Data Preservation Principles

Purpose

CSIRO recognises significant value in the data generated by its substantial investment in research. Durable research data is essential to justify, and defend when required, the outcomes of research. The Data Access Portal (DAP) provides the basis for management of CSIRO's digital research data assets in terms of storage, retention, and accessibility for reference, use or reuse. For the purposes of these Data Preservation Principles, research data assets are those selected for long term storage and management to enable validation of research findings and re-use of high value or unique data. These Data Preservation Principles describe CSIRO's approach to the processes and responsibilities of long term retention and preservation of data assets held in the DAP for use by its user community.

The user community of the DAP includes researchers, students and policy makers from a broad range of the sciences and the wider community. To ensure long term access to the collections in the DAP, CSIRO commits to maintaining the authenticity, reliability and logical integrity in formats suitable for reuse.

Mission

The mission of the Data Access Portal is to publish research data of significance to CSIRO and ensure its ongoing access.

The objective of the CSIRO Research Data Service is to ensure that the organisation captures, publishes and manages the right data to support innovation, collaboration and scientific integrity. The Research Data Service will:

- Ensure research data under the custodianship of CSIRO is securely stored, easily located and where appropriate, accessible to others for reuse;
- Provide the ability to publish data to support reproducibility and research integrity;
- Manage data in an increasingly collaborative research environment; and
- Respond to government and funding body requirements to share the data rising from publicly funded research.

The service is a collaboration between Information Management & Technology teams as well as research partners to deliver a holistic solution for the management of research data in CSIRO. The DAP is the point of capture, and discovery portal, for CSIRO’s research data assets.

Through Data.gov.au, Research Data Australia and other subject discipline portals, CSIRO will make available data collections and/or descriptions of data collections which are across a broad range of disciplines and will increase the availability of the data.

Scope

The scope of these Data Preservation Principles are limited to the collections in the DAP. The DAP includes data assets generated by CSIRO and may include third-party data that has been subject to a data deposit agreement and is aligned with our Data Collection Development Principles. CSIRO is committed to preserving the data that falls within our scope of responsibility.

Objectives

The data collections preservation service at CSIRO is designed to meet these objectives:

- An integrated and interoperable data ecosystem with;
- Established accountability for data at all levels;
- A default assumption of openness, at the same time, ensuring that licensing, ethical and contractual obligations are honoured;
- Supported by data management and data governance tools;
- Acknowledgement of CSIRO culture and specific needs through development of a data governance-oriented social architecture;
- An approach where data is valued by the organisation and is managed in a way that enables us to realise value;
- To preserve the data assets in the DAP permanently; and
- To be a trusted digital repository.

Legal and regulatory framework

CSIRO is an Australian Government corporate entity, with a Board and Chief Executive. We're constituted by and operate under the provisions of the Science and Industry Research Act 1949, which sets out our functions and powers, as well as those of our Minister, Board and Chief Executive. The governance, performance and accountability of our operations, including the use and management of public resources are set out in the Public Governance, Performance and Accountability Act 2013 and related rules.

The legal and regulatory frameworks the DAP follows for preservation and access to datasets include:

- Archives Act 1983;
- Australian Code for Responsible Conduct of Research;
- Copyright Act 1968;
- Digital Continuity 2020; and
- Privacy Act 1988.

Depositors are responsible for:

- Specifying the conditions under which access to a collection is made available to the user community;
- Acknowledging that all necessary permissions for copyright and intellectual property rights have been cleared; and
- A relevant licence agreement enabling the DAP to distribute the collection.

CSIRO Officers acknowledge their responsibilities at the time of deposit by having worked through a Data Deposit Checklist or Software Licence Selection Process. A collection deposited by a CSIRO Officer is subject to an internal peer review. Third-party collections may be included in the DAP subject to assessment by a CSIRO Officer and on agreement and acknowledgement of the data deposit conditions.

Roles and Responsibilities

Staff responsible for implementing these Principles are from the Information Management and Technology department specifically: Information Services, Research Application Development and Data. The Data Management Capability Enhancement Program ensures the DAP operates on best practice principles in data asset preservation and technical capability. This multi-disciplinary team comprises: project board, user reference group, project manager, data librarians, business analysts, software developers, testers, infrastructure specialists and database administrators.

Project Board: Provide guidance on overall strategic direction, scope and priority.

User Reference Group: Provide validation and evaluation of user requirements.

Project Manager: Responsible for the overall success of the activities.

Business Analysts: Produce business requirements in consultation with the user community.

Software Developers: Ensure technical and architectural integrity is developed and maintained. Ensure adherence to standards of best practice.

Testers: Ensure adherence to testing standards are maintained.

Infrastructure Specialist: Provide consultation and expertise for infrastructure requirements and media refreshment.

Database Administrator: Provide database expertise for system functionality.

Data librarians: Liaise with the Data Depositors, Approvers and Data Users. Provide advice and training to Data Depositors and Approvers.

Implementing the preservation strategy

During the development of the DAP in the early 2010s, the Reference Model for an Open Archival Information System (hereafter referred to as OAIS Reference Model) was adopted to inform the processes for preserving originally submitted content.

The DAP's processes are outlined in the RDS Functional Model shown in Figure 1 below.

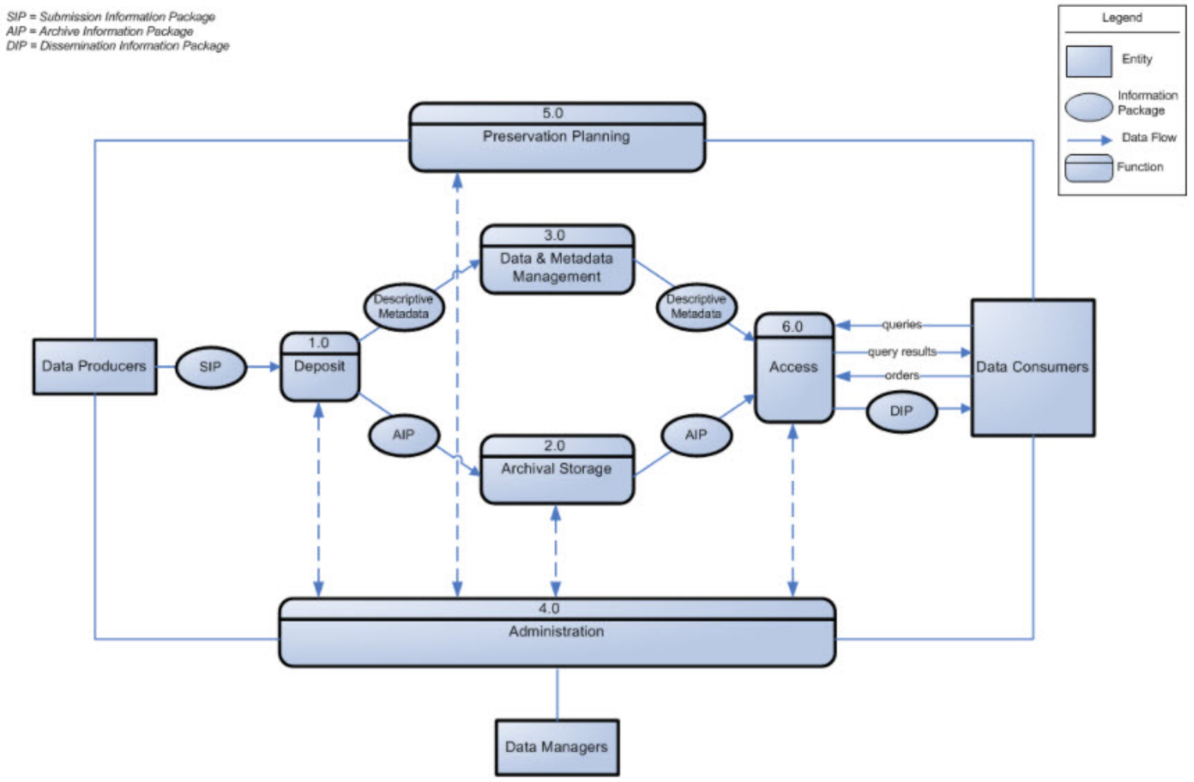

Figure 1 CSIRO RDS Functional Model

Deposit Checklist

To follow the guidance of the OAIS Reference Model, to accept appropriate information, the data depositor and approver are required to complete a Data Deposit Checklist or Software Licence Selection Process. These documents are off-system tools to guide data depositors and approvers through key contractual, legal, ethical and quality questions related to publicly publishing collections. Data depositors produce a draft collection that they submit to an approver. The approver has the appropriate delegation and authority rank and performs the final check before a collection is published. The benefits of the Deposit Checklist are the data depositor and approver consider key issues prior to deposit that impact on quality and preservation. Data depositors and approvers have access to guidance in the form of written guides, training and consultation with staff from a wide range of disciplines within CSIRO including: commercial contracts, legal, ethics, data support and information technology.

Deposit

The deposit function, or ingest function of the OAIS Reference Model, accepts submission information packages (SIP) from the data depositor and verifies the integrity and completeness of the information. The files in the SIP are kept as originally supplied and are retained for preservation. An Archival Information Package (AIP) is generated for the SIP where metadata is extracted for search and retrieval and is added to the Data and Metadata Management database. Files are transferred into Archival Storage. Links between the supplied SIP and AIP are maintained. All processes undertaken in the system are kept in logs. Data depositors and approvers are notified when key processes such as publish are complete.

Archival Storage

The Archival Storage functions include long term storage, maintenance and retrieval of AIPs. This function ensures the SIPs submitted remain the same as deposited and are accessible. Archival storage adds the AIPs to long term storage and provides management over time: manage storage hierarchy, replace media, error checking, disaster recovery and provides files for the Access function.

Data storage is fit for purpose and meets the recommended practice from the OAIS Reference Model. At the collection level, there is geographic diversity of data. A collection is stored in at least two default locations with data depositors able to choose up to another two locations. Collections in each geographic location are generated from the same source copy and part of the process applied by the collection management middleware (LCM - Logical Collection Manager) ensures that the data stored within the collection is a true copy of the source data. If LCM cannot validate this at time of creation an error occurs which alerts an administrator to resolve the problem. Part of this process involves the creation of a cryptographically strong checksum of the stored data as a whole which is validated periodically by LCM on collection mount. Bad storage media is detected from log entries from the hierarchical storage management (HSM) storage backend. When errors are detected, the media in question is manually flagged as suspect and all data is attempted to be migrated from it to a replacement tape, where all files cannot be recovered from the media, the recovery process requires the validation and copying of the collection from a secondary storage location.

Data and Metadata Management

Data and Metadata management functions are responsible for maintaining the integrity of the database and includes descriptive metadata, finding aids and system information. Descriptive metadata describes a collection and is used to support archive operations. This function performs queries requested from Access and generates a result for the user community. Reports are generated for requests from Deposit, Access or Administration and can include: summaries of holdings or usage statistics. Updates are received from Deposit for new AIPs and Administration for system updates.

Version Control

To maintain the authenticity of a DAP collection any alteration is recorded accurately through the use of version control. Changes to metadata and/or files in the DAP create a new version. The previous file(s), AIPs and DIPs are retained. The current version is returned in query results. A new DOI for a collection is minted when the files are changed in any way, or for a substantial change to the metadata fields that make up the attribution statement after review by a data administrator.

Data Collection withdrawal

If the data depositor chooses to withdraw a collection, effectively removing the data from the public domain, a review task is assigned to the data administrator. The data depositor will be contacted to discuss the reason for withdrawing a collection from public access. After discussion with the data depositor and, possibly, data approver, a 'soft deletion' of the collection will occur to withdraw the collection. Hereafter, only the data administrators and data depositor will have access to the collection. DOIs for withdrawn collections will resolve to a page informing the user the collection has been withdrawn. The process of 'hard deletion' of a collection, effectively removing the files, presents risks to other parts of the system and hence 'soft deletion' is the default method.

Access

The Access function of the DAP provides the Data User with an interface to find, access, ask for assistance or provide feedback. The Access function accepts a DIP from the user community and then retrieves the AIP from Archival Storage and moves a copy of the data to a staging area for further processing. The user community accesses the DAP using the online search interface and assistance is available via email and telephone. In addition, the Access function implements the security that a data depositor has applied to their collection.

Preservation Planning

The functions of preservation planning are to monitor the needs of the data depositors, approvers and user community. Additionally, this function monitors technology to ensure ongoing system suitability for the continued preservation and access to the DAP's collections.

Administration

The Administration function includes the day-to-day management of the DAP for Deposit, Archival Storage, Data and Metadata Management, Access and Preservation Planning functions. Managing the interaction with data depositors, approvers and the user community is an additional function of Administration. It is responsible for establishing standards, policies and monitoring system performance. Administration oversees the archiving and access systems of the DAP and ensures they are kept up-to-date.

Content coverage

CSIRO is committed to ensuring the long-term availability of the data it holds by ensuring technology is adapted to changes in storage and application technologies. The DAP includes a broad range of file formats in its collections spanning many scientific disciplines within CSIRO. A wide range of file formats are accepted as preservation is based on the common practices of a discipline. Yet, file formats may become obsolete due to software or hardware dependencies. Preference is given to data formats that are simple and accessible by many software applications and the user guide has a list of preferred formats.

IT Architecture

The IT architecture for the preservation of the collections in the DAP is fit for purpose and developed and maintained by CSIRO. The standards used as reference include the OAIS Reference Model and World Wide Web Consortium. Plans for infrastructure development are considered within the Data Management Capability Enhancement Project.

The application for the DAP is developed within CSIRO and there are at least three releases of the software each year. A record of the software configuration is maintained in a Bitbucket code repository and the build is automated using the Gradle Build tool. The release history of the software is publicly available to the DAP user community. Third party software used in the DAP application are listed in the acknowledgements.

The data assets are stored within CSIRO in multiple geographic locations with two locations by default and up to four in total for a collection. When data assets are retrieved from offline storage from within CSIRO they are checked for data corruption and if found are copied from another location.

Security

Appropriate security measures ensure DAP collections are protected from unauthorised use, accidental modification or loss.

Security for the storage of collections in the DAP is restricted to the Administration team. Their responsibilities are outlined in the CSIRO Information Security Procedure. All administrators have obtained a security clearance that is administered by the Australian Commonwealth Government. Server rooms are located in multiple locations across Australia. Storage of DAP collections are mirrored in two data centres by default and data depositors can choose to mirror a collection in another two centres.

For the DAP application backups are performed on critical information in the system and documentation exists to permit the creation of replacement servers which would allow restoration of service within 2-3 days. Disaster recovery and business continuity procedures have been documented.

Data is secured and users traceable for authenticated access for depositing and accessing restricted collections. The Administration team is able to track user interaction with the application using a combination of application logs, server logs, user access logs, Google Analytics and the DAP database which also records data download requests.

To test the security measures employed by the DAP, the Security team undertakes security tests (including penetration tests) prior to a new release being deployed. The Server Ops and Database team undertake regular maintenance of the underlying infrastructure by applying security patches as required.

Sustainability

CSIRO is an Australian Government corporate entity. CSIRO's parent began as the Advisory Council of Science and Industry in 1916. CSIRO is constituted by and operates under the provisions of the Australian Government Science and Industry Research Act 1949. CSIRO's main funding is directly from the Australian Government and this funding is in excess of 1 billion per year. As this funding is based on a triennial funding program, if funding was to cease, CSIRO would have approximately 3 years notice to plan for the succession of the DAP's collections. CSIRO considers the possibility of it ceasing to operate or changing its scope as highly unlikely.

Acknowledgements

These principles were developed in consultation with the following documents:

Data Archiving and Networked Services (DANS), 2017, Preservation Policy, Version 1.2, viewed 17 November 2017, <https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreservationpolicyUK.pdf>.

Inter-university Consortium for Political and Social Research (ICPSR), 2009, Principles and Good Practice for Preserving Data, International Household Survey Network, IHSN Working Paper No 003, viewed 17 November 2017, <http://www.ihsn.org/sites/default/files/resources/IHSN-WP003.pdf>.

UK Data Archive, 2016, UK Data Archive Preservation Policy. <http://www.data-archive.ac.uk/media/514523/cd062-preservationpolicy.pdf>.