18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Developing data audit trails for the CSIRO Sustainable Yields projects

Hartcher, M.G., Lemon, D. [1]

[1] *CSIRO Land and Water*
Email: Mick.Hartcher@csiro.au

Abstract: As data capture technology improves and computer platforms continue to increase in storage capacity, data management becomes increasingly important. A data management system helps ensure that data has appropriate security, is logically structured, and properly archived with demonstrable integrity. Over the past 2 years CSIRO has taken on some high-profile projects with a high expectation for auditing of final results by various interested parties. A key to the success of these projects has been the implementation of a data management system.

In early 2007, the National Water Commission (NWC) comissioned CSIRO to develop a whole of basin assessment of water availability for the Murray-Darling Basin. This became known as the Murray-Darling Basin Sustainable Yields (MDBSY) project. Following the successful completion of the MDBSY project, CSIRO was contracted to conduct three additional Sustainable Yields (SY) projects across Northern Australia, Tasmania, and south-west Western Australia. The system established for the MDBSY project formed a 'blueprint' for use in these projects.

The MDBSY and subsequent SY projects have involved large volumes of data across various disciplinary project teams. The high public profile and associated level of scrutiny of the results, presented in the final reports, required the establishment of a data audit trail for all project results, so that all values could be traced to their origins. In order to achieve this, a high degree of discipline was required when managing project data. The data management framework developed to support these projects was built on a set of protocols, processes, and standards.

The SY projects have been structured with four disciplinary teams as well as a team focused on data management. The data management team was comprised of a data team leader, a project data manager, and a data coordinator from each of the disciplinary teams. The data team leader and data manager were responsible for developing the protocols and procedures, and for organising the development/acquisition of tools required for organising and managing all of the project data. The key responsibility for the data coordinators was to ensure that their team's data were migrated into the project archive and catalogued appropriately with metadata descriptions.

This paper will outline the development of the data management protocols, with emphasis on the establishment of a data audit trail, and associated tools. It will also highlight how the developments within the SY projects have provided a framework for propagating such a system across all future projects, and how the 'seeds' of a data management culture have begun to 'grow' within CSIRO. Finally it will share some of the learnings from this activity.

Keywords: *Data management, audit trails, protocols, procedures.*

1. INTRODUCTION

As data capture technology improves and computer platforms continue to increase in storage capacity, data management becomes increasingly important. A data management system helps ensure that data has appropriate security, is logically structured, and properly archived with demonstrable integrity. Over the past 2 years CSIRO has taken on some high-profile projects focused on assessing water resources to provide a scientific basis for the federal government to develop water resources policy. These projects have had a high expectation for auditing of final results by various interested parties and have involved multi-disciplinary teams who have each employed and generated large volumes of data.

In early 2007 CSIRO embarked on a complete assessment of water availability for the Murray-Darling basin. This involved modelling historic and future climate scenarios, groundwater and surface water modelling, and environmental analysis of significant wetlands. The project was known as the Murray-Darling Basin Sustainable Yields (MDBSY) project. The results of this work were intended to be used by the National Water Commission (NWC) as a knowledge baseline for setting targets for water usage in the basin. Following the successful completion of this project, three more Sustainable Yields (SY) projects were initiated to cover northern Australia, south-west Western Australia, and Tasmania.

The MDBSY project was a huge and diverse undertaking requiring input from a large number of people from within CSIRO and from various external organisations. The volume and diversity of data, models and reports used or generated within the project, along with the tight timeframes for delivery, required a professional approach to data management. To this end, data management was undertaken as a separate component of the project with close linkages to other teams (Hartcher and Lemon, 2008). The key goals were to ensure that data used or generated within the project was:

- accessible to those who needed it
- safe from being lost or corrupted
- managed according to requirements of data suppliers
- secure from those who did not need it
- endowed with demonstrable integrity

In particular it was essential that it could be demonstrated how each individual dataset within the project was produced and where it came from, in order to provide an audit trail.

By having a distinct data management team, it was possible to establish common protocols across each of the disciplinary teams, as well as establishing a common data repository and a robust set of procedures for managing the data store. These protocols covered data storage, access, security, backup, and archiving and ensured that the integrity of both datasets and documents was, and remains, demonstrable. Protocols for exchanging data with external agencies and sub-contractors, sharing project documents with project team members and disseminating information were also developed and administered. This required taking into consideration various confidentiality and restricted access requirements associated with some data.

Data Management team, consisting of a team leader, a project data manager, the project team data coordinatorsm and some additional data management support staff, was responsible for ensuring a complete audit trail existed for all modelling results, original and interim datasets, software versions, and reports. These were archived in the project data repository, with metadata statements completed and stored within a relational database (Hartcher and Lemon, 2008). The Data Management team took responsibility for:

- providing secure centralised computing facilities (including data storage and processing);
- providing project collaboration tools (including the project SharePoint website, project data catalogue, and data exchange facilities);
- development of a project reporting database;
- ensuring all data collected for the project was appropriately described and licensed;
- ensuring a full audit trail of all steps of the analysis process was captured; and
- ensuring commitments made to third parties with respect to data and models were fulfilled.

The data management protocols, procedures and tools developed for the MDBSY then became a valuable blueprint for the three subsequent SY projects, which are currently in progress. This paper outlines the protocols, the various infrastructure utilised, the tools used for managing data, the development of an audit trail, and the issues of changing an existing data management culture.

2. DATA MANAGEMENT PROTOCOLS, PROCEDURES AND STANDARDS

2.1. Protocols and Procedures

A set of protocols and procedures were developed for management of data within the MDBSY project. Key amongst these was to have project data archived separately from working space, in order to maintain data integrity. The archive provided an organised set of data directories which enabled the development of a structured audit trail for all project outputs. This project archive was only accessible to CSIRO project staff. (Hartcher and Lemon, 2008).

The project archive contained directories for the 18 MDBSY reporting regions, an additional folder for the Snowy region, which was included in the Murrumbidgee region, and another directory for whole-of-MDB data. Each reporting region contained a directory for each project team. The structure within each discipline team's directories varied depending on the tasks being performed, and/or the number of models being applied to the reporting region. Figure 1 illustrates the structure of the project archive directories (Hartcher and Lemon, 2008).

An important aspect of this structure was that individual datasets were contained within their own directory. Each dataset directory was prefixed with an underscore (eg. '*_datasetname'*), allowing metadata tools to identify a new dataset (Hartcher and Lemon, 2008). In some cases, an individual dataset contained thousands of files and it was not deemed necessary or efficient, to describe (via a metadata statement) each data file stored. Therefore metadata statements described the dataset as a whole, encompassing all files contributing to that data set (Hartcher and Lemon, 2008).

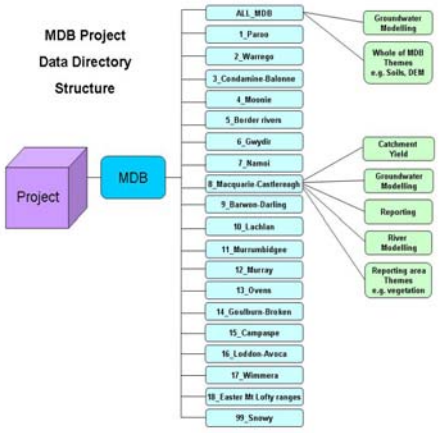

Figure 1. MDBSY project directory structure

Personal work space was provided in separate locations to the project archive directory. This provided space for project team staff to develop data, carry out model runs, create maps, develop report spreadsheets, and to write documents. All final datasets, model runs, etc., were required to be moved from the workspace into the appropriate directory within the project archive, with a metadata statement completed for each dataset (Hartcher and Lemon, 2008).

Core datasets were stored within data volumes allocated for GIS, remote sensing, and time series. Data stored within the project archives could then be migrated, at a later date, into the core data volumes for ongoing use as reference datasets. A series of checks or filters would be applied prior to this data migration, e.g. spatial data must have a reference system defined and a fully populated metadata statement (Hartcher and Lemon, 2008).

Some modelling work required very large volumes of storage in order to run multiple scenarios using, and generating, many thousands of files. A separate working volume was therefore created, and much of the river modelling and catchment yield rainfall runoff modelling work was performed in this space. These datasets were designed to nest within the project archive structure so that the task of transferring the data into the archive directory, once the modelling was finished, would be simplified (Hartcher and Lemon, 2008).

In the MDBSY project each of the key discipline teams (Catchment Yield, River Modelling, Groundwater, Water Accounting and Environmental Assessment) had a data coordinator nominated. It was the data coordinator's role to ensure that the data management protocols were being applied within their team. The data coordinators were data custodians for their respective teams, and had responsibility for ensuring that all project data, software, code, maps, and report elements, were archived in an appropriate location within the project archive. The MDBSY Project data management team included a team leader, a project data manager, the project team data coordinators, and some additional data management support staff (Hartcher and Lemon, 2008).

2.2. Data Security

Data security was an important issue for the project. A number of datasets could only be used within the project if CSIRO could guarantee restricted access. The large storage volumes associated with the project inhibited the use of folder-level permissions for data security. A set of permission 'groups' were therefore created in order to manage the security which allowed specific access within the MDBSY project archive.

Access to the data store was only available to CSIRO team members. The permission groups were created to control data updates, editing, directory structure, and versioning. Read access to project data was provided to all CSIRO project staff. Higher level permissions required specific approval before being granted. There were four permission group levels: Administrators - with full administrative control; Editors - able to create/delete folders and files but not change permissions and ownership; Contributors - able to read/write/create and modify files, but not delete or create folders; and Readers - able to read data and execute software (Hartcher and Lemon, 2008).

2.3. Conventions

While disciplinary teams had specific requirements for data standards, relating to the models being applied, some common data standards were also established across all of the disciplinary teams. Where some standard software products were employed, it was necessary to ensure that such software could access and read data. A key standard was that directory names could not have spaces and so an underscore was used instead, e.g. 02_Warrego. Importantly, some teams were creating data as inputs for other teams. Adherence to common data formats and coordinate systems were necessary in these cases (Hartcher and Lemon, 2008).

Naming conventions were mostly dependent upon requirements for model inputs within each project team. There was adherence to a naming convention required for the reporting elements, so that the reporting team could determine which elements were encompassed within a particular Excel workbook or Microsfot Word text document. This naming convention encompassed the reporting region, the report chapter, the element number for the report, and the version of the element to account for updates if they occurred. For example '02_SW3_37_v10.xls' refers to the surface water results for the Warrego region for elements between 3 and 37, with this being version 10 of the results. The version reference was critical as there were often small changes made to spreadsheets by reporting team members, as well as version updates being requested from project teams (Hartcher and Lemon, 2008).

3. ESTABLISHING AN AUDIT TRAIL

The data audit trail is an important aspect of the SY projects. The MDBSY project reports in particular have been scrutinised by various organisations, state and federal governments, and by the general public. This was expected when the MDBSY began, hence all project results were made traceable via a documented lineage back through the various modelling and analysis stages to the original data inputs and tools. This lineage was described as the audit trail.

3.1. Metadata Catalogue

It was necessary that all data components, ranging from original inputs through to final products, were fully documented with metadata and legally covered by data license agreements where appropriate. Furthermore, it was identified that elements of the project may need to be regenerated to reproduce results and/or be rerun in the future with additional data. It was therefore necessary that any result could be regenerated exactly as it was originally produced with no variation in modelling outputs, and that it was possible to verify the inputs to models and the methods and parameters used.

Therefore all project data, modelling software, model parameters, results, and reports were archived within the project archive directory structure and documented within the metadata catalogue. The metadata cataloguing tool developed for the project required that the lineage of all datasets be described, and that any datasets used as inputs for the development of another dataset be listed within the metadata statement.

The approach taken was to firstly establish that the key modelling datasets for each reporting region were archived. In many cases these were fairly obvious, although project team members responsible for running the models had to be consulted to ensure that all relevant files were included. This process was carried out by using the metadata catalogue to search the project archive for those models known to have been run in each region. In many cases this highlighted gaps in the archive which required further data files to be transferred

from workspace locations. In addition, some post-processing datasets were uncovered which also were subsequently archived, e.g. IQQM post-processing data.

The quality of metadata statements was also checked to ensure adequate detail had been supplied. Quality control of metadata was difficult to enforce as it was often unclear if the detail on a specific dataset was enough to describe all files associated with that dataset. Data lineage quality was less difficult to measure as it was in most cases obvious which datasets were inputs to or derived from another dataset.

Given the large volume of datasets involved in the project, a process of randomly selecting metadata statements from key datasets for each region provided an efficient basis for checking the quality of metadata entries. Feedback to metadata custodians also allowed them to make updates to a range of statements which further improved the quality (Hartcher and Lemon, 2008).

3.2. Reporting Database

A relational database was used to support the generation of reporting elements required for each reporting region. Output data generated for each modelling team was automatically imported from the project archive through the use of the metadata catalogue. This demonstrated that the audit trail was capable of tracing original datasets, and processes applied to the datasets to generate the reported values.

Each disciplinary team was to be responsible for determining the point in their processing at which outputs would be loaded into the database. The role of the data management team was to:

- develop and implement a single data model to hold this data;
- develop tools to load the database from data files provided by disciplinary teams;
- develop query tools for the generation of final indicators for reporting;
- and implement report templates for generation of reports.

The intention was that the final reporting elements (tables and charts) would be directly produced from the database. This would limit the amount of reformatting required to create the final released documents.

An important aspect of the reporting database was that it would form a vital link in the data audit trail. That is, through the database, it was to be possible to directly link reported results to original input data. This would be achieved through capture of some of the final analysis steps in code (queries) as well as storage of links to input data files.

The data audit trail could then be constructed in the following way:

1. Reported results come directly from the reporting database and were generated through stored code.
2. The database also contains a link to the original data input file for every piece of information stored. These files must be stored within the project directory structure.
3. As these files are stored on the project data directory, they were required to have an associated metadata statement.
4. Metadata statements include links to datasets used to create the described dataset.

These input datasets were to be stored within the project directory structure and hence also required metadata statements.

Development of the reporting database proved ambitious given the project timeframes. The river modelling and assessment teams were the only teams to participate in the experiment which had some good early results. A simple data model built around model results was developed, data loaders were built and deployed, data for a number of reporting regions were loaded and early products were delivered (Hartcher and Lemon, 2008).

4. CULTURAL CHANGE

The data management approach taken in the MDBSY project was a significant change to the existing modus operandi. While processes and protocols were defined and roles and responsibilities assigned to project teams, it was difficult to enforce the requirements and to ensure the quality of documentation. In short, a level of cultural change was being attempted during a very difficult project within narrow timelines.

While this was an ambitious undertaking, it may have also been a good opportunity to begin establishing long-term practices or at least principles to guide data management in future projects. Cultural change in any organisation can be a difficult challenge and there appears to be no single route to success. However, there are ways to identify the existing culture and possible steps to enacting change.

4.1. Steps to Cultural Change

Although culture change driven from the grassroots level, whether by an individual employee or a team, cannot be accomplished without support from management, the catalyst is clearly located within the rank and file; the momentum spreads through the organisation from the bottom up. The critical difference between a bottom-up change process and the more conventional top-down approach is lodged in the sharing of responsibility and power between management and grassroots leaders. Top-down companies should adopt top-down culture change strategies; lateral organisations should adopt grassroots strategies; and safety professionals should take the lead where other driving forces are missing. Hybrid organisations, of course, should be guided towards a mix of strategies (Simon, 2001).

The comments by Simon above offer some general advice to enacting cultural change. He also suggests a sequence of steps for enacting change namely, Step 1 - Make the case for change; Step 2 - Establish the vision; Step 3 - Assess the current culture; Step 4 - Develop a strategic plan; Step 5 - Use teams (grassroots and leadership) to implement the plan; Step 6 - Monitor and make course corrections.

These steps are generalised and, of course, require details applicable to the particular organisation. It is worthwhile to examine how such steps may be applied to changing an existing data management culture such as, Step 1 - Conduct a risk assessment of data management based on future requirements for data stores; Step 2 - The need for cultural change in data management is recognised and communicated from the upper management with a clearly defined vision of the characteristics and value such an environment would provide; Step 3 - Review the current state of data organisation and documentation and identify the existing approaches taken by all users; Step 4 - Create an implementation plan which defines how changes will be implemented among the system users and include data management in project planning, i.e. develop data management plans at the start of each project; Step 5 - Provide resources to staff to enable effective data management, i.e. recognised as part of role so that effort can be logged for data management work and charged against projects, and create incentives such as rewards and acknowledgements; Step 6 - Included in performance based measures and annual reviews, as well as creating incentives such as rewards and acknowledgements.

The steps were not undertaken within the MDBSY nor the three ongoing SY projects as cultural change needs to be enacted at an organisational level and this is well beyond the terms of reference for SY projects as well as requiring additional resources not available within the strict timeframes of the SY projects.

5. DISCUSSION AND CONCLUSIONS

The MDBSY project presented significant challenges in storing, processing, and documenting large volumes of data. The methods and tools that were developed to address these challenges provided support to project staff and a common frame of reference for handling the range of data and models acquired for, and developed by, the disciplinary teams. The requirements for developing audit trails imposed additional, and somewhat stringent, data management tasks on disciplinary teams. These tasks and the associated responsibilities had not previously been imposed across such a large project in a consistent and formalised framework. As a result it was difficult to ensure compliance.

The data management framework which was developed for MDBSY has also been applied to the three SY extension project, with very little change to the original framework. These projects are still ongoing, and their respective audit trails are still being constructed. Many of the staff who worked on the MDBSY are also working on the SY extension projects including some who were directly involved in data management. It appears that there is a higher level of acceptance, although not necessarily comfort, with the requirements of data documentation and audit trails, and the MDBSY is constantly used as an example of what is required.

The greatest barrier to developing accurate metadata and having a complete audit trail is the lack of consistent culture for data management. It can be argued that this change should not be incurred during such high-profile projects, yet this may be the best time to begin planting the seeds for cultural change. It appears that some grassroots development has begun and that this may be used as a catalyst for long-term change. However, a top-down approach also needs to begin from upper management within CSIRO.

The top-down approach to cultural change needs to not only include sufficient resources to address data management needs, but must also incorporate formal requirements for the individual staff who will be performing the day-to-day tasks of data management. These requirements should include things such as statements of data management responsibility within annual performance appraisal documents, accurately defined tasks in project planning documents with project sign-off requiring mandatory project management tasks to be completed, and tangible measures of data management adherence. However, such 'big stick' approaches to policing data management alone may also stifle the required changes to culture and possibly reduce the quality of outputs such as metadata. These requirements should also be coupled with appropriate incentives such as recognition through rewards and performance goals for promotion.

The MDBSY project effectively exposed the existing data management culture within CSIRO. The three SY extension projects have utilised the MDBSY data management approach and have tried to improve upon the processes outlined in this paper. However, the existing data management culture within CSRIO requires significant change and currently provides a hurdle to ensuring that long-term data curation, data publishing, and data discovery are quality assured to the level of international standards.

ACKNOWLEDGMENTS

The Data Management component of the MDBSY project could not have been completed without the resourcefulness and commitment of the Data Management Team and Project Data Coordinators. Key contributions came from; Jenet Austin, Phoebe Carmody, Phil Davies, Trevor Dowling, Alex Dyce, Peter Dyce, Peter Fitch, Douglas Kerruish, Tegan Liston, Steve Marvanek, Arthur Read, Garry Swan, Brendan Speet, and Jamie Vleeshouwer. Thanks also to Matt Stenson and Jamie Vleeshouwer for reviewing this paper.

REFERENCES

Hartcher, M., Lemon, D. (2008), Data Management for the Murray-Darling Sustainable Yields project. A Report to the Australian Government from the CSIRO Murray-Darling Basin Sustainable Yields Project, CSIRO, Australia. 34pp..

Simon, S.I. (2001), Implementing Culture Change – Three Strategies, http://www.culture-change.com/3strategies.pdf