netarkivet.dk

Strategy for long term preservation of material collected for the Netarchive by the Royal Library and the State and University Library 2014

Introduction

This document presents a strategy for long term preservation of the material collected for the Netarchive by the Royal Library and the State and University Library.

This document is updated and approved by the boards of directors at the Royal Library and the State and University Library every third year in connection with the updating of the *Policy for long term preservation of digital material collected for the Netarchive by the State and University Library and the Royal Library.*
This document is public and is published on the website of the Netarchive (www.netarkivet.dk).

This document begins with a description of the purpose of the preservation strategy. The next sections describe the framework for, and the requirements of the preservation activities.

Purpose

The preservation strategy covers a series of strategic responses to the issues raised in *Policy for long term preservation of digital material collected for the Netarchive by the State and University Library and the Royal Library.*

The overall purpose of the preservation strategy is to explain the principles and priorities for the long term preservation of the Netarchive's materials, in order to establish a basis for decision making in planning preservation activities.

- *Collection* of material for the Netarchive is not treated in this document, for this topic, see the Netarchive's *Policy for collection material for the Netarchive.*

Access to material in the Netarchive is not treated in this document, for this topic, see the Netarchive's *Policy for access to material in the Netarchive* (in preparation).

Framework for preservation

Responsibility and roles

The management of the Netarchive consists of a steering committee with a representative from each of the three professional groups at the two libraries as well as the operative manager. The steering committee refers to the directors at the two respective libraries.

The common responsibility for operation of the Netarchive is covered by a cooperation agreement which is revised every three years.

Knowledge sharing and development of competence

The Netarchive works to ensure that its staff always have up-to-date professional knowledge. This is done in the following ways:

- Participation in international fora: staff at Netarchive participate in international projects, conferences and cooperation, for example, the workshops held by the International Internet Preservation Consortium (IIPC), and also relevant international conferences intended to develop and share knowledge in this area.

Support of research in digital preservation: Netarchive also works to ensure that the knowledge built up about digital preservation is disseminated to professional colleagues and the public through contributions to the website *digitalbevaring.dk*

Trustworthy Digital Repository

Until 2020 the Netarchive is to be a *Trustworthy Digital Repository*, which complies with the ISO standard 16363:2012 - Space data and information transfer systems -- Audit and certification of trustworthy digital repositories. This is to be attained through close cooperation with a partner organization that works in the same area, so that the Netarchive and its partner can provide mutual assistance to ensure the institutions' status as trustworthy repository.

Standards

The Netarchive uses insofar as possible international standards in carrying out and evaluating preservation-related activities. The Netarchive continues the implementation of the following standards:

- ISO 14721: *Reference Model for an Open Archival Information System* (OAIS) as a reference model in connection with the description and construction of preservation solutions for the archive.

- ISO 14873: *Information and documentation – Statistics and Quality Indicators for Web Archiving* in connection with the description and statistics regarding the content of the archive. Statistics are important in connection with estimations of the cost of preservation and in the choice of functional preservation strategy.

- ISO 28500: *Information and documentation -- WARC File Format.* In this connection decisions must be made with regard to whether the collections of the Netarchive in ARC-format should be migrated to WARC.

- ISO 16363:2012 - Space data and information transfer systems -- Audit and certification of trustworthy digital repositories with regard to achieving the status of a trustworthy digital archive.

- DS/ISO/IEC 27 001 – Information technology - Security techniques – Management systems for information security – Requirements related to information security.

The Netarchive attempts as far as possible to develop and use open, non-proprietary software (open source):

- The Netarchive uses and actively contributes to the development of NetArchive Suite

- The Netarchive contributes through IIPC to the development of relevant tools for all aspects of the task of web archiving, including characterization and validation of WARC, which is part of the ongoing archiving activity.

Data formats
The content of the Netarchive reflects the data existing on the internet at the point at which collection takes place. This means that all forms of data types, formats and versions are collected.

The Netarchive attempts to comply with the following general guidelines:
- Data is stored in original formats.

- The Netarchive attempts to preserve its collections of data in original formats in archiving formats

- The Netarchive follows as far as possible the international standards in this area.

- The starting point is to store data uncompressed. Where for practical or economic reasons the decision is made to use compression of data, lossless compression algorithms are used. Data is stored unencrypted out of consideration for preservation security. In the few cases where access security outweighs preservation security, the issue of encrypting is deliberated. In data transmission of confidential collection data encryption is used if possible.

Legislation
The Netarchive performs bit preservation of data based on the Danish Act on Copyright, § 16 (LBK nr. 202 of 27/2/2010), which permits copying (production of copies) of data for the purpose of preservation.

Costs
The Netarchive continually works to clarify the costs of digital preservation in order to ensure the resources necessary to the task.
If there is a discrepancy between the activities the Netarchive must carry out and what the economic framework allows, the Netarchive informs the boards of the libraries of this by written notice.

Risk assessment
The Netarchive is subject to the risk assessment which the two legal deposit institutions exercise on their other digital collections.
In this connection, Netarchive will carry out risk assessment for the whole operational installation. The risk assessments are presented in writing to the boards of the libraries.

Bit preservation
The most basic form of preservation of digital material is *bit preservation*, a method which strives to secure the originally collected data from destruction. The Netarchive performs active bit preservation on the whole archive. To meet the given political guidelines as well as internationally

recognized *best practice* with regard to data redundancy, the Netarchive's data is stored in the following ways:

- Data management: 2 replicas and 1 copy in the form of a backup
    - The two replicas are *read-only* archives, from which material cannot be deleted. A consequence of this is that collected documents which contain virus or other malware cannot be deleted.
    - The Netarchive does not do either virus or malware checks on the collected material. The two replicas are online copies in independent environments at the Royal Library and the State and University Library respectively. These environments have different hardware and software configurations and are geographically and organizationally separated by a distance of 300 km.
    - The backup copy is on magnetic tape at the State and University Library.
- Integrity check:
    - Continual integrity checks are performed to ensure the validity of the data in the two replicas. The integrity check is based on the comparison of check sums on all the data from the two replicas.

The Netarchive's bit preservation is automatically monitored, currently by the bit preservation software included in NetarchiveSuite.

All the data in the Netarchive is bit preserved by the National Bit Repository software according to the current regulations and the number of copies which are common practice at the two libraries. Regular integrity checks of the collection must be continued. The frequency is set in the Netarchive's annual plan.

Functional preservation

Functional preservation covers a series of different models for the permanent security of *access* to the content of the archive over time.

The Netarchive has not adopted one single model, but uses two primary types, which can also be combined:

- Migration. By migration is understood a preservation and access strategy which ensures permanent access to the content of the archive by continually transforming it from the original to current data formats.
- Emulation/virtualization. This approach has its starting point in the originally collected data and strives to make them accessible through technically re-creating the platforms and program dependencies necessary to run or display the original data.

    The emulation strategy requires a secondary collection of legacy software, operating systems, etc., which can assure running the collected data on the emulated platforms.

In relation to functional preservation of web material the Netarchive attempts to ensure that the following strategic goals are fulfilled:

- Scalability. All aspects of the operational installations of Netarchive must be scalable in order that the preservation activities can be performed in practice.

- Deduplication. The Netarchive deduplicates certain data. This is primarily done out of consideration for the cost of digital storage space. Deduplication must not prevent the "gathering" of a website for dissemination, even though the material may have been collected in different harvestings.

- Data retrieval. It must be possible to retrieve, process and preserve a subset of the archive and give access to it for external users.

- Data mining. It must be possible to make the Netarchive's material available for data mining in part or as a whole.

The Netarchive must ensure the collection of secondary software for use in emulation, possibly through international cooperation.

Metadata and documentation

The Netarchive indexes on the following parameters:

  - URL
  - Time of collection
  - Content (URL browsing and free text index)

The Netarchive documents archive material on a series of different levels:

- Curator-created documentation of the construction of the archive, including documentation of the collection practice (event and special harvesting etc.), registration of important curator decisions regarding inclusion and exclusion of material.

- Automatically generated technical documentation. This documentation consists of generation and storage of log files from the most important collection tools, e.g. Heritrix log files, NAS log files etc. This documentation is stored together with the collected data. Documentation is accessible for all of the Netarchive staff on a Wiki platform or through the programs used to collect material.

The Netarchive's manual and automatic documentation processes must be continually revised in relation to the archive's own needs and those of its users, as well as in relation to international developments, recommendations and standards in this area.

A preservation plan must be made for those parts of the documentation which are not automatically stored in the WARC files.

In the limited cases where the Netarchive uses persistent identifiers (PID) for parts of the collection, the archive must continually maintain these persistent references.

Quality control and Preservation Planning
The Netarchive's curators are responsible for carrying out continual quality control on the material which is collected for the archive.

The collection strategy of the Netarchive is described in *Strategy for collection of material for the NetArchive by the State and University Library and the Royal Library.* The procedures for quality control of the three types of collection are different due to the considerable differences in quantity, e.g. between the selective harvests and the cross-section harvests.

Data in the NetArchive is collected and stored at the URL level in the originally published data formats. Data is thus stored so that the possibility of following links in the collected material is preserved.

The following parameters are documented systematically for the material collected through the selective harvests and *ad hoc* for the large quantities of data in the cross-section archiving.

- Functionality of the individual web page: is it possible to see/run/access text,
- Functionality of the archive's linking of data: is it possible to move from web page to web page in the archive by following links etc. on the collected web pages.
- Functionality in relation to the *type* of website: Static, dynamic, social etc. The goal is to ensure that there are not whole categories of websites that no longer can be accessed.

Selection of material, frequency of collection, collection procedures as well as documentation and quality control of these are described in *Policy* and *Strategy for the collection of material for the Netarchive by the State and University Library and the Royal Library*.

Technology monitoring and preservation plans
The Netarchive will perform technology monitoring by studying international watch reports and participating in conferences and IIPC activities in this area.
The Netarchive must decide how it will work on preservation planning in the future.

Resource persons
The Netarchive carries out its preservation efforts in a running dialogue with the other staff concerned with digital preservation at the two legal deposit institutions. In this way Netarchive ensures that the archive itself doesn't stand alone but is subject to decisions and current activities at the institutional level in both institutions.

International cooperation
The Netarchive will continue to assume a central role in the *International Internet Preservation Consortium* (IIPC) and *Open Planets Foundation* (OPF).
The Netarchive strives to participate in development and research projects of international relevance as well as maintaining a presence at conferences, seminars and the like.

Literature
http://www.digitalpreservation.gov/formats/content/webarch_quality.shtml
http://www.dpconline.org/advice/technology-watch-reports links from here to
Technology Watch Report 13-01: Web-Archiving [874KB] by Maureen Pennock