Preservation Policy Data Archiving and Networked Services (DANS)

Version 1.1 – January 20, 2015

Contents

1. Outline

This policy outlines the principles which underpin the main activities of DANS (henceforth "the Archive") regarding sustainable identification and preservation of, as well as access to digital research data for use and re-use within its user communities. From a preservation point of view this policy generally conforms to the OAIS Reference Model[1], with alterations that are specific to the materials held within the Archive[2].

2. Mission of the Archive

DANS has as its mission to promote sustained access to digital research data. To ensure the continued use of these resources the Archive follows a policy of active preservation with the aim of ensuring the authenticity, reliability and logical integrity of all resources entrusted to its care while providing formats suitable for research for the long term. The designated community of the Archive consists of scholars in Humanities and Social Sciences (the reference to a classification of research disciplines can be found in Appendix A).

The policy codifies long-standing good archival practice at the Archive. In 1964 the first set of data has been archived by the Steinmetz Foundation for Social Sciences, one of the Archive's predecessors. Another predecessor is the Netherlands Historical Data Archive (NHDA), founded in 1989. Furthermore, as of 2007 agreements have been formalised for archaeologists to deposit their data in the e-depot for Dutch archaeology (EDNA), which is part of the Archive. Since 2010 the Archive is gradually extending its domain from Social Sciences and Humanities to Life Sciences.

The formulation and biannual revision of a preservation policy for the Archive are essential steps in fulfilling its strategic aims and responsibilities: it gives strategic direction both to continue initiatives which are necessary for the protection of its collections, and to meet or extend nationally and internationally agreed standards (to be) for digital preservation.

3. Scope and objectives of this policy

3.1. Scope of the policy

The scope of this policy is limited to the Archive. It deals with all aspects of preservation and applies to all materials held by the Archive. This policy does not consider preservation of other materials such as DANS's web pages, internal documents, and the Archive's intranet.

Preservation decisions have an impact on most areas of the Archive and therefore this preservation policy should be read in conjunction with the

[1] References can be found in Appendix A.
[2] This policy has been modelled on UK Data Archive's Preservation Policy (public version 18 May 2011), which is gratefully acknowledged.

DANS Strategy Policy. The preservation policy is equally steered by a variety of external guidelines and standards for digital preservation such as OAIS, Data Seal of Approval, DIN 31644, and ISO 16363.

3.2. Objectives of the policy

The Archive's primary objective is to identify, preserve and make available for use digital research data that have permanent or continuing value. The Archive assumes responsibility for the long-term preservation and accessibility of digital objects. For all practical purposes "long-term" means preservation for at least five years – the minimum retention period for raw research data, according to the Netherlands Code of Conduct for Scientific Practice (2012) – *and* beyond the next round of technical change.

The Archive is also responsible for ensuring the reliability and logical integrity of the data. Any strategy for the long-term preservation of any digital information must address the issue of software dependence. For most digital information it is generally possible to eliminate software dependence by sacrificing structure, but the end products of these transformations are not authentic versions of the original. Thus the primary goal of the Archive's preservation policy is to ensure the long-term accessibility of electronic information while ensuring the highest level of authenticity possible.

The specific aims of the preservation policy are to:

- provide authentic and reliable instances of datasets to researchers;
- maintain the integrity and quality of the datasets;
- ensure that digital resources are managed throughout their lifecycle (e.g. when migrations or changes in metadata are carried out) in the medium that is most appropriate for the task they perform;
- ensure that the relevant level of information security is applied to each dataset;
- and so to be a "trusted digital repository".

4. Requirements

4.1. The Archive's requirements

The Archive has developed a series of requirements which it strives to ensure are followed as closely as possible:

- the data that the Archive acquires are accompanied by adequate documentation to enable their use and re-use for analytical and research purposes;
- the datasets are checked and validated according to strict data ingest procedures (see Section 5);
- the data are professionally catalogued according to appropriate metadata standards;
- the data, documentation, metadata and other representation information are kept in conditions suitable for long-term archival storage;

- the authenticity, integrity and reliability of datasets preserved for future use are retained;
- the actions undertaken by the Archive are uniform regardless of the perceived value of any dataset.

4.2. Legal and regulatory framework

The legal and regulatory frameworks for the management of the data acquired by the Archive are as follows. As an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW), DANS is not a legal entity in itself. Instead, the KNAW is the legal entity under which the Archive functions.

In preserving its datasets and providing access to them the Archive follows:

- Code of conduct for use of personal data in scientific research of the Dutch Association of Universities (VSNU, 2012). This code of conduct is an elaboration of the Dutch Data Protection Act (WBP).
- Copyright act (1912);
- Database act (1999);
- OECD Principles and Guidelines for Access to Research data from Public Funding (2007);
- Privacy regulations for treating personal data from depositors, users, and other third parties.

The relationship between the depositor of a dataset and the Archive is based on a legally-binding deposit agreement and licence (known as the Licence Agreement) which

- confirms the rights and obligations of both parties;
- states the conditions under which access may be given to third parties, as specified by the depositor[3];
- states that the depositor has cleared all necessary permissions.

The Archive will not ingest materials that have unclear ownership or unresolved rights issues.

The relationship between the user of a dataset and the Archive is based on legally-binding General Condition of Use which concern

- the personal use of the data;
- the special restrictions that apply to datasets with personal data according to the Dutch Data Protection Act (WBP);
- the required bibliographic reference to the dataset.

5. Roles and responsibilities

All DANS staff assist in implementing this preservation policy as appropriate to their roles and responsibilities. The Director is responsible for maintaining this policy. Furthermore, a protocol for data processing that pays attention to data provenance is available.

All DANS staff, including temporary staff, visiting fellows and volunteers, are accountable to their line managers for compliance with this policy and

[3] DANS applies the principle 'Open if possible, protected if necessary'.

with related policies, standards and guidelines, including the "Declaration of Confidentiality DANS for employees".

6. Content coverage

Research carried out within the designated community yields a wide range of data types such as texts, spreadsheets, databases, pictures, video, audio, and geographical information. The Archive strives to accommodate this wide range of data types. For any data type various digital file formats exist.

However, all formats of digital files stand the risk of becoming obsolete in the future. The current software will not be able to represent and use the content of the file in the way it was meant to at the time of creation. Moreover, software may be dependent on hardware and the Archive is not in the position to preserve hardware. However, some precautions can be taken. One such measure is to select file formats that have a high chance of remaining usable in the far future. Therefore the Archive has assessed a number of file formats resulting in a list of preferred formats and acceptable formats.

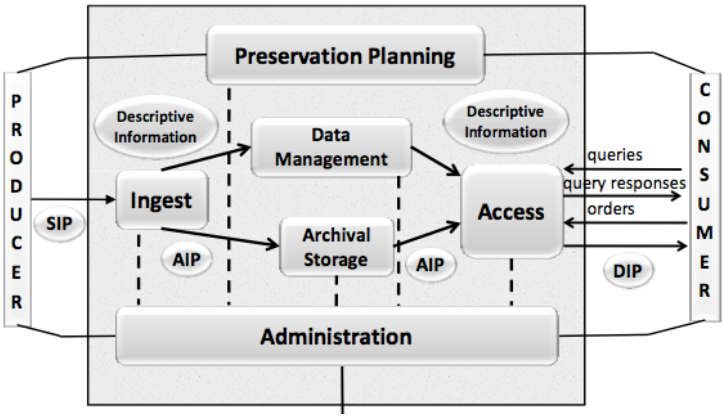
The preferred formats are the file formats which the Archive trusts to offer the best long-term guarantees for usability, accessibility and robustness. Data depositors are strongly recommended to deliver their data in the preferred format corresponding to the type of data. The Archive also allows the use of acceptable formats, but informs prospective depositors about the fact that long-term preservation of these formats is uncertain. This list of preferred formats and acceptable formats changes over time as new formats will be developed and others will fall into disuse.

The data types that the Archive intends to acquire, preserve and make available are of a static nature. When data are changed or extended, the resulting updates are considered as new datasets (see Section 7.4).

Until now the Archive does not acquire or preserve the software that has been used to generate research data, although researchers are encouraged to deposit documentation of the applied software (such as brand, version and configuration parameters) in conjunction with the data.

7. Implementing the preservation strategy

This chapter is structured around the main functional concepts of the Open Archival Information System (OAIS) reference model for digital preservation environments. The Archive's processes are organised according to the OAIS reference model:



7.1. Pre-ingest function

Officially, the pre-ingest function is not part of the OAIS model. However, the
Archive has learned from experience that pre-ingest services help to ensure the usability and accessibility of datasets through the improved quality of metadata and documentation. This way, they also reduce costs within the ingest phase.

In particular, the Archive provides data guides, training and consultancy to groups and individuals about issues such as data formats, data management plans and legal issues.

7.2. Ingest function

Ingest is the first functional component of the OAIS reference model. It includes the receipt of information from a producer and the validation that the information supplied is complete. This process also identifies the specific properties of the information which is to be preserved; it authenticates that the information is what it purports to be. The supplied version is known within the Archive as the "original" version and this is retained for preservation in its original format and stored in the appropriate directory on the preservation system. This supplied version has a close correspondence to the Submission Information Package (SIP) in OAIS parlance.

The Archive staff performs quality control by following a data processing protocol to safeguard that the supplied data will be findable, accessible and comprehensible for the long term.
The ingest function also transforms all elements of the deposited files into a valid preservation format for the specified data type. Furthermore, the ingest function includes the creation of descriptive metadata. The version resulting from the ingest process is an Archival Information Package (AIP).

Upon submitting the original version the depositor is informed that the material has been transferred to the Archive's custody. The Licence Agreement is sent to the depositor along with the unique persistent identifier minted by the Archive's system. When the Archive staff has subsequently processed the dataset and published it to the user community, it informs the depositor about this step.

The Archive will not preserve depositor-submitted media or non-digital documentation in their original format. These will either be returned or destroyed securely.

7.3. Archival storage function

In essence, the purpose of archival storage is to ensure that what is passed to it from the ingest process remains identical and accessible. In the Archive this function receives AIPs from the ingest function and adds them to the permanent storage facility, oversees the management of this storage, including media refreshment and monitoring. This function is also responsible for ensuring that AIPs can be retrieved.

Data storage management has been outsourced. The Archive has a Service Level Agreement (SLA) with its data storage management provider, which includes a confidentiality statement.

7.4. Data management function

Data Management is the third major function of the OAIS reference model. It maintains databases of descriptive metadata; supports external finding aids; and manages administrative metadata which support internal operations, including change control.

Ensuring that any alteration to the preserved version of any part of a dataset is accurately documented is integral to the authenticity of any dataset. The Archive distinguishes between two forms of alteration post ingest:

- New version and therefore a new dataset: when there is a change to data;
- Minor change: when there is a change to metadata, descriptive documents or supplementary files.

When there is a new (version of a) dataset, the Archive recreates all descriptive and structural metadata and retains the old file and the previous AIP within the preservation system. The new dataset is assigned a new persistent identifier. This way, the already existing persistent identifier will continue to refer uniquely to the earlier version of the dataset. The new and the previous dataset are cross-referenced in their respective descriptive metadata.
Alternatively, when there is a minor change, this change is documented in the administrative metadata; no new persistent identifier is minted.

In the case of data conversion to another file format for preservation or access purposes, the Archive maintains the original file(s). The conversion aims to preserve the content of the data, because this is a significant property of the data. Preservation of other aspects, such as the layout of the input format (the “look and feel”) is considered to be of lesser importance.

Deleting data would be an extreme case of data change. However, the Archive does not delete data.

7.5. Access function

This OAIS function contains the services and functions that make the archival collection and related services visible to consumers. End users interact with the Archive to find, request and receive datasets. By default these processes are web-based, but with support by the Archive staff.

Apart from the processes that support these three activities (i.e. find, request and receive datasets), the access function also implements the security that is related to access.

7.6. Administration function

In the OAIS model the administration function manages the day-to-day operations of the Archive. Processes covered here e.g. relate to the negotiation of the license agreement and the general conditions of use (see Section 4.2) and to system engineering functions to monitor the Archive's system operations (see Section 7.3).

7.7. Preservation planning function

The goal of this OAIS function is to ensure that the data in the Archive remain accessible, understandable, and sufficiently usable over the long term. The Archive's preservation strategy is based upon open and available file formats, data migration and media refreshment. Preservation decisions at the Archive are made within the context of the Archive's mission and strategy, balancing the constraints of costs, scholarly value, user accessibility, and legal admissibility.

A crucial fact is that all file formats and physical storage media will become obsolete. The Archive stimulates the use of so-called preferred formats (see Section 6) and has a monitoring process in place to decide when migration to other formats is advisable. Physical data storage has been outsourced (see Section 7.3).

8. Integrity and security

The complete chain of the Archive's custody of all datasets is documented through metadata. All actions are explicit, complete, correct and current. However, only the „original" version can be said to be an integral copy of the version deposited with the Archive. The preservation and dissemination versions are considered to be authentic and there is provenance information of all alterations in the preservation and dissemination versions that relates back to the original deposited version.

The Archive is committed to taking all necessary precautions to ensure the physical safety and security of the data it preserves. This includes a periodical technology vulnerability scan, the SLA with the data storage provider, a procedure for file fixity checking, an annual DRAMBORA Risk Assessment as well as the Declaration of Confidentiality for employees and a periodical safety inventory by the KNAW.

9. Sustainability plans and funding

To fulfil its mission the Archive receives structural lump sum financing from both the KNAW and Netherlands Organisation for Scientific Research (NWO). Should a situation arise which threatens the continued existence of the Archive, these organisations are committed to taking responsibility for the future availability of the data entrusted to the Archive.

Institutional depositors, as opposed to individual researchers, constitute another source of funding, as does participation in (international) research projects. This follows from goals in the DANS Strategy Policy.

10. Appendix A: References

DANS's documents on policy and strategy

General Policy Framework (in Dutch):
http://www.dans.knaw.nl/nl/over/organisatie-beleid/informatiemateriaal

Privacy policy:
http://www.dans.knaw.nl/en/about/organisation-and-policy/legal-information

Preferred data formats:
http://www.dans.knaw.nl/en/deposit/information-about-depositing-data

Provenance and data processing:
http://www.dans.knaw.nl/en/deposit/information-about-depositing-data

Strategy policy:
http://www.dans.knaw.nl/en/about/organisation-and-policy/information-material

The confidentiality document is internal but available on request. The safety analysis initiated by the Royal Dutch Academy of Sciences (KNAW) in 2013 is internal.

Other references

Data Seal of Approval: http://datasealofapproval.org/en/

DIN 31644 - Information and documentation - Criteria for trustworthy digital archives: http://www.nabd.din.de/cmd?level=tpl-art-detailansicht&committeeid=54738855&artid=147058907&bcrumblevel=3&languageid=en

DRAMBORA - Digital Repository Audit Method Based On Risk Assessment:
http://www.repositoryaudit.eu/about/

ISO 16363:2012 - Space data and information transfer systems - Audit and certification of trustworthy digital repositories:
http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510

NARCIS - Classification codes of the research information portal NARCIS:
http://www.narcis.nl/classification/Language/en

OAIS - Reference model for an open archival information system (OAIS):
http://public.ccsds.org/publications/archive/650x0m2.pdf

Steinmetz Foundation (history in Dutch):
http://www.edata.nl/0403_011209/Het_Steinmetzarchief_geboren_uit_een_hausse_aan_veldonderzoek.pdf

UK Data Archive – Preservation policy:
http://www.data-archive.ac.uk/curate/preservation-policy

VSNU - Netherlands Code of Conduct for Scientific Practice:
http://www.vsnu.nl/files/documenten/Domeinen/Onderzoek/The_Netherlands_Code_of_Conduct_for_Scientific_Practice_2012.pdf