ANSTO, AUSTRALIAN SYNCHROTRON, METADATA CATALOGUES AND THE AUSTRALIAN NATIONAL DATA SERVICE*

N. Hauser, S. Wimalaratne, ANSTO, Lucas Heights, NSW 2234. Australia
U. Felzmann, Australian Synchrotron, Melbourne, VIC 3168 Australia

Abstract

ANSTO and the Australian Synchrotron host experiments for domestic and international science community over a wide range of scientific disciplines. Data citation, management and discovery are important to the facilities and the scientists that use them. Gone are the days when raw data is written to a removable media and subsequently lost or locked away in a desk to collect dust.

The metadata catalogue Tardis is being used by both ANSTO and the Australian Synchrotron. Metadata is harvested from the neutron beam and X-ray instruments raw experimental files and catalogued in databases that are local to the facilities. The data is accessible via a web portal. Data policies are applied to embargo data prior to placing data in the public domain. Public domain data is published to the Australian Research Data Commons using the OAI-PMH standard. The Commons is run by the Australian National Data Service (ANDS), who was the project sponsor. The Commons is a web robot friendly site.

ANDS also sponsors digital object identifiers (DOI) for deposited datasets, which allows raw data to now be a first class research output, allowing scientists that collect data to gain recognition in the same way as those who publish journal articles. Raw data is increasingly required by journals to allow for more rigorous referring of publications. Data is being discovered, cited, reused and collaborations initiated through the Commons.

INTRODUCTION

ANSTO and the Australian Synchrotron collect data from neutron and X-ray scattering instruments. The data produced is in several files formats, some of which are not friendly to web robots. Catalogues of these files were made, accessible over http, that can be searched and retrieved either using a web interface, or by a search engine web robot. The high level use case of the metadata catalogues is to provide a researcher with a structured search through scientific datasets.

Placing indexed datasets on the web with a digital object identifier (DOI) can be thought of as the publishing of raw data, which elevates that data to a "first class research output" [1]. Whilst the data itself is not peer reviewed, it is data used to create a peer reviewed publication. Making raw data available allows the analysis of data to be verified by a third party, assuming a journal editor or another researcher has need to do so.

Citing Raw Data

Publicly accessible and catalogued research datasets, combined with a digital object identifier (DOI), allow researchers who publish work that uses these datasets to cite those who created the dataset. This opens up new opportunities in these sample based sciences for experimental scientists to increase their citation index. Rather than citing journal publications only, the raw data used to create publication can be cited.

3 use cases around raw data:

- As a scientist that generates the data, I want to search and manage my data
- As a scientist, I want to analyse data that has been collected by someone else
- As a scientists, I want to collaborate with the scientist that collected the data

Research Funding

Two major funding bodies in Australia are the National Health and Medical Research Council (NH&MRC), and the Australian Research Council (ARC). The *Australian Code for the* Responsible Conduct of Research [2] was developed by the NH&MRC, ARC and Universities Australia. The Code has many facets, but here we will pay attention to the management of research data.

The National Laboratories are now providing more data management services. The benefit to the facility for providing these services is that the body that funds the facility can 'see' the data being produced by the facility. Providing better data management aids researchers, increasing the probability that the researcher will publish, and win for both the facility and the researcher.

THE CODE

Management of Research Data

Under the responsibilities of institutions, the Code requires;

"Provide secure research data storage…"
"Identify ownership of research data…"
"Ensure security and confidentiality of research data…"

The Code requires that data must be in a;
" durable, indexed and retrievable form"
" catalogue … in an accessible form"

"A policy is required that cover the secure and safe disposal…"

Under the Code, National Laboratories only have responsibility for researchers they employ. The Laboratories may decide to provide services to scientists that use the facility. In effect, the National Laboratory does not own the data collected by a user of the facility

ISBN 978-3-95450-139-7

(e.g. a graduate student from University of XYZ), but acts as a custodian of the data, provisioning the requirements of the Code. This type of custodianship of data has been the norm for National Laboratories; however, the Code defines data management guidelines to be met by the custodian. These guidelines are not standards, specifications or tools, so there is a great deal of latitude in the interpretation of these guidelines and in the policy and software tools used to implement them. The authors argue that in this first phase of data management the guidelines provide a necessary step forward. As the implementations mature, policy and tools should emerge as 'best practice'.

Publication and Dissemination of Research Findings

The Code requires a researcher to disclose "all" research findings accurately, whilst protecting confidentiality and managing intellectual property rights of the institution, sponsors and researchers. In the majority of cases, intellectual property rights are secured when a peer reviewed paper is published. At the time of publication, the datasets used for the publication can be made available.

Based on the principles outlined in the Code, a policy is created by each organisation for handling data. The Bragg Institute provides its policy on their website [3]

FORMING COLLABORATIONS

Most collaboration is started when a scientist does a literature search or they meet at a conference. Scientists that make their data accessible via a search engines have another avenue to form collaborations.

Take the case where a scientist collects data but doesn't publish. If the raw data is available via a search engine, a collaborator can find this data and subsequently contact the scientist to get more information. Data that would otherwise not make it to publication could be published.

METADATA CATALOGUES

There are several metadata catalogues applications available, both open source and commercial. Tardis [4] was chosen since it had a similar use case in its original purpose as a database of crystallography datasets.

The database is relatively simple, as is the web application. Scientists have exclusive access to their embargoed data. Non-embargoed data is publicly accessible. Data can be downloaded one file at a time, or in a group.

Ingestion

Scientific instruments often use binary formats for data storage due to read-write performance and space efficiency. However, such formats are not friendly to web indexing and searching. It is preferable to have a database that contains only the metadata about the experiment and the author and a pointer to the raw data file in its binary form. This is done in a two-stage process.

Stage 1, the metadata is read from the binary file and written to a database, along with a pointer to the raw data. This process is called ingestion. This metadata is stored in the Tardis database. The raw data stored on a fileserver.

Stage 2, the database has a process to create an ASCII metadata package that is sent to the ANDS data repository over the Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMH [5]. This metadata package is a subset of the stage 1 metadata, and does not contain the instrumental metadata, rather author and instrument metadata and a pointer to the stage 1 metadata. Stage 2 is done at the end of the embargo.

Embargo

In sample based science, the default data access policy for many facilities is to have an embargo period of between 1 and 3 years, where the scientist who synthesised the sample and the scientist who did the measurements have exclusive access to the raw data and metadata. This allows the scientist to do follow-up experiments, or a series of related experiments leading to a publication. This provides protection of intellectual property rights.

When a publication is being refereed, referees can be provided access to the data. Once the publication is published, the raw data and metadata can be made available to allow the broader community to analyse and critique.

Not all facilities provide an embargo. Conversely, not all facilities provide public access to data. Some argue that the data in most raw data files is sufficiently obfuscated that only the scientist doing the measurement can reliably analyse the data. Self-describing data formats, such as Nexus [6] attempt to mitigate this problem. However, it is at the discretion of the experimentalist how much metadata they will provide. The authors are not aware of any facilities doing audits on the quality or quantity of metadata.

Once the embargo period has expired, the metadata and raw data are made available in the public domain

Public Domain / ANDS / OAI-PMH

Australia is privileged to have the Australian National Data Service (ANDS), which is an Australian Government sponsored metadata repository. ANDS sponsors the build of local repositories that feed the central repository. Researchers for across the full spectrum of public sector research can publish the metadata associated with datasets that have value to the nation. Dataset of interest to the ICALEPCS community, in astronomy, high energy physics, X-ray and neutron scattering are available through the Research Data Commons [7].

79 institutions are contributing to the Research Data Commons. Projects such as the Atlas of Living Australia, which is the Australian Node of the Global Biodiversity Information Facility, are sponsored by ANDS, with datasets that are uploaded by professional and citizen

scientists.

Providing an electronics library of Australia's research effort is an enduring legacy.

DOI

ANDS provides a digital object identifier (DOI) minting service, free to the Australian research community. Data objects that are stored at ANDS have the option of having a DOI minted. The DOI provides an enduring link and method for citation of data.

CONCLUSION

Using the metadata catalogue Tardis, ANSTO and the Australian Synchrotron are providing services allowing the researchers that use these facilities to fulfil the Australian Code for the Responsible Conduct of Research in managing research data. Confidentiality and intellectual property rights are protected, whilst providing a mechanism to enhance scientific collaboration and data reuse. Raw data is accessible, searchable and citable.

ACKNOWLEDGEMENT

ANDS is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy Program and the Education Investment Fund (EIF) Super Science Initiative

REFERENCES

[1] Australian National Data Service. http://www.ands.org.au/cite-data/researchers.html

[2] Australian Code for the Responsible Conduct of Research, http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/r39.pdf

[3] Bragg Institute Data Policy, http://www.ansto.gov.au/ResearchHub/Bragg/Users/DataArchiving/index.htm

[4] Federated repositories of X-ray diffraction images. Acta Crystallogr D Biol Crystallogr. 2008 Jul; D64 (Pt 7):810-4. doi: 10.1107/S0907444908015540. Epub 2008 Jun 18.

[5] Open Archive Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/pmh

[6] Nexus data format, http://www.nexusformat.org

[7] Research Data Australia, http://researchdata.ands.org.au