

## ANDS and Data Storage

### Who needs to know this?

---

This is a general introduction that is likely to be of interest to all users of ANDS, including researchers, their support staff, data centre and repository staff and the general public.

### Data Storage and the *Australian Code for the Responsible Conduct of Research*

---

The Australian Code for the Responsible Conduct of Research (see also the separate ANDS Guide *Research data policy and the Australian Code for the Responsible Conduct of Research* at <http://ands.org.au/guides/code-awareness.html>) was developed by the National Health and Medical Research Council (NH&MRC), the Australian Research Council (ARC) and Universities Australia. Published in 2007, it replaced an older version and introduced the importance of data management for the first time.

Section 2, Management of Research Data and Primary Materials, states

“The central aim is that sufficient materials and data are retained to justify the outcomes of the research and to defend them if they are challenged. The potential value of the material for further research should also be considered, particularly where the research would be difficult or impossible to repeat.” (p. 2.1).

It should be noted that this aim is well aligned with ANDS’ vision for “more researchers *reusing* and *sharing* more data more often”. In pursuit of this aim, two sections of the Code talk in particular about data storage.

Section 2.2 states “Institutions must provide facilities for the safe and secure storage of research data and for maintaining records of where research data are stored”.

Section 2.6 states “Researchers must manage research data and primary materials in accordance with the policy of the institution [and] Retain research data, including electronic data, in a durable, indexed and retrievable form.” ANDS believes that in order to meet the aim above, storage of the data alone is not enough.

In order for the data to be *re-used*, it is necessary to provide discovery information to enable a user to discover the existence of a data collection, assessment information to enable a user to decide if the discovered data is of interest to them, and re-use information to enable a user to make use of the data once they have decided it is of interest to them.

This information is most likely to be accurate, comprehensive and cost-effective if captured as close to the time of creation as possible and managed somewhere. Most data stores are not equipped to manage this kind of information at the level of a data collection. For example, a good collection description might be most simply found in a project plan. ANDS has an active program of work to investigate ways of using existing institutional repositories as metadata stores. ANDS will also be funding the creation of add-ons to existing data store infrastructure as well as a standalone solution.

In order for the data to be *shared*, it needs to be accessible. While this access might be through contacting the data owner, more often it will be via a link to some location where the data is stored. ANDS is not funded to provide data storage, either through the NCRIS funded ANDS activities nor the EIF funded ARDC activities. This means that data will need to be stored in some other location. Four important categories of location are the following.

### Institutional repositories

---

Through the ASHER program (<http://www.innovation.gov.au/Section/science/Pages/asherandiap.aspx>), all Australian universities have implemented an institutional repository. Although these have been designed for document objects, many of them are based on software that allows them to store a range of data objects. As an



example, the ARROW repository at Monash University (<http://www.arrow.edu.au/>) contains both protein crystallography raw image data (<http://tinyurl.com/nl88yg>) and ethno-musicology fieldwork recordings (<http://tinyurl.com/kv9692>). Institutions may wish to explore the option of augmenting their institutional repository to include the complete range of research outputs from their researchers. These stores will often be well suited to storing the kinds of metadata outlined above, but may be less appropriate for large numbers of large data objects.

## Institutional data stores

---

A number of Australian universities and large research institutions are putting in place specialized data stores, optimised for large numbers of large objects. An example is the Monash University Large Research Data Store (LaRDS, <http://www.monash.edu.au/eresearch/activities/lards.html>). These data stores are often based on underlying technologies, protocols and file formats — such as SRB (<http://www.sdsc.edu/srb/>), iRODS (<http://www.irods.org/>), NetCDF (<http://www.unidata.ucar.edu/software/netcdf/>) or OpeNDAP (<http://www.opendap.org/>) — that are very different to those used for conventional institutional document repositories. These stores will often be well suited to storing large amounts of data, but probably will not be able to store the rich metadata at object level outlined above, nor the collection level metadata.

## ARCS data fabric

---

The Australian Research Collaboration Service (ARCS) is funded in part by the Department of Innovation, Industry, Science and Research (DIISR) through its Platforms for Collaboration (PfC) capability within the National Collaborative Research Infrastructure Strategy (NCRIS). In particular, ARCS is funded to deliver both the Interoperation and Collaboration Infrastructure component and the Authorisation Services component of PfC.

As part of its work, ARCS is developing and providing a number of tools that allow researchers and research groups to store, share and transport data across institutional boundaries. The main service through which these capabilities are delivered is the ARCS Data Fabric. The ARCS Data Fabric is available to all Australian researchers and can be accessed through a variety of interfaces including web browsers and operating system integration. For further information see <http://www.arcs.org.au/> or contact ARCS at [help@arcs.org.au](mailto:help@arcs.org.au).

ANDS is working with ARCS to ensure that the ARCS Data Fabric meets the requirements of the Code, as well as the ANDS requirements for data sharing and re-use.

## Discipline stores

---

A number of disciplines have well-established locations for storing and sharing data. This storing and sharing often occurs in combination with the publication process. Examples of such stores are the Protein Data Bank (<http://www.pdb.org/>) and the International Virtual Observatory (<http://www.ivoa.net/>). The Protein Data Bank supports the deposition of coordinate data, structure factor amplitudes and phases associated with a particular protein crystal structure. It does not support the deposition of the raw diffraction images that led to the determination of that structure. The World Data Center for Marine Environmental Sciences (WDC-MARE, <http://www.wdc-mare.org/>) is aimed at “collecting, scrutinizing, and disseminating data related to Global Change and earth system research in the fields of environmental oceanography, marine geosciences, and marine biology”. It focuses primarily on georeferenced data and uses the PANGAEA (<http://www.pangaea.de/>) software for long-term storage. Institutions may wish to develop a position on whether they also seek to store outputs from their researchers whose initial deposit is such a discipline store.



This work is licensed under a [Creative Commons Attribution-NonCommercial-Share Alike 2.5 Australia License](http://creativecommons.org/licenses/by-nc-sa/2.5/au/)