

## ANDS and Data Storage

ANDS and Data Storage .....	1
Who needs to know this? .....	1
Obligations and Expectations: Why Data Storage matters .....	1
The <i>Australian Code for the Responsible Conduct of Research</i> .....	1
ANDS goals .....	2
Metadata and Data Storage .....	2
Data Storage Options .....	2
Types of Data Storage.....	2
Institutional repositories .....	4
Institutional data stores .....	4
National data fabric.....	4
Discipline stores.....	5
Evaluating Data Storage Options.....	5
Further Information.....	6

### 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. It outlines:

- obligations and expectations for storing researcher data;
- what the current options are for researchers to arrange storage for their research data;
- what the policy and metadata consequences are for choosing particular storage arrangements.

Data storage is properly out of scope for ANDS' activities, but it is essential in supporting ANDS' goal of making research data more readily available. Different choices of data storage have implications for metadata management and data access, which are in scope of ANDS' activities.

### Obligations and Expectations: Why Data Storage matters

---

#### **The Australian Code for the Responsible Conduct of Research**

The Code (see the 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 highlighted 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).

In pursuit of this aim, two sections of the Code talk in particular about data storage.



- Section 2.2: “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: “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 goals

The central aim of the *Code* as it applies to data storage aligns well with ANDS’ vision for “more researchers *reusing* and *sharing* more data more often”. But to meet this vision, providing storage of the data alone is not enough.

For the data to be *re-used*, metadata (that is, information about the data) needs to be provided and managed. That metadata includes:

- discovery information, so a user can discover the existence of a data collection;
- evaluation information, so a user can decide if the discovered data is of interest to them;
- access information, so a user can work out how to obtain the data, once they have decided it is of interest to them;
- re-use information, so a user can make use the data they have obtained.

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 through a link to some location where the data is stored. ANDS is not funded to provide data storage, either through NCRIS funding or EIF funding. So data will need to be stored in some other location.

## Metadata and Data Storage

Data is *curated* if someone is taking responsibility for managing ongoing storage and access for the data, and for ensuring that its associated metadata is kept up to date. This metadata is most likely to be accurate, comprehensive and cost-effective if it is captured as close to the time of creation as possible, and managed somewhere. If data is not curated, then no-one is working to keep the data accessible, up to date, or discoverable and evaluated through accurate metadata.

Some data stores allow metadata and the data it describes to be managed together in the same platform. This is common if the data store is specific to a discipline: there will often be a well-defined set of information about the data, which can be provided consistently across different research data collections. This also applies to repository solutions (defined below), although the metadata that repositories manage tends to be generic rather than discipline-specific.

Other data stores do not provide for managing metadata, and the metadata about a data collection needs to be managed externally. There is no reason that metadata and data cannot be managed in separate systems; but managing them separately has more risk of the two falling out of sync, and requires a more complex workflow. Support for metadata is an important consideration for choosing a data storage solution, as we will discuss below.

Many of the data stores that researchers are likely to use will not meet the metadata needs we have outlined. To make managing metadata easier for researchers, ANDS has a Metadata Stores programme, which supports building *metadata stores* to supplement data stores. The current offerings from the Metadata Stores programme are described in the Metadata Stores Solution guide, at <http://ands.org.au/guides/metadata-stores-solutions.html>

## Data Storage Options

---

### Types of Data Storage

To understand how data storage solutions address researcher needs, they can be classified in three different ways.

The first classification distinguishes between *discipline-specific* and *generic* solutions. As we just saw, discipline-specific solutions have their data more integrated with metadata, and that metadata is more detailed.

The second classification is based on how solutions support data curation. *Repositories* store well-curated data, with extensive metadata. Repositories have in the past mostly housed publications, as an extension of the library.



Repositories are increasingly being used to disseminate research data as well; but because of their focus on curation, they are constrained in the size and quantity of data they store. *Data stores*, by contrast, are uncured storage space. They are not constrained in size or quantity of data.

The third classification is based on the scope of solutions: how large a community they store data for. There is a continuum of scope, from catering to a single individual, through to international collaborations. Solutions with narrower scope allow more flexibility in how data is stored and managed. Solutions with broader scope, on the other hand, are better resourced for persistently maintaining the data, and make data available broadly. ANDS' goals of promoting reuse and sharing of data are best met by solutions with broad scope; the following narrow scope solutions are problematic for long term access, and do not satisfy ANDS' goals.

Data storage with *individual* scope covers a single researcher's hard drive. This is the default situation for much research data, once the researcher retrieves it from an instrument. However, with data stores on a hard drive (or DVD), it is difficult to arrange access for research collaborators; impossible to arrange dissemination to outside researchers; and vulnerable to backup and access failure. The linkage of metadata to data using individual data storage is typically ad hoc—to the extent that metadata is collected at all.

The next narrowest scope is the *project-based research team*: storage is shared between the researchers involved in a particular project. This enables collaboration within the team, which makes such storage the default for collaborative research. The storage infrastructure however is only resourced for the duration of the project, and may not be well-resourced for publishing data outside of the project; such publishing typically takes the form of a local web server. Storage is often managed by researchers or research assistants, rather than professional IT staff.

The storage solutions considered in the rest of this document are of broader scope: they carry commitment to persist data and make it widely available, and are resourced to make that happen:

- *Institutionally-supported* systems are resourced by research institutions, rather than teams within the institution. These include systems with scope across the whole institution, but also systems with scope across well-defined organisational units within the institution (such as faculties and departments). Storage is normally managed by dedicated IT staff. Solutions applied across the institution are generic, rather than dedicated to a particular discipline.
- Storage associated with instruments producing data output can be either project-based or institutionally-supported types, depending on who manages the instrument. If the instrument is hosted by a facility, the facility should have dedicated resources for managing data storage in the short term, and sometimes also in the long term. This level of support makes those stores count as institutionally supported. However long-term arrangements for data storage are normally made outside the facility.
- *Government-supported* systems are resourced by governments, either at the national or state level. These systems are also usually generic, rather than specific to a discipline. Because they are resourced outside of any one institution, they support cross-institution collaboration more readily.
- Discipline systems usually have national or *international* scope, and are resourced either by a consortium, or by an institution on behalf of the discipline.

Applying these two classifications gives us the following classes of data storage:

*Generic:*

	<i>Individual</i>	<i>Project-based</i>	<i>Institutionally Supported</i>	<i>Government Supported</i>	<i>Discipline</i>
Repository	—	—	<b>Institutional Repository</b>	—	—
Data Store	—	—	<b>Institutional Data Store</b>	<b>National Data Fabric</b>	—

*Discipline-Specific:*

	<i>Individual</i>	<i>Project-based</i>	<i>Institutionally Supported</i>	<i>Government Supported</i>	<i>Discipline</i>
Repository	—	—	—	—	<b>Discipline Repository</b>
Data Store	Individual Storage	Project Team Storage	Department Data Store	—	—



The cells left blank are possible, but not common arrangements. We consider the boldface classes in more detail below.

## **Institutional repositories**

Through the ASHER program (<http://tinyurl.com/npm46r>), 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 cyclone tracking data (<http://arrow.monash.edu.au/hdl/1959.1/79591>) and ethno-musicology fieldwork recordings (<http://arrow.monash.edu.au/hdl/1959.1/52170>).

Institutions may wish to explore the option of augmenting their institutional repository to include the complete range of research outputs from their researchers. This enables research data to be treated the same way as publications as a researcher's intellectual output—particularly for the purposes of bibliometrics and citation tracking. Institutional repositories 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.

As generalist solutions, institutional repositories will not provide the full range of metadata relevant to a discipline: they are reliable by design as a retrieval location, but they are not as effective for the exploitation of research data. This is because, as generalist stores, institutional repositories are not resourced to support a wide range of discipline concerns. Discipline-based discovery at any rate concentrates on stores with broader scope, which is discussed below.

## **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>) and the University of Melbourne's Storage on Request (SToR, <http://its.unimelb.edu.au/storage/>).

These data stores are often based on underlying technologies, protocols and file formats — such as SRB (<http://www.sdsc.edu/srb/>), NetCDF (<http://www.unidata.ucar.edu/software/netcdf/>), OpeNDAP (<http://www.opendap.org/>) or iRODS (<http://www.irods.org/>), — that are very different to those used for conventional institutional document repositories. These stores will often by design be well suited to storing large amounts of data. However they are not intended to store rich object level metadata or collection level metadata; their metadata support is typically limited to keywords or key–value pairs.

Institutional data stores are primarily intended for storing research data, rather than for disseminating it. This is a key difference from institutional repositories, whose primary purpose is to publish resources. External access may be provided for data stores, but it requires authorisation and authentication management: it is not provided as a default.

Metadata stores are intended to bridge the gap in metadata coverage, for both data stores and repositories, by supporting metadata storage for a range of disciplines independently of data storage, and by allowing a range of metadata formats.

## **National data fabric**

The Australian government has recognised effective research collaboration depends on access to shared data stores, where research data can be readily accessed, analysed and re-used. Such data stores can support the retention and integration of nationally significant data assets. A national data store solution is needed, to complement institutional solutions, for storing and publishing data outside the confines of individual institutions.

The Australian Research Collaboration Service (ARCS) was funded in 2007 to provide infrastructure support for research collaboration; this encompassed tools to store, share and transport data across institutional boundaries. The ARCS Data Fabric (<http://www.arcs.org.au/index.php/arcs-data-fabric>) is the service through which the collaboration capabilities is delivered, and includes a common storage component; the Data Fabric is available to all Australian researchers, and can be accessed through a variety of interfaces, including web browsers and operating system integration.



ARCS is concluding operations in mid-2011. The collaborative capabilities of the Data Fabric may continue to be available through other bodies, such as the National eResearch Collaboration Tools and Resources (NeCTAR, <http://www.nectar.unimelb.edu.au/>), although that decision has not been made as of this writing.

The Research Data Storage Infrastructure (RDSI) project to develop a national network of distributed data stores was announced in January 2011. RDSI has the University of Queensland as its lead agent, and is undertaking sector consultations in the first half of 2011; recommendations for an initial set of storage nodes will be made in September 2011. The Data Sharing program within RDSI (DaSh) will be the likely vehicle for that consultation: see <http://www.usyd.edu.au/news/eresearch/2339.html?newsstoryid=6347>.

As of this writing, the national infrastructure for data storage is in transition; this guide will be updated as more details come to light. If users are concerned for long-term storage of data, institutions are currently the best positioned to provide such a guarantee, although persistent storage is within RDSI's intended remit.

## Discipline stores

A number of disciplines have well-established locations for storing and sharing data, managed typically by a consortium of institutional members. This storing and sharing often occurs in combination with the publication process. For example:

- The International Virtual Observatory (IVOA, <http://www.ivoa.net/>) coordinates registries of astronomical data, through its Registry of Registries (<http://rofr.ivoa.net/>). The participating registries are managed by member institutions. IVOA is an example of a federated data store: data policies and metadata are managed centrally, but storing the data itself is shared among members of the consortium.
- The Worldwide Protein Data Bank (<http://www wwpdb.org/>) accepts deposits of structural descriptions of proteins; it is a consortium whose members are themselves collaboratories, and is funded by a variety of sources. The WWPDB data model is more centralised than IVOA's, with a single archive mirrored between its members. The WWPDB does not contain raw image data underlying the structural analysis, but can link to experimental data stored externally.
- 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. This allows it to archive collections for ongoing access.

Discipline stores are often the repository of record for a discipline, and institutions should develop a position on how to interact with discipline stores. Institutions may agree to store experimental data, with the discipline store referring out to the data as a registry. Institutional repositories may store data which has also been registered in discipline stores; if they do, the associated metadata may need to be adjusted for a more generalist audience, and the relative priority of deposit will need to be addressed.

## Evaluating Data Storage Options

---

The goals of "reusing and sharing data more often" are met by storage solutions which make data discoverable and accessible over the long term. That means integration with metadata through curation, for discovery and evaluation of data. It also means that the data storage has to be maintained, to continue providing access: the storage media should be updated, backed up, and reliable.

These solutions described above are not mutually exclusive, but complementary: they are designed to address different needs. In designing a data management strategy, institutions need to harness all the available types of storage. Focusing on just one (such as institutional data stores or repositories) foregoes the advantages of the others, and does not satisfy all the requirements of researchers themselves. For example, institutional solutions are more sustainable than cross-institution discipline solutions, because they have concrete institutional commitments guaranteeing long-term storage: this satisfies the requirement for long-term access. However discipline solutions are where researchers search for data by default, and they provide a more discipline-appropriate environment to do discovery: they are a better fit for the discovery requirement.

We can score the types of solution according to how they satisfy various requirements. Solutions are arbitrarily scored from 0 (poor) to 5 (excellent).



	Supports large datasets	Sustainable	Reliable	Supports curation	Supports researcher discovery	Supports bibliometrics
Individual Data Store	3	1	1	0	0	0
Institutional Repository	2	5	5	5	3	5
Institutional Data Store	4	5	5	0	0	0
National Data Fabric (= Data Store)	5	3	5	0	1	1
Discipline Repository	1	1	4	3	5	3

The only type of solution which does not satisfy the goals of reuse and sharing data are the narrow-scope, individual and project-team solutions. The reason they remain so widely used is their convenience and low cost. To steer researchers away from such quick fix solutions, institutions need to encourage researchers to plan at the beginning of a project for how they will store data, and to budget for it. Such planning takes place in a data management plan, and is described on other ANDS guides; the *Data management planning guide* <http://ands.org.au/guides/data-management-planning-awareness.html> is one starting point

## Further Information

---

ANDS Guides and other Resources: [www.ands.org.au/guides](http://www.ands.org.au/guides)

