MANAGING ANTARCTIC DATA—A PRACTICAL USE CASE

K Finney1*

1Australian Antarctic Division, Australian Antarctic Data Centre, Kingston, 7050 Tasmania, Australia
Email: kimfinney@gmail.com

ABSTRACT

Scientific data management is performed to ensure that data are curated in a manner that supports their qualified reuse. Curation usually involves actions that must be performed by those who capture or generate data and by a facility with the capability to sustainably archive and publish data beyond an individual project’s lifecycle. The Australian Antarctic Data Centre is such a facility. How this centre is approaching the administration of Antarctic science data is described in the following paper and serves to demonstrate key facets necessary for undertaking polar data management in an increasingly connected global data environment.

Keywords: International polar data management, Scientific data curation, Data administration, Preservation, Data reuse

1 INTRODUCTION

The Australian Antarctic Data Centre (AADC), which has been operating for 16 years as the primary data repository for the Australian Antarctic Science program (AAP), has been gradually refining its policy base, working to integrate data services into the science program workflow, and continuously developing underpinning data infrastructure. Each of these activities is designed to improve data management services available to Antarctic researchers and to lift the volume and types of science data that are publicly accessible for reuse.

The AAP is a competitive research program involving scientists from the Australian Antarctic Division (AAD), the Commonwealth Scientific and Industrial Research Organisation, Australian state/federal government agencies, the university sector, and international institutions. The AADC coordinates the archiving and publication of data derived from AAP Antarctic and Southern Ocean-based research according to the open data principles of the Antarctic Treaty System (Antarctic Treaty Secretariat, 1959). In performing its functions, the centre works as part of the international network of Antarctic Data Centres, co-ordinated under the auspices of SCAR (the Scientific Committee on Antarctic Research), and was admitted to the International Council for Science – World Data System (ICSU-WDS) in 2011. ICSU-WDS is an international federation of global data centres and data service providers. Australia’s ability to contribute to such global systems and to reuse data within the AAP and beyond is dependent upon scientists paying adequate attention to data management tasks that need to be performed within individual science projects and upon easy researcher access to core data management infrastructure. This paper describes how the AADC has been approaching polar data administration and how it is developing infrastructure to support AAP science. Whilst there is still much room for improvement, the combination of activities, practices, and policy described here present a useful example of how polar data management can be coordinated to scientific and national advantage.

2 SCIENCE APPLICATION PROCESS AND AAP DATA POLICY

In 2010, the AADC conducted an audit of the data it had received from past science projects implemented under the umbrella of the Australian science program in all of its previous guises, since the establishment of the AAD in 1980. In this audit there was a specific focus on those projects that commenced after the creation of the AADC (in 1996). Not surprisingly, it was found that a large number of projects had not submitted any data for archiving, despite a long-standing policy (first formalised in writing in 2004) that ‘all data should be deposited with the AADC’. Three critical issues were identified as contributing to this poor level of compliance:

1. A lack of implemented penalties for non-compliance (even though sanctions, such as the right of the chief scientist to deny a chief investigator access to AAD logistical support, were informally touted within the program).
2. No prior understanding by the AADC of specifically what datasets should be delivered from approved AAP projects and hence a limited ability to chase outstanding data submissions.
3. An inadequate set of utilities available for the AADC to administer policy compliance and too few tools and assistance for scientists to comply with many of the (post 2006) data policy obligations.

PDA8
Recognising that reforms were necessary, development of the new 2011–2021 Antarctic Science Strategic Plan (Australian Antarctic Division, 2011) offered an opportunity to revise and strengthen the current AAp data policy (AADC, 2013) to more closely align it with the science project assessment process and to begin targeted upgrading of the AADC toolset. These policy changes and science project assessment alignments are described in the next few sections and characterise the AADC’s approach to scientific data administration.

2.1 Data submission history assessment criterion

Since the introduction of the 2011 Science Strategic Plan and the drafting of the new data policy, a public call is made every two years for science proposals. Submitted proposals are subject to peer review using a new ministerial-approved assessment process that now includes specific reference to the AAp data policy. Within this process, project proposals are rated based on a range of criteria associated with the quality and relevance of the proposed science and the competence of the listed research team. An important change in the new assessment criteria is that a chief investigator’s previous history of data submission is now taken into account in the scoring. Although only three points (out of one hundred) are allocated to data submission history, because the program is highly competitive, these relatively few points have the capacity to influence the assessment outcome. Research scientists with no previous history of participation in the AAp as a chief investigator and those with an excellent data submission history get allocated the full three points. Those with a particularly poor track record of data and metadata submission are allocated zero points. Performance variations in between are assigned either one or two points.

It is already evident from the number of people who have contacted the AADC to submit old datasets since the policy was marketed that this approach provides a good incentive for scientists to make sure that they have sustainably archived their data. It is however readily acknowledged that by applying penalties anchored to the proposal assessment process, we are really mainly affecting those researchers who have a repeat history of working in Antarctica (or within the AAp grant scheme). Because the majority of chief investigators in the Australian program do have a long and active connection to the AAp, most will have a vested interest in maintaining a good data management record.

By including data submission history as part of the assessment criterion used to judge the competence of the chief investigator and his/her team to conduct the science proposed, we are reinforcing the expectation that science professionalism involves maintaining good data management practice.

2.2 Data management planning

The newly strengthened data policy also includes a provision that successful AAp projects must now submit a data management plan, to be delivered to the AADC by a chief investigator within the first six months of receiving project approval. Assistance with producing these plans is provided by AADC staff (in their roles as Science Project Liaison Officers: SLOs), and plan creation is standardised and made easy by using an online tool. Plans, once submitted, are versioned and reviewed to ensure they meet guidelines and then remain active for the duration of the project. Development of these plans is considered to be the first milestone in all approved AAp projects, and implementation progress is tracked through a formal project monitoring and review process conducted annually by a science review committee (the Antarctic Research Assessment Committee; ARAC), which has an independent chair, external to the AAD (Australian Antarctic Division, 2012).

Within the plan, project team members must identify what datasets will be collected, when these data will be ready for submission to the AADC, who in the team will be responsible for their submission, and the likely volume of data that will be deposited. Under normal circumstances investigators must submit all project data to the AADC (or an alternate sustainable repository) by a project’s end date. For the first time since the centre’s inception in 1996, it is now possible to forecast the type and approximate quantity of data that will be generated annually from Australian Antarctic research. This information enables the AADC and its parent institution, the AAD, to improve the management and growth of expensive information technology infrastructure (e.g., digital storage area networks) and science facilities (e.g., on and offsite storage for biotic and geologic specimens/samples and ice cores). Better facilities planning should lead to enhanced services for research projects.

2.3 Data citation

Whilst it is not yet mandatory in the AAp Data Policy for AAp scientists to formally cite data in authored research publications, it is now strongly encouraged. If scientists cite their own data it becomes more visible and more widely accessible, and options for using both datasets and paper publications as measures of professional achievement become possible. For many scientists, particularly those engaged in observational and monitoring
science, a significant proportion of their life’s work is invested in capturing and collating datasets whose value becomes more apparent through time. The number of publications possible from such data may be limited in the early phases of their research due to the need to establish temporal trends, variability, and baselines before publishing. Being able to demonstrate the various uses of their data (through reviewing citations) should be an important factor in determining the impact of researchers’ scientific activity in conjunction with their publication history. But most fundamentally, citation involving online, accessible data provides an open mechanism for scientific verification and validation (The Economist, 2013).

Compliance with this relatively new citation policy element is being monitored by ARAC, with input from the AADC. The AADC is able to supply persistent addressing for formal dataset citations, namely, digital object identifiers (DataCite, 2013) minted by the Australian National Data Service (ANDS, 2012), and provides guidance for scientists on emerging citation standards (Kotarski, Reilly, Schrimpf, Smit, & Walshe, 2012) by automatically marking-up deposited data for online publication using these standards. Recognising that there is a strong cultural element to this policy principle, and because global ‘systems’ are not yet in place either within many existing repositories or within the publishing sector, a ‘soft’ approach is being taken to shepherd AAp scientists into citation as a practice.

3 MYSCIENCE

Figure 1. Screen snapshot showing a portion of a MyScience project record

To successfully implement the new data policy, the AADC rearchitected some of its infrastructure so that: (a) the AADC could monitor policy compliance and feed this information into the governance framework established for monitoring AAp projects and (b) AAp research scientists had utilities that enabled them to readily comply with policy directives. With a keen desire to minimise application maintenance overhead, it was decided that the primary tool used by the AADC to administer policy compliance would also be a utility that
could be used by AAp scientists to manage their individual project-based resources (i.e., metadata records, datasets, associated documentation, publications, and Data Management Plans). The Web-based application developed to fulfil this function is called MyScience (see Figure 1).

3.1 Resource administration through MyScience

MyScience is accessible via secure login and is available to any scientist with an internet connection, a browser, and one or more registered AAp projects (past or current). It provides a single interface for scientists to access functionality and content from separately developed, mainly pre-existing AADC systems and data stores. The system is project-centric in that each MyScience record relates to a single AAp project that is usually associated with multiple scientists and support staff. MyScience accesses information from corporate databases of registered AAp scientists, project proposals and progress reports, publications, metadata, and data. From this interface an investigator can:

1. Create metadata records
2. Deposit datasets and associated resources and link these to existing metadata records
3. Register publications
4. View summaries of project activity timelines, team composition, and project resources that have been registered with the AADC
5. Access metadata and data associated with the project

AADC staff, in their roles as SLOs, can also insert annotations anchored to various elements of the MyScience record (i.e., ‘to-do’ messages, see Figure 1) as data administration reminders for project team members. This messaging facility is only activated when logged into MyScience using an SLO role. Since most AAp research teams are from institutions outside of the AAD and are distributed across Australia, this communication channel, centred on a compendium of a project’s resources, has proven an effective way to reach project members regarding data administration issues.

Apart from functioning as a portal for project resource management, MyScience can produce reports for the AADC that are used in governing the AAp (e.g., information for the AAp science project review and assessment processes and program performance monitoring activities). The remaining functionality inherent in MyScience pertains to the creation and management of data management plans.

3.2 MyScience and data management planning

AAp data management plans are designed to assist project teams in thinking about likely data flows and any associated ‘within-project’ data management early on in their project’s life-cycle. The plan’s function is to educate project teams about available services, facilities, and obligations under the AAp data policy. It is also a vehicle for encouraging teams to identify, before field work commences, what data ‘agreements’ might need to be put in place with collaborators who are external to the AAp. Explicitly performing this particular task can prevent the conflict over data access and publication that often arises in science programs due to misunderstandings over implicit agreements about data application and ownership.

The online data management planning utility, accessible from within the MyScience application, is essentially a planning template. It contains three different types of information:

1. Project-based information already registered in other AAD systems
2. Preformulated text that the AADC automatically inserts into the plan (usually basic guidance on data management issues)
3. Information provided by project team members in response to data management questions

Questions in the planning template contain pick-lists and checkboxes where possible, and information being sought through the plan has been winnowed down to only those things that the AADC considers essential in order to reduce the administrative burden on those developing plans.

The template uses a range of controlled vocabularies for inserting content in specific sections (see Figures 2 and 3). Investigators are encouraged to supply new vocabulary terms when there are deficiencies in seed lists, and AADC staff then receive automatic notifications to moderate terms. Unmoderated terms can still be used to populate the plan template in real time, but if a term is later changed after moderation, the term is updated in the plan (and the plan creator is notified if he/she has not already been contacted during the term moderation process). The plan’s vocabulary seed lists are pre-populated, where possible, with terms reused from existing
domain vocabularies. The data management planning process is, however, being used by the AADC to build comprehensive and relevant vocabularies for AAp science because there are currently no vocabularies available that fulfil all of the program’s requirements.

The vocabulary terms captured are being reused in other AADC core infrastructure to mark-up AAp metadata and data that are exchanged within global data networks. The ultimate benefit of these activities for scientists is that datasets described using rich, standardised, and mapped vocabularies can be discovered and accessed with much higher precision and recall than poorly and inconsistently described data. Domain vocabulary development and harmonisation is a relatively new activity within scientific data administration and is currently being pursued across many scientific disciplines globally. This is because formalising the description and definition of scientific concepts facilitates other desirable activities such as automated data extraction, integration, and manipulation. The AAp data management planning process provides a very structured way for the AAp to fill gaps in existing polar science vocabularies.

![Figure 2. Screen snapshot showing a portion of the data management planning tool](image)

Finally, all submitted plans are considered fluid, in that they can be added to or changed over time. This fits with the dynamic nature of scientific research and the ever-changing logistics of operating in a harsh Antarctic environment. Plans are versioned and logged, and old and new versions are permanently accessible online.
CONCLUSION

The data administration changes, facilities, and activities outlined in this paper have already resulted in long-outstanding datasets being deposited within the AADC. These previously hidden data are now available for reuse. The centre is currently a far better placed to administer the AAP policy, and scientists are being supported to comply with policy obligations. Investigator co-operation is helping to build better infrastructure, which is more closely meeting scientific data publication, discovery, and access requirements. Our experience has shown that data policy, promulgated through a resourced governance framework that is tied into science program and project administration, can lead to better data management outcomes. In the long term this can only be beneficial for scientists and national science endeavours, particularly in disciplines such as Polar Science, where data capture is such an expensive activity.
5 ACKNOWLEDGEMENTS

AADC staff (i.e., Miles Jordan, Mat Oakes, Matt McKellar-Spence, Ursula Harris, Dave Connell, and David Smith) have contributed in various ways to the design and implementation of the systems and procedures described in this paper.

6 REFERENCES


(Article history: Available online 23 September 2014)