

Guidance for researchers seeking long-term data storage

The Pawsey Supercomputing Centre has multiple research data storage options available to the research community. Access to long-term research data storage at the Pawsey Supercomputing Centre is governed by the Pawsey Supercomputing Centre's Data Storage and Management Policy which was approved by the Pawsey Supercomputing Centre Board. The Data Storage and Management Policy is available on the Pawsey Supercomputing Centre website, and all researchers using the Pawsey Supercomputing Centre's data storage facilities must comply with this Policy. All data storage allocations, even where individuals currently have data stored with the Pawsey Supercomputing Centre (for example on Cortex), require researchers to complete the online application form, that is available on the Pawsey Supercomputing Centre website.

- **What is meant by long-term storage?**
Generally, long-term storage is for individuals who have justification for storing research data for more than 12 months. This may be due to a project continuing to create new data for a significant time, or for ongoing analysis purposes, or where researchers are collaborating with others to continue a field of investigation.
- **Can the Pawsey Supercomputing Centre long-term storage be used to store any data?**
All of the Pawsey Supercomputing Centre's long-term storage facilities (funded by the Pawsey and RDSI projects) can be used in the creation or development of any data that is to be used continually, both currently and in the future, for research purposes – in essence, to support “data intensive research”. This means that the Pawsey Supercomputing Centre's long-term data storage facilities cannot at any time be used as a scratch space, or for backup of non-research data (such as laptops, even if these machines are used for research purposes) or codes (even if the codes are for research purposes), or for data protection purposes only (i.e. where there is no plan to actively use the data), or for any other ‘ephemeral’ purpose.
- **What volume of research data will the Pawsey Supercomputing Centre accept for long-term storage?**
The data collection to be stored must already exist (even if it is still under development) and be greater than 5TB¹, before an application is made to the Pawsey Supercomputing Centre. Future predications as to what the volume of data might become are not acceptable. All applications received for storage in excess of 100TB will be automatically referred to the Pawsey Supercomputing



Centre's Data Allocation Committee for review.

^{1*} The Pawsey Supercomputing Centre Executive Director may, in some circumstances, approve access for collections that are less than 5TB, particularly where multi-participant access to the research data is required.

- **Can multiple projects be covered in a single application, and who should get access?**

Only one project will be permitted per directory on any Pawsey Supercomputing Centre long-term data store – irrespective of the status of the researcher or their group. On the conclusion of the application process, the principal data custodian will be asked to sign a declaration indicating who should get access to the research data and under what conditions.

- **What if a researcher joins or leaves an institution and needs access?**

Additional researchers can be allowed access to a specific directory only on receipt by the Pawsey Supercomputing Centre of a signed declaration from the PI who owns the project space/is the principal data custodian. If a researcher leaves an institution and access is no longer required, the PI should inform the Pawsey Supercomputing Centre immediately to allow that access to be removed.

- **What storage is available?**

Data Stores @ Pawsey Supercomputing Centre includes storage from Research Data Storage Infrastructure (RDSI), which is a national collaborative initiative providing significant resources across Australia to the research community. More detail about the Pawsey Supercomputing Centre's role in the RDSI Project is available on the Pawsey Supercomputing Centre website. Storage provided as part of the Pawsey project will also be available soon.

- **Does the Pawsey Supercomputing Centre create an off-site backup?**

Currently, the Pawsey Supercomputing Centre does not create any backup (offsite or otherwise), due to the total data volume involved. In the event of a catastrophic failure, the Pawsey Supercomputing Centre will be unable to recover the data. All researchers are therefore asked to ensure that another copy of their research data is held elsewhere in all cases.

- **Once I have an application, how to I ingest data to the Pawsey Supercomputing Centre?**

LiveARC (also known as Mediaflux) is the principal research data ingest tool for all Pawsey Supercomputing Centre long-term storage funded by the Pawsey and RDSI projects. Specifically, LiveARC:



- Integrates with all of the Pawsey Supercomputing Centre’s long-term storage, meaning researchers can access any Pawsey Supercomputing Centre storage with ease via one “front door”.
 - Can be developed to support many researchers’ workflows via an easy to use web interface.
 - Is provided by many other RDSI nodes, hence researchers can use this tool no matter which node they are using in Australia.
-
- **What is happening with RDSI funded storage?**

The RDSI Project ended in December 2014. At the Pawsey Supercomputing Centre, storage for all WA research data collections owned by researchers based in Pawsey Supercomputing Centre partner organisations and WA Government will be provided on a free of charge basis until at least early 2017. From January 2015, the position regarding the provision of research data services and charging models at other RDSI nodes in Australia is not known. It is the responsibility of researchers to ascertain the status of data services and fees at other nodes independently of the Pawsey Supercomputing Centre.
 - **Can research data be replicated from another RDSI node?**

Research data can be replicated from another RDSI node on a case-by-case basis. Any arrangement is subject to prior agreement by the nodes participating and may include network connectivity (between nodes), to allow research data to transfer securely and efficiently.
 - **How does copyright apply to research data held at the Pawsey Supercomputing Centre?**

Data (wherever it is stored) can be protected under copyright and can be licensed under an agreement with the data author. In most cases this will be the researcher’s employer (e.g. a University or Government body). Access to data under copyright to others is the responsibility of the data author; the Pawsey Supercomputing Centre has no role to play in any data copyright agreement.
 - **Is copyright the only legal restriction on research data?**

Ensuring confidentiality, contractual obligations and requirements that govern the appropriate/ethical use of research data (sometimes known as sensitive data or personal information) is the sole responsibility of the principal data custodian. Please refer to the Pawsey Supercomputing Centre Data Storage and Management Policy for further advice.

