netarkivet.dk

Policy for long term preservation of material collected for the Netarchive (Netarkivet) by the Royal Library and the State and University Library 2014

Introduction

This document states the preservation policy for material collected for the Netarchive by the Royal Library and the State and University Library.

The Royal Library and the State and University Library each have their own digital collections which are covered by the respective policies and strategies for preservation of digital collection materials at the two institutions. However, they share responsibility for the materials collected for the Netarchive, a common collection of internet materials that is collected in accordance with the "Act on Legal Deposit of Published Material", [translation of] Act No. 1439 of 22. December 2004 (Legal Deposit Law), chapter 3. The Netarchive was established in 2005, and is based on the Legal Deposit law of 2004. It is managed as a virtual center. The Netarchive is thus not a legal entity, but is based on a permanent, close cooperation between the two institutions.

This document starts with a description of the goals of the preservation policy and the activities it covers. Next it explains the background for the development of this policy, including the content of the archive. A series of political implementation principles follow, and the document concludes with a brief literature list and an appendix that describes some of the specific library and technical challenges that form the basis for the development of this policy.

This document is updated and approved by the directors of the Royal Library (KB) and the State and University Library (SB) every three years.

The responsibility for revisions rests with the person who is responsible for Netarchive information at each institution.

This document is public and is published on the website of the Netarchive (www.netarkivet.dk) in a Danish and an English version in order that it be included in the general international evaluation of policies in this special area.

Goals and Limits

The preservation policy constitutes the overall framework for the work of the Royal Library and the State and University Library in preserving the collected material for posterity.

- *Collection* of material for the Netarchive is not treated in this document. For this, reference is made to the Netarchive's *Policy for collection and discarding of material for the Netarchive [in preparation]*.

- *Access* to material in the Netarchive is not treated thoroughly in this document. For this, reference is made to the Netarchive's *Policy for access to material in the Netarchive* [in preparation].

This document belongs to a group of documents that, in addition to those mentioned above, also includes

- *Strategy for long term preservation of material in the Netarchive,* in which this preservation policy is specified in concrete, implementable strategic elements.

Background
In relation to the traditional spheres of activity of preservation institutions, the preservation of digital material in general, and internet material in particular, represents a new field of work. The character of the material means that it is necessary to initiate an active preservation effort at an early stage.

In order to fulfill the goal of secure long term preservation of internet material, it is necessary that a series of policy considerations be satisfied. These issues are treated in this document.

An overview of the basic challenges in this task is briefly presented in an appendix; challenges to which this policy attempts to respond appropriately. The presentation of the challenges provides a brief programmatic statement of the necessity of an active preservation effort and a current point of departure for the continually updated preservation policy.

The content of the Netarchive
The material and the collections which are covered by this policy are:
Material harvested as a part of the Netarchive's three-pronged collection stategy:

- Selective collection of selected sites
- Event harvesting
- Bulk harvesting of all websites in the .dk domain, as well as domains and URLs outside the .dk domain that target a Danish public

- Material collected from content portals, where material is provided via other interfaces than web interfaces, e.g. ftp
- Material acquired by the Netarchive in the form of larger, closed collections of web documents, collected by automated processes through others' initiatives
- Documentation material related to collection, e.g. lists of URLs where collection has been attempted and logs of these collection processes.

Principles of implementation

Responsibility and roles

The Netarchive is run via interaction between IT-technical operations, IT-development and web curators. Both the State and University Library and the Royal Library contribute to each of these three roles.

Knowledge sharing and competence development

Both the Royal Library and the State and University Library ensure that the Netarchive staff continually have sufficiently updated expertise to maintain and carry out the preservation policy responsibly, whether through outsourcing or in-house.

The aim is to accomplish this development and maintenance of competencies as far as possible through internal education of staff, participation in national and international conferences, cooperation and partnerships, as well as through relevant research projects.

Trustworthy Digital Repository

The State and University Library and the Royal Library want the Netarchive to achieve the status of Trustworthy Digital Repository[1] and thus live up to internationally recognized standards. This ensures that digital preservation is carried out with integrity and authenticity and with the correct and necessary metadata, that the library follows relevant legislation and relevant contracts, and that the Netarchive continually updates its policy and strategy for digital preservation.

This status as Trustworthy Digital Repository must be revised regularly and the process of revision must be as far as possible supported by internationally recognized tools.

Use of standards and open software

An essential condition for obtaining the status of Trustworthy Digital Repository is the use and implementation of international standards in this area.

As far as possible the Netarchive uses international standards in the execution and evaluation of tasks related to preservation.
To ensure transparency and mutual exchange of experience, the Netarchive endeavors as far as possible to use open, non-proprietary software (open source).

Data formats

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

Legislation

The preservation activities for the material in the Netarchive follow the legislation and guidelines which the libraries are subject to at any time. The libraries each continually keep track of relevant legislation and guidelines and cooperate on a coordinated, common administration of these in relation to the Netarchive.

Costs

[1] http://digitalbevaring.dk/ord/trustworthy-digital-repository/

The preservation activities for the material in the Netarchive are subject to the existing economic and budgetary framework.

Risk management
The preservation activities for material in the Netarchive are based on principles regarding risk management with particular focus on IT-security.

- The Netarchive must carry out and/or update its risk analysis for digital preservation every third year in relation to current legislation and international standards.

Bit preservation
The preservation activities for material in the Netarchive are based on *bit preservation*, a basic technique to ensure the preservation of digital material[2].

- The Netarchive peforms bit preservation of the whole collection and its documentation material.

- The Netarchive attempts, as far as possible, to store copies of data in independent environments (geographic, technical and organisational), insofar as economy permits.

Functional preservation
The preservation activities for material in the Netarchive serve to ensure continued access to the material. For this purpose the Netarchive intends to preserve the digital collections' functional characteristics, i.e. to preserve as far as possible the original expression, as experienced by the user at the point of collection.

- The Netarchive follows the state of the art according to the most appropriate methods to ensure functional preservation.

Metadata
The Netarchive collects and preserves metadata about the digital collections.
- Metadata is stored as far as possible together with the data, so that metadata and data are closely related.

The Netarchive documents all processes.

Quality control

The Netarchive continually ensures the quality of the collected material.

- Random manual quality control is carried out on the selective harvested material

- In cross section archiving, manual quality control is performed in the form of sampling, until automated processes are available.

[2] http://digitalbevaring.dk/bitbevaring/

Cooperation with other institutions
The Netarchive works to achieve appropriate coordination of the national efforts to preserve the digital cultural heritage material among the Danish state preservation institutions.

- Division of tasks and responsibilities, primarily with regard to choice of strategy for dataformats

- Sharing of operations

International cooperation: The Netarchive must through strategic cooperative projects achieve a central position in international networks that deal with digital preservation, so that the institution is at the same level as comparable European national libraries legally, politically, and in terms of knowledge, methods and tools.

Literature:

Beagrie, Neil et al (2008): *Digital Preservation Policies Study, 30 October 2008,* http://www.jisc.ac.uk/publications/publications/jiscpolicyfinalreport.aspx
Consultative Committee for Space Data Systems (CCSDS) (2012), *Reference Model for an Open Archival Information System (OAIS), Magenta Book, Recommended Practice, CCSDS 650.0-M-2, 2012 (ISO14721:2003*), http://public.ccsds.org/publications/archive/650x0m2.pdf
Det Kongelige Bibliotek (2012): *Politik for langtidsbevaring af digitalt materiale på Det Kongelige Bibliotek*
http://www.kb.dk/export/sites/kb_dk/da/kb/downloadfiler/BevaringPolitikDigitaltMateriale_21092012.pdf
Digitalbevaring.dk (2012): *Hvordan laver man en bevaringspolitik og -strategi?*
http://digitalbevaring.dk/bevaringspolitik-strategi/
Statsbiblioteket (2012): *Statsbibliotekets Politik for digital bevaring*
http://www.statsbiblioteket.dk/om-statsbiblioteket/filer/politik_digital_bevaring

Appendix:

Challenges related to Methods

Collecting of internet material for the Netarchive is carried out via ”net harvesting”, i.e. automated searches and downloading processes. This method of collecting is established in the *Bekendtgørelse om pligtaflevering af offentliggjort materiale [Regulation regarding legal deposit of published material], BEK nr 636 af 13/06/2005* (Pligtafleveringsbekendtgørelsen [Legal Deposit Regulation]) § 2, stk. 3. The method serves, among other things, to ensure that the large quantities of relevant material can be collected and therefore to a great extent avoids manual processes and curating. Collection of internet material is described in more detail in *Policy for collection and discarding of material in the Netarchive* and *Strategy for collection and discarding of material in the Netarchive.* The automated collection results in a series of challenges with regard to the long term preservation of the materials:

- Lack of detailed knowledge of the contents of the archive.
    - Material published on the Internet has great diversity and is only rarely accompanied by descriptive metadata, information which in ”traditional” library and archival operations contributes significantly to identification and recognition of complex materials.
    - The large quantities of data make manual curating and cataloguing impossible *in toto*. The current collection policy and strategy does ensure, however, that a minor portion of the material in the Netarchive is the product of an active, manually curated collection process.
- Undesirable, superfluous content.
    - The automatic identification and download of internet material results in the collection and storage of a quantity of ”noise”, materials which are not desirable in the Netarchive, neither from the perspective of space or that of the cultural heritage. Among other things, there are the so-called *crawlertraps*, infinite loops of references that force automatic download programs into a loop that collects large amounts of superfluous data.
- Incomplete data
    - The Netarchive is not a complete collection of Danish internet material. There are deficiencies in the form of whole web sites, which are not collected, and in the form of only partially collected web sites. There are several reasons for this; among the primary ones are:
        - Lack of awareness of web sites with Danish content: the Netarchive receives information about *all* .dk-domains; but all Danish content on other top-level domains needs to be identified and added manually to the harvesting definitions. This is the case for such as the top level domains .com, .org og .eu. At this time there is no automated work procedure for this task.
        - Lack of curating: Most of the content in the Netarchive has been collected through periodic automated cross section archiving. In this method, such great

amounts of data are collected from so many web sites, that a manual process of curating is in practice impossible. Sampling is done for quality control; but in practice it has turned out that there is no way to avoid either deficiencies in collecting or redundance.

- The technical complexity of the Internet: the automated tools employed in net archiving can only with difficulty – and often not at all – deal with embedded and /or dynamic data on web pages. This means that these data most often are not collected. For example, embedded interactive elements, embedded audio and video or, material behind ”login”- or paywalls. Many of these materials can be collected through a manual effort; but in practice it is not possible to ensure that all web sites are examined manually.

- Deduplication

  - To reduce the expense of storage, the Netarchive uses *deduplication.*Through this process certain duplicate data is eliminated in the Netarchive, and replaced by a reference to the first instance of the data. Deduplication is also used for PDF-files, images and video. Html and text files are *not* deduplicated. On an average cross section harvesting (autumn 2013) Netarchive ”saves” close to 9 TB storage space in this way, or to put it another way, nearly 30 % of the total amount of data collected in one cross section harvesting.

    Deduplication presents a series of challenges for dessemination, since data from a potentially large number of harvestings must be connected to display a web page correctly.
    Deduplication also influences the statistics of the Netarchive, since the statistics are solely based on the number of files present, and does *not* take into account the deduplication references.

Above and beyond the challenges mentioned above, which to a great degree arise from the collecting conditions under which Netarchive was set up, another series of issues relate more directly to the character of the material already collected and stored in the archive:

- Despite the lack of detailed knowledge of the overall content in the Netarchive, analysis of even minor samples shows that a great deal of the content is indeed highly complex:

  - On a higher level a single archived home page can be made up of a larger complex of material brought together from various data sources and displayed parallel as a whole for the internet user. The identification and thereafter re-assembling of these individual parts can be a very difficult operation.

  - It is possible to publish all digital material types via the internet, either as ”living”, embedded data or as data packs for download, and therefore the Netarchive contains both music numbers, videos, computer games, apps, database content and computer programs. Each one of these types of material offers separate challenges to long term preservation, which must be accommodated if the content of the Netarchive must be capable of display for users in the future. The preservation institutions already have preservation policies for a series of the types of materials mentioned, but the placement

and the coupling of these with other materials in the Netarchive increases the need for an active preservation policy effort.

- The size of the Netarchive

    - Despite the short life of the Netarchive, it already represents a collection of considerable size. The amount of content published on the internet, which is collected by the Netarchive, grows year by year. The size of the volume of raw information found in the Netarchive makes the archive one of the Royal Library's and the State and University Library's largest collections. The size of the archive combined with the rapid growth creates a significant challenge for the preservation of the archive; not just in relation to the data, but also in relation to the necessity of ongoing maintenance and replacement of equipment.

- Embargoed archive content

    - The Netarchive is an "unfiltered" archive in the sense that the automated harvesting, the large amounts of data and the great diversity of the material results in an archive that contains both personally sensitive and directly illegal material. The special data protection considerations which apply to such material, result in a series of challenges in relation to long term preservation, including the possibility of limiting access to confidential, personally sensitive and/or illegal material.