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

Implementation of a reporting workflow to maintain data lineage for major water resource modelling projects

Merrin, L.E.[1] and S.M. Cuddy[1]

[1.] CSIRO Land and Water GPO Box 1666 Canberra 2601
Email: linda.merrin@csiro.au

Abstract: In early 2007, the Australian Government commissioned the Murray-Darling Basin Sustainable Yields project to quantify current and likely future water resources throughout the Murray-Darling Basin. It was a large multi-disciplinary modelling project with more than 150 people contributing to the modelling and the production of over 80 public reports. Managing the design, authoring, revision and production of these reports produced over 10,000 files requiring 18.4 GB of storage. Contributions to the reports came from six discipline-based teams supported by coordination, data management and reporting teams.

In early 2008, the Council of Australian Governments (COAG) expanded this study to include three new sustainable yields projects, namely the Tasmania Sustainable Yields project, the South-West Western Australia Sustainable Yields project and the Northern Australia Sustainable Yields project. These new projects are also producing more than 80 reports, and large volumes of reporting data.

In the Murray-Darling Basin Sustainable Yields project the reporting team was responsible for managing the production of reports. This involved developing, implementing and managing a system for linking model results and text from each discipline-based team to their corresponding reporting and analysis in the reports. A reporting workflow was developed using the concept of elements (plots, tables, maps and diagrams) and text, which allowed for creation, and then continuous revision, throughout a formal three-stage review process, while maintaining data lineage.

Elements and text were handled separately in the workflow. Elements were organised in Microsoft Excel workbooks, with one worksheet per element. These workbooks also contained the raw outputs from the models and therefore allowed for the post-processing of results in an organised fashion while maintaining data lineage.

Text was written into a base Microsoft Word document, which contained all of the styles, headings, subheadings, and a list of elements and their associated captions. These Word documents were managed using Microsoft SharePoint web collaboration software. Using strict versioning and a check-in/check-out protocol, SharePoint was used to control access to the documents (commercial-in-confidence until government release date) and to alert the reporting team when new versions of documents were uploaded for incorporation into the workflow. Adobe Professional was used to prepare reports for publication.

All of the Excel and Word documents contained a version log at their beginning. In this log, the details of the changes made to each of the versions, the date and the person responsible, were recorded. This not only traced the development of the documents, but also revealed the stage of progress through the workflow.

This paper reports on the workflow described above. The workflow process and related tools allowed for the production of the reports to be successfully managed while maintaining data lineage. The use of familiar office tools (Word, Excel, Adobe Professional) ensured that the learning curve for authors and modellers was minor, and this allowed training to focus on the use of SharePoint and the need to adhere to formal protocols and conventions. This paper also identifies areas that would have benefited from more structured management. Particular reference is made to the synchronisation of the region and summary reports and to the production of maps, and how the three new sustainable yields projects are overcoming these issues. The decision to separate the modelling activities from reporting on those activities was an innovative one, and the paper reflects on some of the implications of this decision.

In conclusion, the reporting workflow processes developed for the Murray-Darling Basin Sustainable Yields project have proved to be effective, adaptable and scalable as illustrated by their adoption in the three new sustainable yields projects.

Keywords: *Data lineage, data provenance, data pedigree, audit trail, Murray-Darling Basin Sustainable Yields project*

1. INTRODUCTION

The Murray-Darling Basin Sustainable Yields project was commissioned by the Australian Government in early 2007 to quantify current and likely future water resources throughout the Murray-Darling Basin (Figure 1). It was a large, multi-disciplinary modelling project with more than 150 people from a number of government and non-government organisations forming six discipline-based modelling teams, supported by a coordination team, data management team and reporting team.

The Murray-Darling Basin was divided into 18 reporting regions. For each of the regions, a main report, summary report, and factsheet were produced. In addition to these region reports, a whole-of-basin report, a whole-of-basin summary report and a number of technical reports were also produced, resulting in the production of more than 80 public reports during the project. Each of the reports went through a formal review process, with the review process for the region reports consisting of three formal stages. Managing the design, authoring, revision and production of these reports produced over 10,000 files requiring 18.4 GB of storage.

In early 2008, the Council of Australian Governments (COAG) expanded this study to include all major Australian water systems. As a result, three subsequent sustainable yields projects, namely the Tasmania Sustainable Yields project, the South-West Western Australia Sustainable Yields project and the Northern Australia Sustainable Yields project (Figure 1) have been commissioned and are due for completion by the end of 2009. These new sustainable yields projects, like the Murray-Darling Basin Sustainable Yields project, have multiple disciplined-based teams located around the country. These projects are also producing multiple region and technical reports subjected to the same multi-staged review process as the Murray-Darling Basin Sustainable Yields project. This, like in the Murray-Darling Basin Sustainable Yields project, is resulting in the generation of large volumes of reporting data.

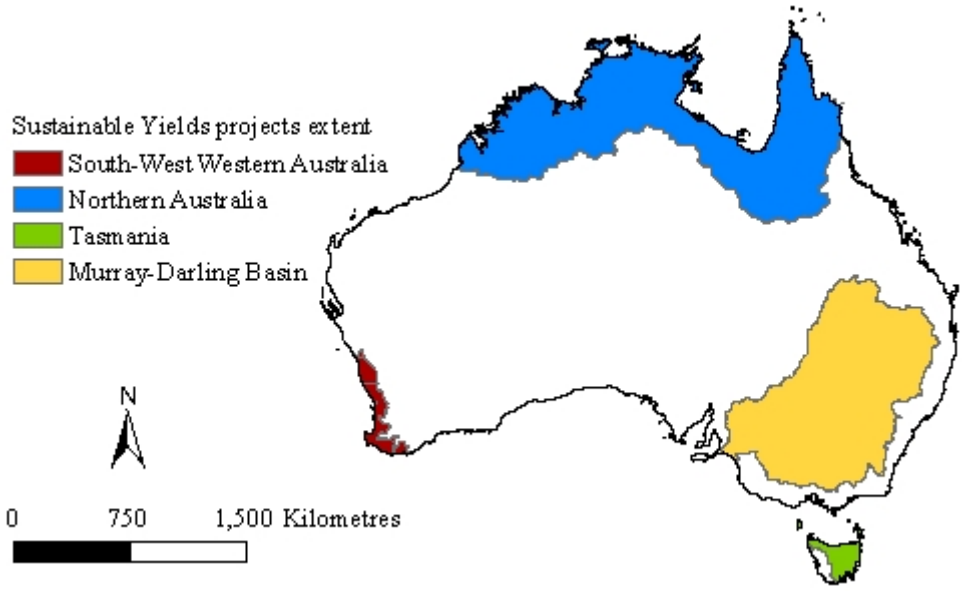

Figure 1. A map showing the location and the extent of the four sustainable yields projects.

Managing the report production was the responsibility of the reporting team. This team, consisting of a data coordinator, production support assistants and editors, was responsible for collating elements (plot, tables, maps and diagrams) and text from each of the modelling teams into the multiple reports before and during the review process. Throughout this process, the ability to trace these elements in the final reports back to the raw data and modelling runs was critical.

The tracking of data from its original source and processing history is called data lineage (also referred to as data provenance and data pedigree) (Bose, 2002; Simmhan et al., 2005a; Simmhan et al., 2005b). This information is recorded as metadata (information describing data) and supports a number of applications including:

- data quality – data lineage can be used to determine the reliability and usefulness of data
- audit trail – allows data and the analytical processing to be audited
- repetition – recording the processing steps allows for repetition to verify results, as well as modification of the process, or updating the results with the latest data
- attribution – used to determine ownership and copyright of the data
- informational – used to locate datasets of interest and to assist with the interpretation of data in the intended context (Simmhan et al., 2005a; Simmhan et al., 2005b).

One way of maintaining the data lineage of a dataset is through a workflow. Such a workflow was developed by the reporting team to track text and elements from their source through the review process to the final products. The workflow was required to be easy to implement, needing no specific software, as the number of contributors across multiple organisations, and the tight project deadlines, greatly hindered the possibility of training. This reporting workflow, originally designed for the Murray-Darling Basin Sustainable Yields project, is outlined below, along with its adaption to the current sustainable yields projects.

2. THE REPORTING WORKFLOW AND ITS IMPLEMENTATION

The reporting workflow describes the stages of development from the individual elements (plots, tables, maps and diagrams) and text, though to the final product (Figure 2).

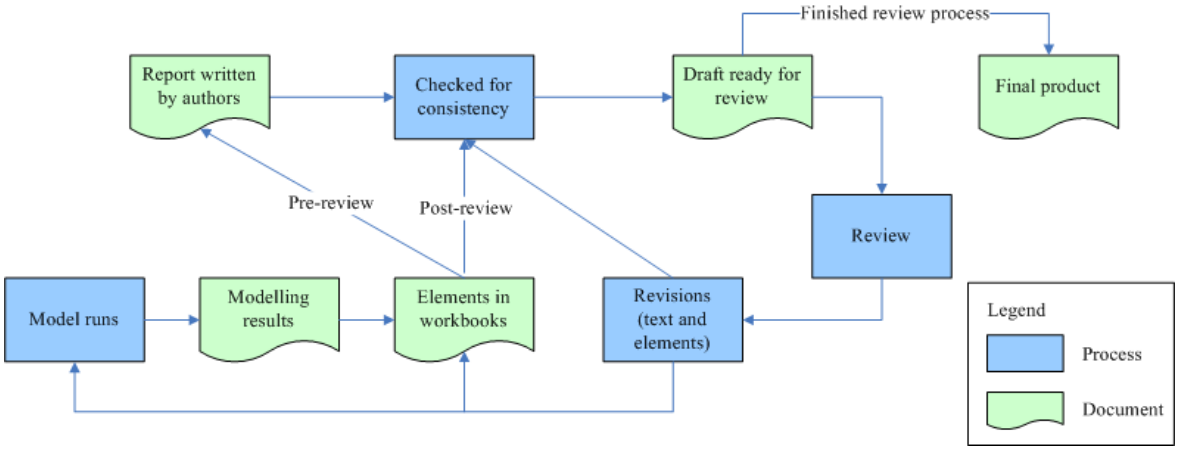

Figure 2. The reporting workflow showing the development of text and elements, including their revision through to the final product.

The steps in the workflow are:

- Report written by authors - the reporting workflow begins once the model runs have been completed and the authors have completed the first draft of their report section. Each report section is then handed over to the reporting team and is said to be under product control (under revision/reporting control).
- Checked for consistency - once in the reporting domain, a production support assistant checks and updates the elements from the modelling team workbooks as required, and updates any numbers referenced in the text. The report section is then handed over to the content editor to ensure readability, followed by the copy editor to check and correct consistency of language and numbers between report sections, and between reports.
- Draft ready for review - the report sections are then collated and are ready for review.
- Review - the report goes out to review.
- Revisions (text and elements) - once back from review, the elements and text are updated in line with reviewer comments. For the elements, this may involve re-running models (model runs → modelling results), or simply making adjustments to the post-processing (elements in workbooks). The text is returned to authors for revision. Once the edits are incorporated in to the report section, it is returned to the production support assistant to update any revised elements and numbers referenced in the text (consistency checking) ready for the next stage of review.
- Final product - once the review process is completed, the final product is produced.

Elements (plots, tables, diagrams and maps) and text are handled separately in the workflow. Elements are stored together in a Microsoft Excel workbook with each modelling team having their own workbook. Each element is contained within its own worksheet, with the worksheets containing the plots and tables also

containing the underlying data used to generate them. This allows the reporting team full control over the appearance of these elements. The worksheets for the diagrams and maps also contain a link to the image file, and in the case of the maps, a link to the mapping project file.

Each element is given a unique identifier (element number) so that it can be tracked though the workflow. This number contains two letters, indicating the team the element belongs to, followed by two numbers to individually identify each element (e.g. SW14, the 14th element from the surface water team). Element numbers are used to label the Excel worksheets, image files and map project files for easy identification.

The Excel workbooks contain a version log as the first worksheet. This log contains a history of the development of each of the elements. In particular, it contains information regarding which model the input data to the element came from, and the subsequent processing history. It also records the date and the person responsible for the update. The log is updated for each new version, and is also used by the reporting team to determine which elements have been updated in the Excel workbooks and therefore need updating in the reports (Figure 3).

(a)

	A	B	C	D	E
1	**Pre-Version 10**	**Version 10**	**Version 11**	**Version 11b**	**Version 12**
2	18/12/2007	18/12/2007	20/12/2007	7/01/2008	9/01/2007
3	From Nicola and team	From Nicola	Modified by Geoff	David	Nicola
4	Loading data from model runs	Workbook handed to Geoff for checking	Updated Scenario Cdry data for SW10	SW16 Made y-axis consistent	Corrected the information in SW08
5				SW25 Modified y-axis	
6				SW34 Modified y-axis	

(b)

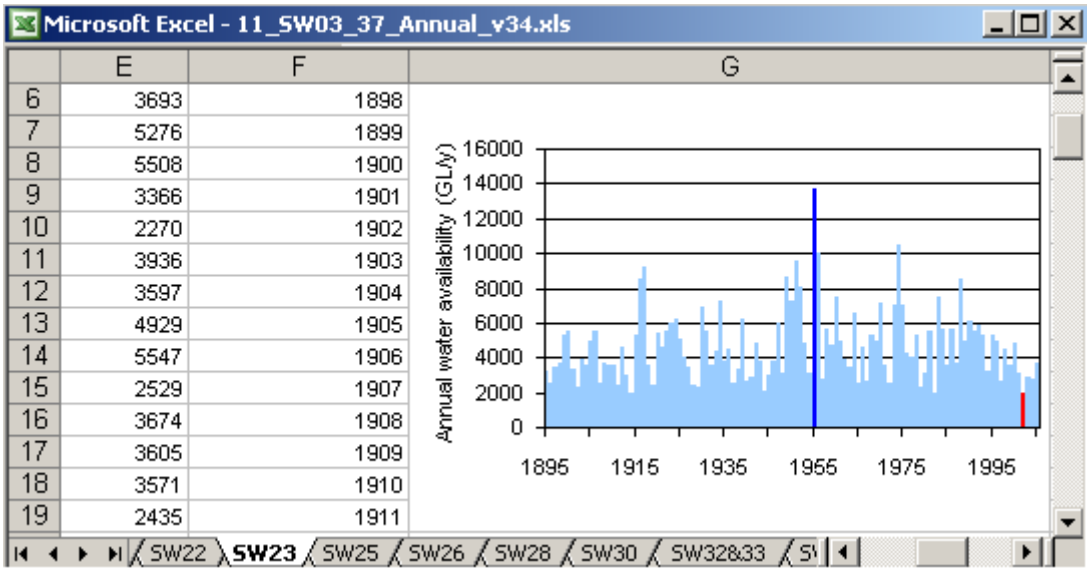

Figure 3. An Excel workbook illustrating (a) the version log sheet showing the edits and therefore the processing history of each of the elements, and (b) an element with the data that generates it (after CSIRO, 2008). Note also that the worksheet tab is labelled with the element number, and the workbook file name is labelled with the elements it contains (in this example, elements 3 to 37).

Unlike the word documents, the Excel workbooks are stored on the network drive. This is due to large file size and the time required to up-load and down-load to the SharePoint Internet site. This ensures that there is separation in the storages used for the Word documents and the workbooks, which allows work to continue if one storage system becomes temporarily unavailable.

Text is written by the authors into a base Microsoft Word document. The base document is a structured Word template containing the headings, subheadings, element lists and associated captions, along with the proscribed formatting and stylistic conventions agreed upon for the reports (Figure 4). The base document also forms an important link in the data lineage for the reports, as it is the only document that links the elements in the final report back to the workbooks. An early decision was made in the Murray-Darling Basin Sustainable Yields project to use element numbers instead of figure and table numbers to track the elements. These element numbers are independent of the position of tables and figures in the reports.

Water availability

Element	Form	Content	Length
SW24	Text	General text	¼ page
SW26	Table	Table 4-7. Annual water availability relative to Scenario A under scenario C and D	½ page
SW23	Plots	Figure 4-3. Annual water availability under Scenario A	1 page
SW25	Plots	Figure 4-4. Change in annual water availability relative to Scenario A under (a) Scenario C and (b) Scenario D	¼ page

Table 4-7. Annual water availability relative to Scenario A under scenario C and D

Figure 4-3. Annual water availability under Scenario A

Figure 4-4. Change in annual water availability relative to Scenario A under (a) Scenario C and (b) Scenario D

Figure 4. A section of a base document used in the Murray-Darling Basin Sustainable Yields project showing the section heading, elements (referred to by element number) and their captions. Element tables are dispersed throughout the document within the relevant section. Section 3 discusses the evolution of the base document in the three current sustainable yields projects.

At the beginning of the Word document there is a version log. The log records the details of the changes made and who authored them. A new version number is created each time the document is updated. For example, if an element in the Word document is updated, the log would identify the element and its source workbook. The log also indicates the document's stage of development through the workflow. All previous versions of a document are archived to record document development, and also as protection against accidental corruption of the latest version.

Word documents are stored online using the Microsoft SharePoint Internet collaboration software. SharePoint permissions are used to control access to the documents until their public release date. The check-in/check-out facility is used to ensure that only one person may work on the document at any one time.

This process, as outlined above, was used in the Murray-Darling Basin Sustainable Yields project to maintain data lineage for both the region reports and the region summary reports. These summary reports were largely formed from text and elements that were copied directly out of the region reports. The elements for each of the summary reports were contained in their own Excel workbook. A base document for each of the team-based sections contained the text from the region reports, with minor re-working. These documents were then sent to an external design company to produce the summary reports.

This workflow was implemented in two main ways. In the initial stages of the project, a series of workshops and face-to-face meetings were held. In addition, a series of information sheets were developed. These information sheets provided instruction in: the construction of an Excel workbook and a base Word document; the use of Word templates within a document; image management; and collating Word documents into a single report using Adobe Professional.

The second key strategy used to implement the workflow was to strengthen communication between each of the modelling teams and the reporting team. To facilitate this, each modelling team had a designated production support assistant. Forming the link between the reporting team and the respective modelling team, this person ensured the latest element versions were contained in the reports before editing and review. The production support assistant also carried out minor formatting corrections to the elements, clarified editor queries with the authors, and ensured that their modelling team followed correct process so that data lineage was maintained.

3. DISCUSSION

The reporting workflow was originally designed for the Murray-Darling Basin Sustainable Yields project and with only minor changes, has been successfully adapted to the Tasmanian Sustainable Yields project, the South-West Western Australia Sustainable Yields project and the Northern Australia Sustainable Yields project. The changes included the treatment and storage of maps and diagrams by each of the modelling teams, the structure of the base document, and the handling of edits between region reports and region summary reports.

In the Murray-Darling Basin Sustainable Yields project many modelling teams treated the maps and diagrams differently to the plots and tables. While some teams stored the maps and diagrams in the Excel workbooks as outlined in the process above, other teams did not. Where maps and diagrams were not stored in workbooks, no information regarding their evolution was recorded. Although versioning was used, it was more difficult to know whether a new version of the document had been created, as the files were not stored in an organised fashion. The method developed for the three new sustainable yields projects allows map changes to be recorded. It also means that the production support assistant need only check one document per modelling team to locate and insert updated elements into the report, and that new versions are less likely to be missed.

In the Murray-Darling Basin Sustainable Yields project maps and data layers were also stored in a different directory structure to that of the workbooks. This made it difficult to locate some of the maps and make minor changes to the documents if some of the data layers were stored offline. An established directory structure should be used when archiving the documents for long-term storage, as the software package used to generate the maps (ArcMap) saves the pathway to the data layers, rather than embedding the data in the project file. Therefore if any mapping files are moved, the data links are broken.

The structure of the base Word document has been changed for the benefit of the new sustainable yields projects. The base document used previously in the Murray-Darling Basin Sustainable Yields project had tables of elements listed throughout (Figure 4). As the document began to be populated these were deleted, resulting in information regarding the element numbers to be lost. This maintained the need for a separate, unpopulated base Word document to link the final elements in the reports back to the elements in the Excel workbooks. In the current sustainable yields projects the list of elements are at the beginning of the document (Figure 5), and can remain throughout the document's development. It can also be customised to the individual document, eliminating the need for a separate document to maintain data linage.

Element	Form	Content	Section
RR01	Map	Figure 1. Map showing reporting regions, catchments, and calibration catchments	3.1
RR02	Plot	Figure 2. Mean annual rainfall, areal potential evapotranspiration and modelled runoff	3.2
RR03	Plot	Figure 3. Mean summer (DJF) rainfall, areal potential evapotranspiration and modelled runoff	3.2
RR04	Plot	Figure 4. Mean winter (JJA) rainfall, areal potential evapotranspiration and modelled runoff	3.2

Figure 5. The base document as it is used in the three current sustainable yields projects. This base lists all the elements at the beginning of the document, allowing the table to be kept for the duration of the document. Captions and section headings still remain in the body of the document.

In the Murray-Darling Basin Sustainable Yields project each of the regions had a region report and an associated summary report. This summary report was composed mostly of elements and text that were copied directly out of the region report. Keeping both documents synchronised proved challenging, as the summary reports were produced by an external design company. During the design process, the design company was notified of minor edits either by phone, or by marked-up copies sent by email. These changes did not always flow back into the original electronic version of the document. When major changes were made, a new electronic version was sent, resulting in some of the earlier minor edits needing to be re-done. While these errors were identified during quality control checking, a more disciplined approach to editing these documents will ensure that this will not happen in the three new sustainable yields projects and that data lineage is maintained.

As with the development of any framework, whether it be modelling or reporting, it is important that there is sufficient flexibility built into the framework so that it can accommodate improvements without

compromising the longer-term benefits that come from consistency and conformance to standards. The changes implemented for the three new sustainable yield projects significantly improve the framework, with no negative effects.

4. CONCLUSION

The decision to create a separate team to manage the report production and delivery was an innovative one, given that most project team members were researchers more used to writing and producing their own reports. However, this decision was taken due to the enormity of the underlying modelling exercise, the highly political and high profile nature of the project, the short-term and multiple delivery deadlines, and the style of reporting required. It also allowed for additional resources and capacity in the reporting area. The development of, and rigorous adherence to, the reporting workflow that has been described in this paper allowed for the delivery of high quality products on time and to specification.

The separation of the reporting team from the modelling teams in these large modelling projects has proved to have multiple positive outcomes. Not only did it result in efficient and effective report delivery, it also exposed researchers to the benefits of using templates and processes, maintaining data lineage, and the experience of working with content and copy editors, and dedicated production assistants. As the Murray-Darling Basin Sustainable Yields project progressed, the simple modelling team tasks were transferred to the team's production support assistant as they became so familiar with the reports' contents that they could identify and fix errors and omissions, under minimal guidance from the modelling teams, resulting in greater efficiency.

In conclusion, the reporting workflow processes developed for the Murray-Darling Basin Sustainable Yields project have proved to be effective, adaptable and scalable as illustrated by their adoption in the three new sustainable yields projects. Using common office tools and requiring little training, this workflow can be implemented quickly across multiple organisations.

ACKNOWLEDGEMENTS

The Murray-Darling Basin Sustainable Yields project was commissioned by the National Water Commission on behalf of the Australian Government under the Raising National Water Standards Program in collaboration with the Australian Department of the Environment, Water, Heritage and the Arts (DEWHA). The North Australian Sustainable Yields project forms part of the Northern Australia Water Futures Assessment and is commissioned by the National Water Commission on behalf of the COAG and in consultation with the DEWHA. The Tasmanian Sustainable Yields project and the South-West Western Australia Sustainable Yields project are commissioned by DEWHA on behalf of COAG.

The authors would also like to thank the Murray-Darling Basin Sustainable Yields project reporting team and the Water for a Healthy Country Flagship, the umbrella organisation within CSIRO for the four sustainable yields projects.

REFERENCES

Bose, R. (2002), A conceptual framework for composing end managing scientific data lineage. Proceedings of the 14th International Conference on Scientific and Statistical Database Management.

CSIRO (2008), Water availability in the Murrumbidgee. A report to the Australian Government from the CSIRO Murray-Darling Basin Sustainable Yields Project. CSIRO, Australia. 155pp.

Simmhan, Y. L., B. Plale, and D. Gannon (2005a), A Survey of Data Provenance in e-Science. SIGMOD Record, 34(3) p31-36.

Simmhan, Y. L., B. Plale, and D. Gannon (2005b), A Survey of Data Provenance Techniques, in Technical report TR-618 Computer Science Department, Indiana University.