Preservation

Policy number: PP 10
23th March 2018

Table of Contents

Outline 2
Background 2
   Organisation 2
   Funding 3
   Roles and Responsibilities 4
   Task Forces 4
External Auditing and Certification 6
Ingest 6
Privacy 7
Storage and Integrity 7
Data Recovery 9
Security 9
Audit 9
Ensuring Future Access, Usability and Understandability 10
Data withdrawal 11
Relocation or transition of activity 12
Outline
This document describes the policies and procedures at Digital Repository of Ireland (DRI) that support ongoing preservation of deposited data. The DRI takes a broad view of preservation, recognising that it is not merely a matter of secure, reliable and error-free storage of data. It also encompasses provision of retrieval and access to data, now and in the future, as well as issues of ongoing understandability and reusability of that data. Furthermore, ensuring long-term preservation also requires consideration of issues of organisational continuity and sustainability, and planning for the transfer or hand-over of responsibility for data if it were to transpire that the DRI could no longer fulfil its preservation function.

This document is broken into sections that address the relevant areas and are described as high-level activities within and involving all areas of the Repository. Other documentation has been developed that outlines the more operational, low level activities and procedures and are referenced where relevant. In addition, an overall view of the Repository Infrastructure is given in the report Building the Digital Repository of Ireland Infrastructure1 for those interested to learn more about the specifics of implementation.

These preservation functions are constantly in development and this document will be updated accordingly in future.

Background
The Digital Repository of Ireland is the national trusted digital repository for Ireland. The DRI Mission Statement is as follows:

The Digital Repository of Ireland is a trusted national infrastructure for the preservation, curation and dissemination of Ireland’s humanities, social sciences, and cultural heritage data2.

The repository links and preserves data held by Irish institutions, providing a central internet access point and interactive multimedia tools. As a national e-infrastructure for the future of education and research in the humanities and social sciences, DRI is available for use by students, scholars as well as the public and acts as a focal point for the development of national guidelines and policy for digital preservation and access.

Organisation
DRI is made up of three long-established and respected partner institutions with expertise in key areas of archiving, repository infrastructure, and HSS domains: the Royal Irish Academy, Trinity College Dublin and Maynooth University. DRI staff are located across these three institutions with each institution and staff member providing unique experience and access to a wider group of experts who actively support work undertaken by DRI. The repository is hosted on servers in

---

1 Digital Repository of Ireland. Building the Digital Repository of Ireland Infrastructure, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.qr474f68n
2 http://www.dri.ie/about Accessed 12th October 2017
Trinity College Dublin with mirror servers in Maynooth University and supported by software developers in Trinity College Dublin. The Policy Manager is based in Maynooth University. The Director, Digital Archivist, Education and Outreach Manager and Programme Manager are based in the RIA, and are supported by additional, project-based staff recruited to expand DRI capacity while contributing to core work. Institutional IT departments provide access and support to email, networking and other IT support services for staff.

The principal governance structure of the DRI is the Board. The DRI Board is comprised of members from across the Irish Archiving, Digital, Education and Cultural sectors who provide advice on the implementation of DRI’s core activities and future endeavours. The principal management mechanism of the DRI is the Core Implementation Team (CIT). The CIT is responsible for the day-to-day operational management of the DRI in addition to overseeing requirements and developing the DRI strategy. The CIT is comprised of the DRI Director (Chair) and DRI Principal Investigators and the DRI Programme Manager.

The DRI mission statement is part of DRI’s business plan, which has been approved by DRI’s funders – the Department of Education and Skills, via the funding mechanisms of the Higher Education Authority and the Irish Research Council – and by DRI’s governance bodies - the Management team (‘Core Implementation Team’) and the founding Board.

DRI runs a number of partnership leveraged projects targeting development of additional features and collections. DRI has a comprehensive education and outreach program that includes training courses, seminars, workshops, conferences, fact-sheets, user guides and other publications.

Funding
From 2017 DRI has received its core funding from the Irish government. The Digital Repository of Ireland is named as a key national infrastructure in the Department of Business, Enterprise and Industry Innovation 2020 strategy on research and development. Through this recognition the Irish government has committed to long term funding of DRI.

DRI’s core funding is augmented by the acquisition of leveraged research project funding, as well as a revenue-generating membership model, which was introduced in 2018. The membership model focuses on increasing the volume of material deposited in the archive, but also on the diversity of member organisations. This contributes to the overall sustainability of the repository by raising revenue which can be reinvested into the repository. In addition, regular

3 http://www.dri.ie/events Accessed 11th January 2018
4 http://www.dri.ie/publications Accessed 11th January 2018
engagement with our members (e.g. via the Biannual Members’ Forum) will ensure we are meeting the long-term needs of the humanities and social sciences in Ireland.

Roles and Responsibilities

**Director**
The director controls and governs DRI and ensures that legislation is complied with, sustainability is ensured and that policies are implemented.

**Programme Manager**
The Programme Manager works closely with the Director and is responsible for project planning, preparing reports for project stakeholders, budgeting and financial records maintenance, and meeting administration.

**System Developer**
The developer maintains and develops the repository code in accordance with requirements and policies. He/she collaborates where necessary with the System Administrator on implementing policies and disaster response.

**System Administrator**
The System Administrator carries out day-to-day infrastructure activities including ensuring the repository is functioning correctly, software and hardware are up-to date, alerts are responded to and policies/procedures such as backup, preservation, security and disaster recovery are followed and monitored.

**Digital Archivist**
The Digital Archivist maintains and develops collections by interacting with collection owners and promoting digitisation and preservation.

**Training and Outreach Officer**
The Training and Outreach Officer is responsible for fostering and developing a preservation community in Ireland and running workshops and training and education activities related to preservation and digital archiving.

**Policy Manager**
The Policy Manager researches, develops, delivers and advocates for DRI policy, particularly pertaining to long term digital preservation and research data management in line with national and international practice.

**Task Forces**
DRI has a number of permanent or ad-hoc task forces and working groups convened to deal with specific cross-functional areas of concern. These often include non-DRI staff recruited from among our member organisations who provide expertise in particular areas. At present these Task forces are as follows:
- Requirements Task Force - Responsible for ongoing requirements gathering, specification and analysis. DRI's original Requirements are published in the DRI Requirements Specification document.
- Workflows Working Group - Responsible for developing requirements into actionable Feature Specifications.
- Business Models - Responsible for developing the DRI Business Model.
- Business Records - Responsible for keeping DRI business records.
- Trusted Digital Repository Task Force - Responsible for TDR compliance and certification.
- Metadata Task Force - Responsible for making recommendations on Metadata and File Formats.

Other short-term task forces and working are created and dissolved as required.

An overview of how some of these task forces work together and with other functions of DRI is shown in Figure 1.

---

Figure 1: The DRI Organisational Structure

---

6 Digital Repository of Ireland. DRI Requirements Specification, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.wd37kb264
OAIS
The DRI project broadly follows the Open Archival Information System (OAIS) reference model. The Reference Model establishes a common framework of terms and concepts which constitute a digital repository and lays out the functional components and responsibilities of such a repository at an organisational level. These terms and concepts are used in this document.

External Auditing and Certification
The DRI is committed to becoming a Trusted Digital Repository (TDR). This requires certification and/or auditing by external bodies or institutions. The European Framework for Audit and Certification of Digital Repositories has been chosen as our certification path. This framework has three steps: first gain the Data Seal of Approval (DSA), then self audit ISO 16363 (Audit and Certification of Trustworthy Digital Repositories) and finally external audited and certified ISO 16363. Currently DRI is at step 2: DSA certified and are self auditing ISO 16363. In 2018, DRI applied for the Core Trust Seal, the successor to the Digital Seal of Approval.

Ingest
DRI’s Collection Policy provides an overview of the types of data that we aim to preserve and informs potential depositors on appropriate content for ingestion. Although DRI provides recommendations on formats for ingest, non-standard formats can also be ingested. Valid metadata with all mandatory fields filled, together with institutional affiliation are required. The Repository supports the most commonly used metadata standards in Ireland: Qualified Dublin Core, EAD, MODS and MARC21 encoded as MARCXML. Information on the data producer is captured in the descriptive metadata.

In addition, associated assets and contextual information may be deposited where applicable and according to the rules and ethics of individual disciplines. These provide information on the data creation process which allows future researchers to more fully understand them. Where the digital objects are generated through research conducted on people the depositor is asked to include the legal or ethical approval they obtained. DRI provides guidance outlining the types of documentation which would be useful for the depositor to include with their contextual information documents, along with information on how to deposit them within the system.

---

7 https://public.ccsds.org/pubs/650x0m2.pdf Accessed 22nd March 2018
8 http://www.trusteddigitalrepository.eu/ Accessed 22nd March 2018
10 http://www.iso16363.org/standards/iso-16363/ Accessed 22nd March 2018
12 Digital Repository of Ireland. DRI Collection Policy, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.s465jx541
13 Digital Repository of Ireland. How to DRI: Contextual Information, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.sn00qc64j
Before ingesting into the Repository, the DRI Deposit Terms and Conditions\textsuperscript{14} must be signed, via a tick-box. This agreement defines the role and responsibilities of the depositor and DRI and details the guarantees on the deposited data. The DRI Depositor Agreement ensures that the original creators of all deposited data retain ownership, copyright and associated intellectual property rights. In signing the DRI Depositor Agreement, the depositor warrants that they are the owner of the copyright and any associated intellectual property rights of the digital objects; or they are duly authorised by the owner, or owners of these rights, and capable of granting a licence under the terms outlined in the agreement.

During ingest, technical metadata is generated by a file characterisation process and the asset is saved to preservation storage. Metadata is saved to the Repository and is then pulled from Repository data streams and added alongside the asset to create the Archive Information Package (AIP). Descriptive, preservation, provenance, technical and version metadata are saved as well as checksum manifest files.

All objects ingested into the repository can have a persistent identifier assigned. This takes the form of a Digital Object Identifier (DOI)\textsuperscript{15} minted by Data Cite.\textsuperscript{16} DOIs assist in preservation by providing a permanent identifier and a link to the object which is more stable than a url.

Privacy

The Repository allows a data owner to set access restrictions on their data in accordance with the DRI Restricted Data Policy.\textsuperscript{17} These access restrictions are implemented by storing the permissions as technical metadata alongside each object or collection. Users may request access to data via the web interface.

Storage and Integrity

Preservation requires that the ingest, storage and integrity of deposited assets is handled correctly and to a high standard. Bit-level preservation is built into the system at multiple levels using, from the lowest level to the highest:

1. The Ceph Storage System\textsuperscript{18} supporting data scrubbing, checksums, replication and erasure coding
2. An Archival Information Package (AIP) format supporting manifests and checksums
3. External data integrity checks to ensure all data remains intact
4. Backup procedures to ensure that recovery is possible in the event of failures

\textsuperscript{14} Digital Repository of Ireland. DRI Deposit Terms and Conditions, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.rj43ck509
\textsuperscript{15} http://www.doi.org/ Last Accessed 22nd March 2018
\textsuperscript{16} https://www.datacite.org/ Last Accessed 22nd March 2018
\textsuperscript{17} Digital Repository of Ireland. DRI Restricted Data Policy, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.sb39mq22h
\textsuperscript{18} http://docs.ceph.com/docs/master/ Last Accessed 22nd March 2018
Archival storage is provided by the Ceph Storage System in a storage pool using erasure coding and replication. Erasure coding is used to provide RAID style data security. Ceph regularly ‘Scrubs’ its data to check and fix bit-flip/decay errors.

All of the metadata and data comprising an ingested digital object is stored in an Archival Information Package (AIP). The AIPs are stored on the replicated Ceph storage pools and use the Moab Versioning format\(^\text{19}\) shown in Figure 2. This is a file packaging format for organising files and metadata on disk or in transport. It is similar to the BagIt\(^\text{20}\) format but supports built-in versioning.

The Moab AIP contains the data, descriptive XML metadata, and technical, administrative and preservation metadata for the digital objects. It also stores various manifest files listing all files for each version of the digital object and their checksums. A manifest checksum file contains checksums of the Moab manifest files, giving a double-layer integrity check.

When a new version of an asset or metadata file is ingested, it is added to the package alongside previous versions. The package is addition only: no files are ever updated or deleted.

![Figure 2: The Moab Versioning Format](image)

The integrity of AIPs is of the highest priority. A process checks each package by opening it and recalculating its checksums. This is then compared with the stored manifests. Checksums are

---


also stored in a centralised database and the process compares against this to guarantee that storage has not been tampered with. In the future, the results may be written as preservation metadata into the package, although at the moment they are simply available as a report which allows the system administrator and the asset owner to view the integrity check results in the object audit.

AIPs reside on a Disk/Disk-to-Disk-to-Tape system. An AIP is first created on a fault-tolerant hot storage pool. This pool is replicated to a second data centre that is geographically distant. The pool is backed up daily to cold storage, first to disk then to tape. These tapes are removed offsite to a secure safe. Tapes are monitored, cycled and tested to minimise defects. Between hot and cold storage, the archive packages are stored in 4 locations. All locations are secure data centres that have restricted and monitored access, fire suppression systems and battery backup power.

Data Recovery
When a data integrity incident is reported, our data recovery procedures are followed. In general, the response involves the administrator of first contact assessing if a single package or many are impacted. The scale of the incident is assessed and the data is recovered from cold storage. Data is recovered as a new version into the AIP alongside corrupted version(s) and version/provenance metadata updated accordingly. Collection managers are notified on detection of and response to a data integrity incident.

Security
Preservation of repository assets requires that access to and protection of the infrastructure is controlled and monitored. This area is well covered by general IT standards and guidelines which are being investigated and will be applied as appropriate. The security procedures are detailed in the DRI Security Procedures document (Internal use only).21

Physical servers are hosted in modern data centres where access is controlled and monitored. Servers are protected to industry standards including access and monitoring. Patching and updates are carried out according to a regular update schedule.22 The infrastructure is modelled on well-defined web application infrastructures that reduces the risk of attack. Encryption is used for data in transport to the repository. DRI staff handle data and access to it sensitively and abide by DRI’s code of ethics.23 Risks to the infrastructure are identified and assessed in accordance with our Crisis and Disaster Management Policy.24

Audit
The repository maintains an audit of all changes made to the metadata or data assets of objects and collections. Much of this audit information is stored within the Moab versioned AIP although

---

21 DRI Security Procedures Document: Internal use only
22 DRI Systems Upgrade Operational Procedure: Internal use only
24 DRI Crises and Disaster Management Policy (internal use only)
some additional records are stored in the Fedora Repository System and in the System Database.

A report is available to Manager Users that enables them to see all past versions of the object along with information about the user who made any changes.

User actions within the system are also audited. A report is available to Organisational Managers which shows all changes to each user’s permissions within the system.

**Ensuring Future Access, Usability and Understandability**

As described in the Ingest section of this document, DRI collects descriptive metadata from the data owner at ingest time. The DRI’s Training and Outreach Officer is active in ensuring that data owners understand the importance of providing rich and detailed descriptive metadata to aid discovery and future access and reuse. Data owners are also encouraged to provide contextual information and documentation for their collections including details about how the data was gathered, the survey or research methods used, the software used to create certain data assets, and so on. In addition, the data owner must specify rights and reuse information, funder information and access controls. The Repository then automatically generates administrative, technical and provenance and preservation metadata. This wealth of metadata for collections and digital objects greatly enhances the future understandability of the collections along with their reuse potential.

In addition to metadata, DRI understands that file formats may fall out of use and become harder to open and manipulate, especially where proprietary formats are involved. DRI recommends certain well-supported, preservation quality formats for ingest into the repository. A list of these formats is published in the DRI File Formats Fact Sheet, which is regularly reviewed. DRI does not, however, mandate that these formats be used.

To ensure continued access, DRI undertakes a yearly format watch process whereby the status of formats contained within the repository is reviewed and a migration plan developed and implemented for any found to be at risk. This process involves first generating a report of all formats currently in the repository, and then investigating these formats on a case-by-case basis. A variety of resources are consulted to help with this undertaking, for example, PRONOM, the Library of Congress Digital Formats website and the Digital Preservation Coalition Technology Watch Reports.

---

25 [http://fedorarepository.org/] Last Accessed 22nd March 2018
26 Digital Repository of Ireland. DRI Factsheet No 3: File formats, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.rj43ck402
When a format is identified as at-risk, format migration is considered as the first approach. This involves creating a new copy of the asset file in a more current standard which maintains all of the essential characteristics of the original file (insofar as this is possible in the new format). New formats are chosen carefully on a case by case basis. Migrated files will be stored alongside the original files and will form a new version. The technical metadata of the object will be annotated to indicate the reason for the change and any necessary changes that may have been made to the file in the process of migration. The repository provides a report for the administrator user showing the number of files of each format within the Repository so that objects in need of migration can easily be identified.

A secondary approach being considered is emulation. This involves re-creating and packaging an environment, including software and potentially even hardware, to allow the original file to be accessed even if the end user does not have access to the software and hardware used to create and view the original file. This is a more complicated and technical process, but as emulation and virtual machine tools and software develop it is becoming a more viable option.

Emulation is likely to be particularly suitable for certain types of digital object which do not easily lend-themselves to format migration, such as software packages. To date, DRI members have not attempted to ingest any data which would not be relatively straightforward to deal with using format migration. As this changes in the future, more concrete procedures for emulation will have to be developed for the Repository.

Whether format migration or emulation is used, preservation metadata and other technical metadata will contain all of the relevant information about any changes to the object. This ensures provenance of the data is stored and will aid in future understandability.

Data withdrawal
DRI policy distinguishes between ‘soft deletion’ (‘unpublish’) whereby certain references to the withdrawn content are deleted, but not the content itself, and ‘hard deletion’ whereby the content and all references to it are deleted, including back-ups. In the case of soft deletion, the data is only accessible to DRI staff and the depositor. Soft deletion is the default method of withdrawal provided as hard deletion is a more expensive process and could present unacceptable risks to other parts of a collection.

DRI may in exceptional cases undertake hard deletion of collections. There may be an administrative charge to cover the cost of this service. Applications to hard delete collections will be reviewed to determine acceptance/non-acceptance.

- In cases of the withdrawal of data, the administrative metadata is updated, and the external view of the catalogue record is updated to reflect the change of status of the

---

30 Digital Repository of Ireland. DRI Withdraw Data Policy, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.rb699s823
collection (with the dates of its availability and where appropriate the reasons for withdrawal).

- In the case of hard deletion, the DOI will refer to a landing page that states that the object was removed from the Repository by the depositor.

Relocation or transition of activity
The legal agreements signed by depositors grants to the repository the right to assign its benefits, rights and obligations to a third party in a situation where it is necessary. Should DRI funding cease, the repository partners would determine a final location for the repository and its deposited data, either at one of the partner’s institutions, or at one of the many institutions that DRI partners with, such as the National Library or the National Archives (both legally mandated heritage institutions). The repository and national infrastructure that constitutes DRI is maintained by three national partners with technical and domain expertise in digital archiving: the Royal Irish Academy, Trinity College Dublin and Maynooth. The repository’s servers and digital expertise are located in Trinity College Dublin, and backup servers are located in Maynooth.

Should any depositor voice dissatisfaction with this plan, the DRI’s partner institutions, in collaboration with owners/depositors, would provide download access to their data and its deletion from the DRI Repository thereafter. This is also outlined in the Organisational Manager Agreement, which must be signed before any data is deposited.