Collection Ingest, Management and Preservation

DARIAH 2010

DARIAH
Digital Research Infrastructure for the Arts and Humanities

Table of Contents

Executive Summary 3

Purpose 3

Scope 4

Audience 5

Responsibilities of a data managing/preservation organization 5

Data Ingest 6

Pre-ingest 6

Ingest responsibilities 6

Ingest procedures 7

Management 9

Preservation Strategy 9

Archival Storage 9

Data Management 10

Safety and Security 11

Executive Summary

DARIAH's mission is to foster and support high quality digital research in arts and humanities. Part of this remit involves promotion of the importance of sustaining and preserving data for future use, so that scholars can benefit from and build upon earlier work in their field. The maintenance of data which is used and created by arts and humanities scholarly working in the digital realm is of paramount interest across the academy, whether in regard to access to online journals or primary source material held in academic or cultural institutions. Without adequate planning the data and outcomes of digitally-enhanced research are unlikely to survive for the long term, and will not be available to future generations of researchers.

This policy will provide a framework for understanding and evaluating the management and preservation of data. It will outline the fundamental processes of this crucial activity, with the aim of allowing data creators, users and curators to become partners in the long term maintenance of the data. In doing so, this policy will expand knowledge about the necessity of data preservation, and provide DARIAH members and users with a basis for understanding the fundamental processes involved in data management. In the long term the aim is to have a positive impact on the lifecycle of digital research data to the benefit of the whole of the research community.

Purpose

Many of DARIAH's constituent organizations are directly engaged in the collection and preservation of data. All of the users of DARIAH will have an interest in the maintenance of the data which they use and create. The activities of both providers and users of data require policies in the areas of collection management, ingest of data, and preservation. This policy sets a standard for DARIAH organizations to adhere to, and allows users to understand the processes for the long-term management and preservation of their data.

The preservation of data is at the core of the enterprise of digital arts and humanities. Without robust policies for collection management and preservation data will not be available for later use and referral by future researchers, nor will it have the necessary elements to allow it to be discovered and shared through the DARIAH infrastructure.

This policy is a codification of good archival practice across the network of DARIAH partners and in the wider digital arts and humanities community. It also draws on wider experience of preservation within the academy more broadly.

This preservation policy is based upon some basic principles:

- In order to fulfil its value to the research process data must have a significant life cycle beyond its initial use, and therefore remain must remain available for a significant period after creation.
- Without planning and intervention digital objects naturally deteriorate or become obsolescent.
- It is the responsibility of archives and repositories to both preserve the data and make it accessible.

Scope

The scope of this policy is fourfold and it focussed on building an understand of the processes for data management with the aim of providing a basis for cooperation and collaboration between creators, users and managers of data.

- Provide a framework for understanding the processes of the management and preservation of digital data.
- Outline fundamental processes for ingesting, managing and preserving data for use by DARIAH's user community.
- Allow users and organizations to be effective partners in the preservation process.
- Expand knowledge about the necessity of long term maintenance of data for arts and humanities research and educate researchers about their roles and those of the repositories.

The long-term management of data and the associated issues of preservation must be built around a partnership between the creator or depositor (this may or may not be the same) of the data and the organization which is entrusted with its management. The earlier in the process that this association commences the greater the chances that the data will be successfully preserved for use by future researchers. Each half of this partnership brings

benefits to the association. The preservation organization will have (or at least have access to) both a broad and detailed knowledge of the state of the art with regard to how best to create data to make preservation possible. The researcher (creator/depositor) will understand the particular needs of the data and the communities of practice which provide the context. Only through the amalgamation of these two concerns will preservation be possible.

Audience

This policy is primarily aimed at partners within DARIAH who are responsible for the collection, management and preservation of data. Repositories which are part of the DARIAH will be expected to adhere to the policies set out here in order to make sure that the processes for ingest and long term management are of the highest possible quality which also ensuring at least a minimum level of uniformity across the partnership. This policy specifically applies to organizations which have accepted the responsibility to preserve digital information and data and make it available to users.

In addition, this public document will allow users of DARIAH to understand the procedures and processes which underlie the management of their data. DARIAH will work to make all of the procedures for ingesting, managing and preserving data transparent and comprehensible, to enable DARIAH users to make informed decisions about how to best provide for the long term preservation of their data. As such, users will also have responsibilities with regard to preservation, including, but not limited to, making their data available in accepted formats, and providing accurate archival and preservation metadata, using accepted media for transferring their data to the archive.

Responsibilities of a data managing/preservation organization

- Work with data producer to incorporate information into archive using agreed best practices
- Manage the data following policies and using agreed procedures to ensure long-term preservation

- Make the data available through means approved by DARIAH and provide access to the data to the community of users

Data Ingest

Pre-ingest

Engagement with data and data creators in pre-ingest stages is recommended and can greatly improve the quality of the data and the process of preservation. This is not just a matter of involvement once the data has already been created. DARIAH partners responsible for the management of data, as well as those who work directly with researchers, should aim to be active in the development of projects from their earliest stages. Pro-active activities in the planning stages of data creation projects, and throughout the data lifecycle prior the depositing of data and resources is beneficial for the researcher as well as the organization responsible for maintaining the data. It has cost-saving benefits because it ensures the quality and suitability of the data to be deposited thus requiring less processing at ingest.

In the case of new data creation projects, consortium partners should therefore aim to be engaged from before the data collection phase of the project. It is understood, however, that this is not possible in all situations. Early engagement will make possible an approach which embeds good data management practices at the most basic level, avoiding problems which may otherwise arise. Important things to consider include data formats, copyright considerations, and the embedding of appropriate metadata from the start.

Ingest responsibilities

The key responsibilities of organizations which are part of DARIAH with a responsibility for data management and preservation are as follows:

- **Data Quality:** It is the responsibility of the data manager to ensure that data is received from the creator in a format which makes long term preservation possible. Data may be obtained in from depositors and creators in a range of styles and formats, some of which are unsuitable for long term preservation. In such cases before data can be ingest it will be necessary to transform the data into accepted sustainable preservation formats.

- **Validation of data:** When data is received from the producer it is the responsibility of the organization which will manage it to verify its validity. It should be supplied in an uncorrupted form and should be a complete representation of the data without

errors. Verification that the data is both complete and uncorrupted should be done using current best practice. A clear audit trail must be maintained with sufficient documentation to provide evidence that accepted procedures have been followed. Procedures for undertaking these crucial tasks should be clearly outlined in a publicly available handbook.

- **Metadata assessment:** The data manager should gauge the level of metadata provided by the depositor and work with the creator to ensure that the metadata provided is of sufficient quality for the preservation of the data. In particular at this stage it is crucial than the necessary preservation metadata is provided or created. Preservation metadata should includes all of the necessary information to make the information accessible including: details about the format of the files, instructions for use and re-use, documentation of all actions that have been performed on an object since its creation, information about the authenticity and provenance of a digital object, and rights information.

Ingest procedures

1. **Accession:** Data to be ingested by the data manager has to be accessioned through a series of recognized steps and following agreed procedures. Data is likely to be deposited in one of a number of portable media (e.g. CD, portable hard drive, etc.) It should be checked for viruses, media corruption, data corruption and completeness. In addition the metadata supplied should be reviewed at this point. The depositor should be notified if any of problems come to light from these procedures.

2. **Documentation:** Adequate documentation procedures must be in place and all activities in ingest should be fully documented. In addition to the organization's own documentation all DARIAH documentation should be completed and checked. It is crucial that this process involve the notification of DARIAH of all newly accessioned material in order to provide complete data for registries and catalogues.

3. **Acquisition:** All data provided by the depositor on portable media must be copied to the organizations own storage. Data should be stored in its original formats in preparation for transformation into the formats in which it will be preserved. Once data is ingested into the system it needs to be checked for consistency and accuracy. This should include checking that the files conform to the attributes of the recorded formats.

4. **Rights and confidentiality:** It is vital that all rights be cleared prior to ingest of data. The depositor should be responsible for clearing rights and providing documentation. In addition the depositor should inform the repository of any issues of confidentiality related to the data and its potential use.

5. **Metadata:** Ensure that all necessary metadata is provided with the data. If some areas of metadata are inadequate, then steps should be taken at this point to ensure that it can be completed. In particular we are concerned with preservation metadata, which focuses on the information necessary to support and document long term preservation of the digital object or collection. This should include information about provenance, authenticity, preservation activity, technical environment, and rights. Some of this will be added to and developed over time as the material is preserved. It is important that all preservation metadata is kept up to date for the life of the object. Mechanisms for doing so should be built into the preservation process.

6. **Create preservation version:** This version is the primary vehicle for ensuring that the content in the deposited objects is accessible in the long-term storage, and that it is possible to migrate the data when this is deemed necessary. It should aim to make preservation as straightforward and inexpensive as possible, while retaining the essential elements of the original digital object. Ideally the preservation version should have the following characteristics :

    a. Software and hardware independence - In order to minimize the possibility of obsolescence the data should be preserved in formats which are not software or hardware dependent.

    b. Simple data structures and open file formats – This will minimise the need for future migration and maximise the possibility that the data will be accessible as the technology changes and the potential for corruption and inability to access the data sets in.

    c. Long-term preservation planning – The chosen preservation method should be based upon current best practice for the type and format of the data to be preserved, and a written plan based on those practices should be set out at this stage. If there are any peculiarities of the data that will make it difficult or complicated to preserve, this should be included in the planning.

7. **Dissemination:** In some cases it might be necessary to create a version or versions for dissemination. Often, the original version can be utilized in the short-term, but as changes in technology occur and users are no longer able to utilize the original version, then it may be necessary to create versions which preserve more than just the data, and allow the user to experience the data in ways which are identical or close to the original experience of using the digital materials deposited.

Throughout the ingest process it is crucial that procedures are followed, actions taken are documented, and all preservation metadata updated. Procedures should be in place to demonstrate that the objects are accurate representations of the original digital object.

Management

Long-term management and maintenance of the data is the second core function of a DARIAH repository. This is about maintaining and providing access to the search information about the data, custody of the digital objects and metadata, storing the data in suitable conditions, ensuring the viability of the original deposited information and migrating the data to new formats when necessary.

Preservation Strategy

All decisions about preservation strategy must be taken with consideration of the other key DARIAH policies, including policies on Data Creation, Collection Scope and Coverage. The Trusted Digital Repository policy should also guide any specific decisions about policy within a given archive or repository.

Archival Storage

Archival storage is the primary component of data management. The purpose of this is to ensure that the data deposited and the information about that data is available to users of the archive in the long-term. This process requires that the data remains identical to that which was created during the ingest process. It is also an absolute requirement that the data be accessible to the archive's users. Data integrity must be monitored in order to ensure that the content is preserved. There are three main components to this process:

1. **Data preservation and storage**: Each data management organization will have its own procedures for archival storage and DARIAH does not aim to be prescriptive in this area. However it is crucial that storage and preservation policies are coherent and transparent. Guiding principles and policies should be published and frequently reviewed.

2. **Monitoring**: It is crucial that data be checked on a regular schedule for media wear and tear and data corruption. This should be done with the aim of discovering issues before they develop into severe problems that will jeopardize the viability of the data. This allows the organization to move the data to new media to prevent future wear and tear. The organization should have policies in place regarding the methods and frequency for checking media and its associated data.

3. **Refreshing & migration**: If problems are discovered there should be a means of updating the media through recognized procedures. There should also be routine maintenance and refreshing to new media undertaken to avoid any problems. Where the data is compressed, preservation standard compression formats should be used. In addition to media migration, where necessary, format migration should be undertaken on the data. The criteria for format migration are similar to those governing the initial creation of data during ingest. Primarily data should be migrated only when it is necessary to either (1) prevent obsolescence or (2) maintain the data in formats which are accessible and useful to the user community.

Data Management

This is the second major aspect of the overall management function of a DARIAH archive. The requirement for management of the data includes such facets of archival practice as maintenance of metadata (including databases of descriptive information about the data), supporting external finding aids, and manages version and change control.

All preservation decisions should be monitored and recorded in a centrally-held database in which all changes and versions created during the management of the digital data objects. The accurate documenting of any alterations to the digital objects is crucial for ensuring the integrity of the documents and therefore the value of the object to the research community which the archive serves. Where there are questions about the impact of changes on the value of the object to researchers the archive should endeavour to consult members of its

user community about those changes. It is always necessary to keep the needs of researchers at the forefront of considerations regarding the data.

Researchers interaction with the data held by the digital archive is a crucial part of all preservation. This is the one of the areas in which the organization will interact most closely with the end user of the data. The repository must provide the means for finding information, requesting information and transmission of that information to the end user. In order to achieve this the archive must maintain information about the data it holds, and finding aids to allow for discovery. It should also provide support for external finding aids.

Safety and Security

An essential aspect of all preservation activities is physical safety and security of the data. The organization responsible for managing the data should be able to demonstrate policies and facilities commensurate with the needs of maintaining security of both software and hardware systems. Data security requires encryption and passwords in order to prevent access to files without authorization. Safety relates to protecting the data from physical harm, through is maintained through robust procedures and facilities for fire prevention and protection, intruder prevention, and environmental control systems.