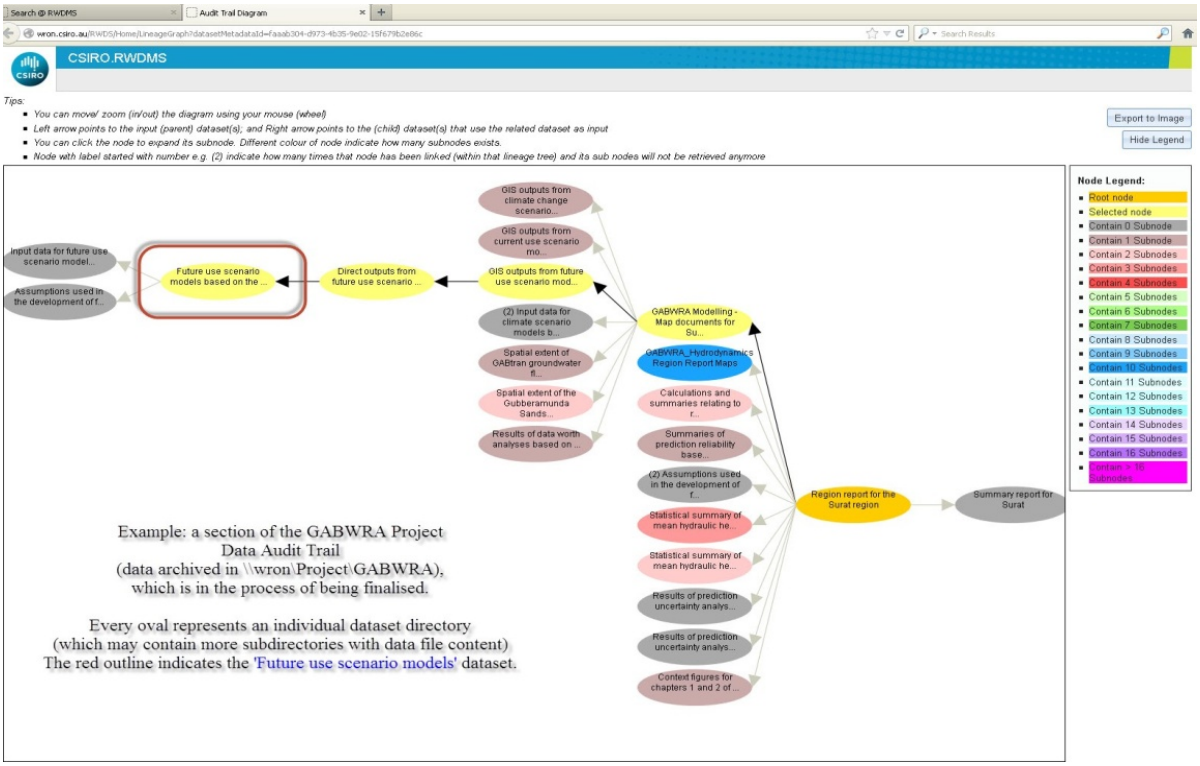

Figure 4. A Data Audit Trail diagram.

5. PROVENANCE CAPTURE

The RWDMS provides a mechanism to create data audit trails using the metadata records' *Lineage* field which can be used to illustrate the provenance of reported results. In addition, a *History* field is provided within ANZLIC metadata records within which provenance descriptions can be entered as free text. However this is a limitation in the data governance chain as it does not enforce any compliance to a standard measure provenance description thus quality is not assured. In some cases the content of this free text description has been filled with much detail, such as clear explanations of the steps used to create a dataset, but in other cases the content has been very poor having virtually no value at all as a provenance description. This is a key issue which needs to be addressed with further developments in the area of workflow provenance capture and automated provenance reporting systems.

6. DISCUSSION AND CONCLUSIONS

The Governance Framework that has been developed for managing data and creating data audit trails in large multi-disciplinary projects within the CSIRO WfHC Research Flagship has proven to be an effective one. It has provided an organized approach to data management with defined roles, clear objectives and accountability which has moved the data management culture away from one dependent upon individual methods to one that applies a more enterprise-based approach with a long term focus on data reuse, repeatability, and transparency for the methods of scientific analysis that yield project results. The tools and

technologies are continuing to be developed as are the protocols and processes, while an improved data management culture is developing, albeit slowly.

An organized Governance Framework creates the necessary platform to develop data audit trails for large multi-disciplinary projects which can be used to defend the methods used for generating scientific results. While these developments have proven to be a significant improvement on past practices there are still gaps in the generation of audit trails, such as the lack of automated provenance capture and barriers in the further development of a fully coherent data management culture. It is hoped that such limitations can be overcome by focusing future development on the quality of audit trail descriptions via automated provenance capture being reported directly into the RWDMS metadata *History* and *Lineage* fields. This would then provide a far more robust system for ensuring that audit trails are defined in totality and are truly and utterly defendable.

ACKNOWLEDGMENTS

The author would like to acknowledge the CSIRO Water for Healthy Country Research Flagship for funding the development of the RWDMS system and supporting its application in projects. I would also like to acknowledge the many strategic developments contributed by Dr David Lemon, especially in the establishment of data management teams in the Murray Darling Basin Sustainable Yields (MDBSY) project, which have provided the foundation on which to build the governance framework that has been established. I would also like to thank the original architect of the MDBSY metadata catalogue, Garry Swan, as well as Jamie Vleeshouwer and Arthur Read for their work in helping to develop the catalogue further. Thanks also go to Teddy Wijaja, Rohit Anantaram, and Garry Swan and Andrew Freebairn for their work in building the RWDMS and the ongoing debugging/enhancement of the tool. Finally, I would like to acknowledge the work of the many Data Coordinators across the various SY projects including (alphabetically) Jenet Austin, Heinz Buettikofer, Phil Davies, Trevor Dowling, Anne Henderson, Geoff Hodgson, Steve Marvanek, Linda Merrin, Gail Ransom, Andrew Taylor, and Chris 'Tunza' Turnadge.

REFERENCES

ANZLIC core metadata requirements, Version 1.1 (2007). http://spatial.gov.au/system/files/public/resources/anzlic/ANZLICmetadataProfile_v1-1_2007.pdf.

Hartcher, M.G., Lemon, D. (2008), Data Management for the Murray-Darling Basin Sustainable Yields Project – A report to the Australian Government from the CSIRO Murray-Darling Basin Sustainable Yields Project, November 2008. http://www.csiro.au/Organisation-Structure/Flagships/Water-for-a-Healthy-Country-Flagship/Sustainable-Yields-Projects/DataManagementMDBSY.aspx

Hartcher, M.G. and Lemon, D. (2009), Developing data audit trails for the CSIRO Sustainable Yields projects. In Anderssen, R.S., R.D. Braddock and L.T.H. Newham (eds) 18th World IMACS Congress and MODSIM09. Modelling and Simulation Society of Australia and New Zealand 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