Data Management Plans – Are They Working for the Australian Antarctic Program?

Presenter - Dave Connell[1]

Author - Dave Connell[1], Ben Raymond[2]

[1]Australian Antarctic Division, Kingston, Australia, dave.connell@aad.gov.au
[2]Australian Antarctic Division, Kingston, Australia, ben.raymond@aad.gov.au

The Australian Antarctic Data Centre (AADC) implemented a Data Management Planning system for the Australian Antarctic science program in 2012. Since then 112 data management plans, covering two distinct rounds of Antarctic project submissions have been lodged into the AADC MyScience data management application. Data recorded in these data management plans are now being archived at the AADC, and the utility of the process can begin to be assessed.

DATA MANGEMENT WITHIN THE AUSTRALIAN ANTARCTIC PROGRAM

In order to facilitate effective data management within the Australian Antarctic program (AAp), the AADC was established in 1995. Since then the AADC has created infrastructure to support the archival of data and the creation of metadata records, as well as value-adding to Australian Antarctic data through the use of GIS and mapping tools, the creation of targeted databases, and data-analysis activities. However, until 2012 the AADC lacked the capability of directly managing the data archival needs of each AAp science project. In 2012, the AADC launched the MyScience project management application (Finney 2014) [1], which included the capability to create data management plans.

Each AAp science project is required to complete a data management plan (DMP) within its first six months. So before the scientific work is already carried out, AAp scientists are required to think ahead and inform the AADC of their likely data management needs. DMPs are reviewed by AADC staff before approval to ensure they adequately capture the data-related activities and outputs described in the project application.

The existence of these DMPs allows AADC staff to efficiently keep track of AAp science projects and ensures that all expected datasets are accounted for, and that AAp scientists are not unduly pestered for data before the expected due dates.

TEETHING PROBLEMS

The first version of the DMP tool in the MyScience application was built around vocabularies. Extensive vocabularies were developed for the tool based on the Global Change Master Directory science keywords [2] and the British Oceanographic Data Centre's parameter dictionary, but the tool was also designed to allow users to contribute their own keywords. Vocabularies based on scientific instruments and platforms were also included.

The intention was to eventually incorporate the parameter, instrument and platform keywords into metadata records describing the datasets, which would allow very powerful search tools to be built around them – both for human and machine users.

However, in many cases the vocabulary-oriented process did not sit comfortably with scientists. Firstly, many scientists do not tend to think of their research activities in terms of parameters, instruments, and platforms, but rather in terms of logical data groupings (e.g. they might consider a "data set" to be a set of related variables, or a single season's data collection). Further, many DMPs became overly complicated - instead of making a single entry of "meteorological data from Casey Station", a scientist might be required to make multiple entries of "mean temperature data from Casey Station", "mean air pressure data from Casey Station", "wind direction data from Casey Station", etc. In some cases this meant that a DMP had numerous entries that were largely duplicates of each other (but varying in parameter or location details, for example).

This also placed a large burden on AADC staff during the DMP approval process, as well as during the management of data delivered by project personnel against those approved DMPs. A given data set (either described in a project proposal, or submitted by a project team) effectively mapped to multiple entries in a DMP, and working out which ones were relevant was not always straightforward. Partly for these reasons, the process of creating a DMP was often a long-drawn out affair, requiring extensive back-and-forth dialogue between AADC staff and AAp scientists.

WHAT'S CHANGED

For the second round of AAp science projects that required DMPs, the form was changed so that data would be managed at the dataset granularity, rather than the parameter granularity. Scientists were no longer required to specify individual dataset parameters that they planned to archive, but instead could identify a whole dataset. Hence, using the example

above, scientists could now indicate that they would archive a dataset of "meteorological data from Casey Station" (while still providing details of what that dataset would comprise).
This has resulted in a significantly easier and faster process. Anecdotal evidence suggests that AAp scientists found the form much easier to complete, and much more reflective of the work they were actually doing. It has also helped speed up the process due to the fact that the review component by AADC staff is greatly simplified by the "whole of dataset" approach, and less back-and-forth dialogue is required.

BENEFITS AND WHAT'S NEXT

The benefits of the DMP system, even in its early form, were quickly apparent when it came time to archive data resulting from AAp projects. The DMP provided AADC staff with not only a checklist of what data to expect, but when to expect it and who to expect it from. This has greatly improved communication with AAp scientists surrounding these matters, and also allowed AADC staff to more accurately report data management outcomes relating to AAp projects. The MyScience data management tool and the associated DMPs are fully accessible by the responsible scientists. This provides AAp scientists with a means of easily assessing their projects in relation to data management, and allows them to quickly access their data, their data citations, and their dataset DOIs.
Further improvements to the system are already planned. Improved reporting tools need to be implemented - AAp scientists are obliged to appropriately manage their data (according to the AAp data policy), so a data management tool that allows AADC staff to easily produce a report detailing the data management status of each project will allow the AADC to provide more accurate and detailed advice to the AAp review panel. Furthermore, greater integration with data submission tools is required – ideally when scientists submit data to the AADC, the submission tools will link to the DMP and allow users to specify which item from their DMP they are archiving.
We will also consider the feasibility of re-integrating data vocabularies at the data submission stage, so that their advantages are retained while still maintaining the simplicity of the current DMP process.

REFERENCES

1. Finney K (2014) Managing Antarctic Data – A Practical Use Case, Data Science Journal, 13 PDA8-PDA14, doi.org/10.2481/dsj.IFPDA-02
2. Global Change Master Directory (GCMD). 2015. GCMD Keywords, Version 8.1. Greenbelt, MD: Global Change Data Center, Science and Exploration Directorate, Goddard Space Flight Center (GSFC) National Aeronautics and Space Administration (NASA). URL:http://gcmd.nasa.gov/learn/keywords.html

ABOUT THE AUTHOR(S)

DAVE CONNELL

Dave Connell completed a Bachelor of Science (honours) degree at the University of Tasmania, and has been working at the Australian Antarctic Division since 1998 and as the metadata officer since 1999. His role is to catalogue and archive all scientific data collected by the Australian Antarctic program – specifically to ensure that scientists write high quality metadata records and archive their data in a timely manner. During his time at the AAD, he has overseen the transition from ANZLIC metadata to DIF metadata, and also developed tools for converting DIF metadata into various profiles of the ISO 19115 metadata standard. Dave is also very active in the Australian Government metadata space – reviewing and adapting ISO 19115 metadata standards for use in Australian scientific organisations. Recently he has begun working with the Ocean Acidification – International Coordination Centre to develop an ocean acidification metadata profile.