1640

THE NATIONAL LIBRARY OF FINLAND

Web Archiving in Finland

Memorandum for the members of the CDNL

13 December 2010

Esa-Pekka Keskitalo

Table of Contents

1 Summary 3
2 The National Library of Finland 4
3 Background of Web Archiving 4
4 Legal Framework 5
5 Strategies and Policies 8
6 Organization and Resources 9
7 Web Harvest in Practice 10
8 Process of a Harvest 12
9 Contents of the Web Archive 13
10 Interpretation of the Legislation 14
11 Sensitive Data 15
12 Using the Web Archive 16
13 Preservation 17
14 Future Directions 18

Contact Information

Esa-Pekka Keskitalo (Mr.)

Email: esa-pekka.keskitalo@helsinki.fi

Tel. +358 9 191 44487

Address: Kansalliskirjasto
PL26
00014 HELSINGIN YLIOPISTO
FINLAND

1 Summary

This report describes web archiving in the National Library of Finland.

The National Library of Finland has been archiving Finnish web on a regular basis since 2006. Web archiving is an important part of the Library's endeavours to collect and preserve Finnish published cultural heritage.

In 2010, the amount of harvested data was 200 million files, or 25 Terabytes.

The report takes the reader through

- the relevant legislation;
- internal plans and policies;
- funding and their allocation;
- the practices of web archiving;
- arrangements for the use of the archive; and
- issues rising from data security, sensitive materials, &c.

The *Cultural Materials Depositing and Preservation Act* regulates legal deposit and web harvesting. The Library shall collect a representative sample of different kinds of web materials available, over time. No content type or genre of materials is given preference. Archiving shall cover not only freely accessible materials but also those that require registration or purchase. Not included are materials that are not intended to the general public but to a specific community – e.g. intranets and extranets.

Also the *Copyright Act* is important, giving the Library necessary privilege in making copies of web materials, and in setting the conditions of using the web archive.

The Library aims to preserve the look-and-feel of the materials archived. In other aspects, too, care is taken about securing the long-term preservation of it. The *National Digital Library Project* will in the future provide a robust system of storage, risk analysis and preservation tools.

The National Library considers the annual general harvest of freely available web resources the backbone of web archiving. It creates a relatively large and representative basis for the collection. The harvest covers all *.fi domains, and all servers in Finland the Library is able to identify. The annual harvest must be supplemented by theme-based harvests, and other means of acquiring contents.

2 The National Library of Finland

The National Library of Finland[1] is responsible for preservation of published Finnish cultural heritage.

The National Library is established in the Universities Act (558/2009)[2]. It is responsible for storing, managing and making accessible national cultural heritage within its scope of interest (that is, cultural heritage in form of publications). Further, the National Library shall provide services to academic and public libraries, and promote national and international cooperation. The Library is in charge of legal deposit

The annual budget of the Library is 26 million Euros, and it has about 280 employees.

The Library is an independent institute within the University of Helsinki[3]. The National Library is the major research library in many fields, such as history, art history, history of science, philosophy, and Russian studies. There is a separate Helsinki University Library that provides library services to the wider university community.

The Library counts its foundation from 1640 when the Royal Academy of Turku – with its library – was founded. The Academy was later transferred to Helsinki and became what is the present-day University.

3 Background of Web Archiving

From 1707 on, an obligation to deposit printed matter has existed in Finland. The original regulation under Swedish rule had its followers in the laws of the autonomous Grand Duchy of Finland as a part of Russia (1828), and finally in the laws of independent Finland (1919). In 1980, legal deposit was expanded to cover audiovisual publications. It has always been the National Library (under various names) that is responsible for legal deposit.

The need to revise the provisions of the Legal Deposit Act (420/1980), because of the emerging electronic publishing, was recognised early in the 1990s. In 1997, the Ministry of Education created a working group that reported in 1998. Another report was produced in 2000[4].

The revision took a long time to get through. There were two reasons for this. The European Union started a process of harmonizing member states' copyright legislation. This process took precedence, especially because changes in the Copyright Act would have been necessary for the archiving of electronic resources, anyway. Another factor was the decision to amalgamate legislation concerning the National Audiovisual Archive to the same piece of legislation; it is responsible for preservation of films and broadcast materials.

[1] National Library of Finland. - <http://www.nationallibrary.fi/>

[2] Universities Act. Unofficial translation. - <http://www.finlex.fi/fi/laki/kaannokset/2009/en20090558.pdf>

[3] University of Helsinki. - <http://www.helsinki.fi/university/>

[4] Committee Report / Committee preparing the Legal Deposit Act 2000. - Abstract in English.
- <http://www.minedu.fi/export/sites/default/OPM/Julkaisut/2000/liitteet/tr13_2000.pdf>

In the meantime, the National Library was able to do research and make preparations so that web harvesting could start as soon as the law would allow it. The Ministry of Culture and Education made necessary funding available.

The new Copyright Act came into force on January 1, 2006. It may be called the birthday of the Finnish Web Archive, as the Act made it possible for the National Library to start archiving web pages.

What the Copyright Act made possible was made an obligation by the Cultural Materials Depositing and Preservation Act (1433/2007) that replaced the Legal Deposit Act of 1980. It came into force on January 1, 2008. We will call it briefly the Cultural Materials Act.

4 Legal Framework

The Copyright Act and the Cultural Materials Act set up the system of electronic legal deposit, of which web archiving is a part. Personal Data Act also has an effect on the entire web archive. Some other pieces of legislation may be relevant concerning the subject matter of special items in the web archive.

4.1 Copyright Act

The Finnish Copyright Act was thoroughly revised in 2006. The Section 16 b described below was added then, too.

4.1.1 Right to Make Copies

For web archiving, the crucial provision is that the National Library is granted permission to "make copies of works that have been made available to the public in data networks" (Section 16b).

4.1.2 Right to Make the Materials Available

The Copyright Act further regulates the way these copies may be used (also Section 16b). There are three kinds of restrictions: where the materials are available; what the user may do with the materials; and for which purposes the materials may be studied.

There are eight organizations, indicated in the Act, where the copies may be used. They are

- the National Library;
- the National Audiovisual Archive;
- the Library of the Parliament; and
- libraries where deposited printed matter is also being stored.

The Act decrees that use must take place on "workstations set aside for the purpose of transmitting to the general public". Making digital copies and retransmitting the materials must be prevented.

"Research" and "private study" are the legitimate purposes of using the materials.

Very similar provisions apply to the radio and television archive, maintained by the National Audiovisual Archive. The archive and the National Library's materials may be available on the same workstations.

4.2 Cultural Materials Depositing and Preservation Act

4.2.1 Purpose of the Act

The purpose of the Cultural Materials Act is to preserve for future generations, and make available to scholars and others who need them, the materials of national culture made available to the general public in Finland. In addition, the Act also regulates the work of the National Audiovisual Archive in the field of films and broadcast materials.

4.2.2 Scope and Definitions

The Act applies to "web materials" that are located on servers in Finland or that are intended to be available to the general public in Finland.

With "web materials" the Act means any kind of material available through a data network.

A "web publisher" is an entity by whose initiative and responsibility web materials are made available to the general public.

There is an exception: the Act does not apply to archival records. This limitation is based on the desire to protect privacy. It also makes clear that the National Finnish National Archives Service is responsible for the preservation of archival records.

The Finnish text does not use the term "web" but more generally "data networks". We use "web" in this report for brevity.

Practical questions concerning the scope of the Act are discussed on page 14.

4.2.3 Collecting Web Materials

According to the Act, the National Library shall, employing software, collect web materials available to the general public. In this collection the Library shall include materials from different points of time, in a representative and multi-faceted manner.

The Library may use outside help in these operations.

The Act assumes that collecting happens first and foremost "automatically", that is, without interaction with the web publisher. However, if the materials cannot be collected so, the web publisher has an obligation to assist. The publisher has two options:

- to enable collecting, or
- to deposit.

Enabling the collecting could mean e.g. adding the National Library to the list of accepted IP addresses. Depositing would mean that copies of the materials are transferred to the Library, not by fetching from the website but in some other manner – making them available on a SFTP server, by sending them on a hard drive, &c.

Although the word "deposit" might suggest otherwise, the National Library becomes the owner of all web materials collected by virtue of the Cultural Materials Act.

The legislator has felt that the obligation to assist needs to be limited in a couple of ways. The obligation does not apply, if

- the factual, visual or sound contents of the web materials are particularly insignificant;

- if the materials are or are contained within a newsgroup, discussion forum, or suchlike[5];
- if enabling or depositing is technically impossible; or
- If enabling or depositing would constitute an undue burden because of the large size of the set of materials in question.

4.2.4 Quality and Authenticity of Web Materials

The National Library shall store the web materials so that their authenticity can be verified. The date of collection and the original location must be stored.

When deposited, the materials must be

- complete;
- free from technical defects;
- true to what was made available to the general public; and
- Not protected by DRM or other methods.

When deposited, the material must be accompanied, in a digital form, with descriptive metadata as well as with information about copyrights pertaining to it.

4.2.5 Right to Circumvent DRM Methods

It must be possible for the Library to copy materials from one memory device to another and migrate then from one file format to another. In case such actions are prevented by a technical method, the Library has – quite exceptionally – the right to circumvent them.

4.2.6 Supervision by the Ministry of Education and Culture

The National Library shall from time to time make a plan that sets out the volume of archiving of web materials, and the practices of depositing web materials. The Ministry of Education and Culture will give the plan its approval. The plan shall take into consideration the technical tools and funding available to the Library and the needs of research and study. The plan must treat all web publishers on an equal basis.

The Library has submitted a plan once and is considering submitting a new one in 2011.

The work done and the quantitative results are reported to the Ministry annually as part of the normal reporting.

4.3 Personal Data Act

Personal Data Act (523/1999) is intended to safeguard, in the processing of personal data, the protection of private life and the other basic rights which are related to the right to privacy. "Personal data" means within the Act any information on a private individual and any information on his/her personal characteristics or personal circumstances.

Obviously, the Web contains huge amounts of personal data. In spite of being nothing like an ordered register, list, or database, the Act does apply to processing of web pages.

Processing personal data is allowed only for purposes defined in the law, for example if processing is based on the provisions of an Act or it is necessary for compliance with a task or obligation to which the controller is bound by virtue of an Act (Section 8). The Cultural Materials Act gives the National Library duties that require personal data processing, insofar that data is included in the web archive.

[5] According to the preparatory works of the Act, these kinds of materials should not be in the focus of collecting, although they are not strictly excluded from the collectable materials. It was felt that web users often do not realize that what is said fleetingly on these forums can actually get stored for a long time. Chat room contents are never collected by the Library, due to technical reasons. Some other types of discussion may well be included. The border between permanent and ephemeral web contents has blurred over the years.

4.3.1 Other Legislation

The Criminal Code may have effects on the web archive in cases including but not limited to

- incitement to ethnic hatred;
- disseminating depictions of brutal violence;)
- disseminating morally offending images of children, violent acts, or zoophilia;
- disseminating information violating privacy; and
- libel.

In these cases mere holding is not an offence, but dissemination is. Holding is a crime in case of recordings of actual sexual or morally offending activities involving children.

A work in the web archive may turn out to violate someone's copyright or related rights. These cases are covered by the Copyright Act.

5 Strategies and Policies

5.1 Collections Policy

The National library adopted its current Collections Policy in 2008. The general principles of collection planning and accruing are perseverance, scholarly depth, and interaction with the scholarly community.

The specific principles concerning web harvesting are those that are adopted in practice. These practices are described on page 10.

5.2 Preservation Policy

The National Library adopted its current Preservation Policy in 2009.

Some of the general principles relevant to web archiving are:

- The original look and feel of a publication is an important information carrier.
- Digital preservation requires attention from the moment a digital object is created.
- The Library shall ensure adequate storage facilities.

Following policies of digital preservation are adopted:

- Bit-level preservation shall be ensured by suitable storing methods, back-ups, and checksums.
- Materials and their metadata should form an independent information package. Metadata should be as rich as feasible. The policy acknowledges that metadata in the web archive cannot be as complete as in e.g. digitized collections.
- Migration is the preferred method of keeping materials usable over time. Original will be kept always.

Digital preservation and its prospects are explained on page 17.

6 Organization and Resources

6.1 Organization

Work on electronic legal deposits is shared between two functions of the Library. ("Functions" are the three main branches of the Library's organization.) The Research Library Function has the overall responsibility of legal deposits, including both the printed matter and electronic materials. This function has the leading role in policy-making. Actual web archiving and other tasks that require extensive IT input are taken care by the Library Network Services Function.

6.2 Funding

Although an institution within the University of Helsinki, the National Library of Finland is partly funded directly by the Ministry of Culture and Education.

The actual costs of electronic legal deposit are not easy to determine. However, it is safe to give an approximation of 600,000 euros per annum. Salaries and related costs make about 80 % of the total.

6.3 Human Resources

There are altogether ten employees that are funded with resources that are earmarked for the fulfilment of obligations set down by the Cultural Materials Act.

- Two IT specialists focus primarily on web archiving. One concentrates more on the actual harvesting, the other taking care of indexing and other post-processing tasks, and the user interface, search engine, etc.
- One IT specialist is responsible for infrastructure of servers, disk space, back-ups, etc. He provides these services to other tasks, too. He also takes care of the dedicated workstations and secure connections to them.
- Three IT specialists work on other aspects of collecting electronic legal deposits, and on the planning of long-term preservation of digital collections.
- One person has a supervising and planning position at the Library Network Services
- Three librarians at the Research library Function work almost entirely with electronic resources.
- One librarian works more generally on digital collections from the services' point of view.

The actual work descriptions are not always as clear cut as what appears above. The Library encourages use of employee's special skills flexibly over units and projects.

6.4 Software

The National Library uses Heritrix web crawler software for web archiving. It is an open-source project of the Internet Archive. Current version (December 2010) is 1.4.2.

Also the user interface is based on software coming from the Internet Archive, namely the Wayback Machine. From the same source, NutchWax search engine is used.

6.5 Hardware

The necessary IT infrastructure is outsourced to the University of Helsinki. The Library is indeed part of the University but, beyond the basic personal computers, the Library manages its IT infrastructure independently from the IT Services Department of the University. A lot of services are provided by CSC – IT Centre for Science Ltd that is a state-owned non-profit company. In the case of electronic cultural materials, the Library has felt a need to retain a closer control of the system that has been more easily achieved within the University and indeed within the same building. These arrangements are likely to change as the National Digital Library project proceeds (see page 17).

Current hardware situation in December 2010:

- *Disk System: EMC CX-4 240. At the moment 110 disks, i.e. 120 Terabytes (gross).*
- *Web Crawler and User Inteface Server: 2 x HP BL460c, 8 core X5550 (2.66GHz), 42GB RAM.*
- *Vmware Cluster: 4 x HP BL490c, 8 core X5570 (2,93GHz), 72GB RAM.*
- *Tape Libraries: HP MSL 6030 LTO-3 & HP MSL 4048 LTO-5*

7 Web Harvest in Practice

7.1 Annual General Web Harvest

Once a year, the National Library makes a general sweep over "the Finnish Web", covering in principle all "Finnish" servers that it has knowledge of. In this context, a Finnish server would be one

- hosting websites within Finnish namespaces;
- being located in Finland; or
- hosting contents intended for use in Finland.

7.1.1 Finnish web domain names

There are two Finnish web domains, .fi and .ax (autonomous Aland Islands).

A list of domain names in .fi namespace is provided once a year by the Finnish Communications Regulatory Authority.[6] This list, so called WHOIS database, is also preserved for future reference. It contains information about the owner of the domain name. The Aland authorities are unfortunately not able to produce such a list.

7.1.2 Servers located in Finland

The National Library has taken measures to search for servers located in Finland. Information about them is not readily available. As imperfect as the result is, it adds considerably, approximately 50 %, to the total count of servers included in the general harvest.

First, we use GeoIP database in order to find IP addresses located in Finland. Then we go through these servers and check for the openness of the standard http port. The order is randomized, because several checks in a row might cause undue load or might be interpreted by security software as malicious port scanning.

In the next step, we reject hosts that do not have a DNS name. Also hosts that have names with many numbers, certain words (e.g. "adsl"), or apparently random sequences of characters are rejected. With these restrictions we are able to leave out almost all hosts that might cause problems, i.e. private hosts that are misguidedly left open, interfaces to technical systems7, etc.

[6] http://www.ficora.fi/en/

[7] Using this methods, we have come across e.g. an unprotected web interface for management of processes in a printing press.

7.2 Theme-based harvests

Theme based harvesting is intended first and foremost to amend the list of sites to be harvested by human effort.

Such list has two important criteria: we want to find sites that

- would be otherwise excluded from the general harvest; or
- would be harvested but that are so rich and deep in content that they should be harvested more thoroughly than is the usual case.

Themes can be based on several criteria:

- Genres, as blogs. Many Finnish blogs are hosted by large international services and thus outside the usual scope of the general harvest.
- Subject matter, for example "immigration, emigration and expatriates", "photography", "antique and collectibles", "ornithology and bird-watching", "climate change", etc.
- Events such as sport events, Eurovision Song Contest, political summits, etc.
- Current events. We do not have as yet set procedures for covering sudden major news items but on the voluntary basis some such events, e.g. notorious Jokela school shooting in 2007

Sites based on themes are submitted by the National Library employees. In some cases we have used outside help and intend to use it more in the future. For example, harvesting of resources in Sami languages would not have been possible at all without help from Lapland Regional Library. Another case is the Turku, the Cultural Capital of Europe in 2011. The project organization sends links to related resources in e.g. media to be harvested by the Library. It is our intention to find more partners like these.

We find that theme-based thinking helps to concentrate efforts and enables to get assistance from outside experts of the several fields. We recognise the problem of underlying bias. Looking back to the themes they might be deemed to be concentrated on arts and social issues and lack in the fields of e.g. technology. It should be stressed, that the themes are not an end in itself but only a way to add to the completeness of the general harvest. The results of theme-based harvesting are embedded into the same web archive. We preserve documentation bout the crawls.

7.3 Frequently Harvested Sites

At its best, web harvesting can only capture a fraction of information available on the web over time. Even so, it has been felt useful to harvest some sites more often. At the moment, we harvest a number of online newspapers on a weekly basis, some even daily.

7.4 Requests to Archive

As the web archive is becoming more known, web publishers have started to request harvesting of their web sites. As yet this occurs one, two times a month and the requests can be easily taken care of. Most often the web publisher is about to revamp their web site, and wants to make sure that the old version will be preserved.

8 Process of a Harvest

8.1 The Crawl

Heritrix is given a list of so called seed URLs that will be the starting points of web crawling. It is also configured for the crawl. This means giving it rules to follow:

- How many steps it will proceed from a seed URL.
- Will it remain within the domain, or not, etc.

After the crawl is ended, the results are examined before further processing. We check that all the seed URLs have been crawled. We ensure that the amount of data is what was to be expected. We go through list of sites that do not allow harvesting and evaluate the need to include them anyway in further rounds.

Sometimes, after these checks it is necessary to redo the whole crawl.

General harvest is done without deduplication, i.e. even unaltered files are harvested again. In other occasions, deduplication is activated.

Log files will be compressed and stored elsewhere, after the data are added to a database.

Current status of a harvest process is tracked manually on an internal wiki page.

8.2 Complaints from Webmasters

Web crawler may sometimes cause problems on the sites it is going through. Over the period of five years, there has been perhaps a dozen of cases where a web publisher or system administrator has been in contact concerning the web crawler. In all cases harvesting has caused a peak in load, often related to a loop.

All complaints have been resolved amicably. No one has approached the Library with demands to cease harvesting their materials.

8.3 Indexing of URLs

First, a script writes out paths for each WARC file. These paths will be needed in indexing and later in retrieving material for use.

We use a modified version of Wayback indexer. We have chosen not to use automated indexing tools in Wayback. It has been necessary to keep some manual control over the process.

The result of the indexing is an unsorted CDX file that is appended to older CDX files. The result is then sorted anew. We keep the CDX index in three parts. This is a compromise between quickness of searching (one file) and manageability.

The public index of the web archive is updated with new data from which references to the actual contents are first stripped.

8.4 Full- text indexing

NutchWAX is used for full-text indexing. At the moment, we must take care not to handle too large segments of data at the same time. It has happened that indexing process fails after days of processing because one or another resource is overloaded.

We are examining ways to improve this process.

The index fields will be boosted for better search results. A page rank file is produced from the material and used as a feed in boosting.

8.5 Back-ups

At this point, we make back-ups to a tape library. Back-ups are not automatic but are done as is felt necessary. Back-up tapes contain both (W)ARC files and log files. We have a secure storage space for back-ups.

9 Contents of the Web Archive

The size of the archive is counted as the number of files. It is unfortunately hard to say, how this translates to "documents", as an HTML page can embed anything from zero to hundreds of files. Every small graphic element is a file. – The size of the archive in uncompressed state is also estimated in the table below.

Annual			**Accrual**		
Year	**Million Files**	**Terabytes**	**Year**	**Million Files**	**Terabytes**
2006	20	3	**2006**	20	3
2007	50	7	**2007**	70	10
2008	70	10	**2008**	140	20
2009	150	20	**2009**	290	40
2010	200	25	**2010**	490	65

Table 1. Volume of web harvesting.

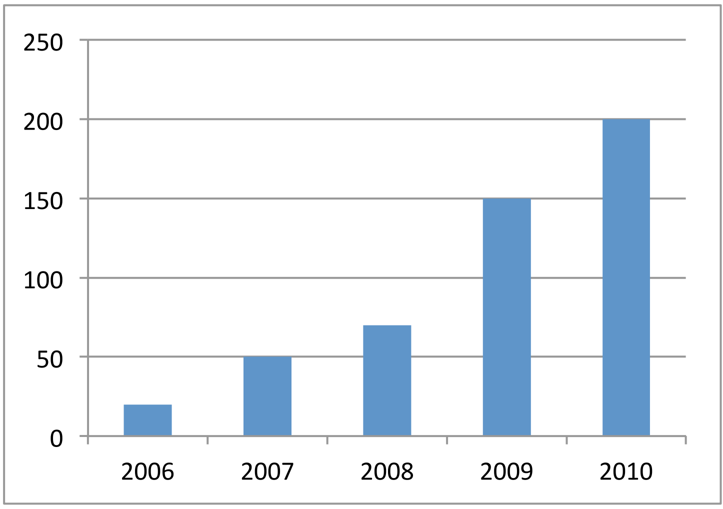

Figure 1. Volume of web harvesting.

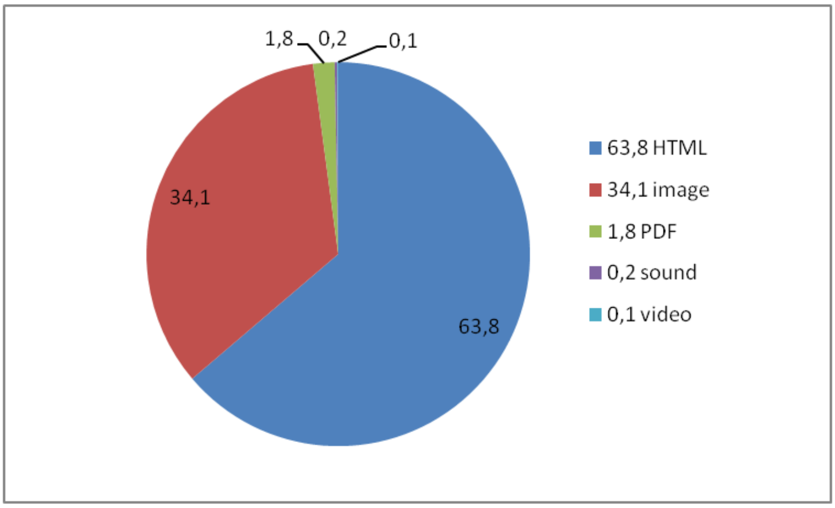

Fig. 3. Percentage of various file formats in the Web Archive

10 Interpretation of the Legislation

10.1.1 Connection with Finland

As mentioned in 2.4.2, the Act applies to materials that are on servers in Finland or that are "intended to be available to the general public in Finland".

We interpret that intention must be specific, i.e. the web publisher thinks especially the Finnish audience. Certainly most web sites do not mind being available also in Finland but obviously are not covered by the Act.

10.1.2 How to define the Public?

The Act is aimed at preservation of cultural materials that have been made available to the public. The web is used for all kinds of private purposes, too, and the definition of "the public" is therefore important.

We work on the theory that restricted access as such does not mean that material is not available to the public. Rather, the question is, whether a person needs to have specific qualifications in order to be granted access.

Materials on sale to anyone are deemed to be available to the public. Also a site that requires a registration but where everyone is able to register is deemed public.

On the other hand, even large amounts of web materials may be non-public. That is the case if the require that the user

- is an employee in a company;
- is a member of an registered association;
- is in business relationships with an company, &c.

Web materials must be effectively restricted to the intended audience, otherwise they are considered public. The fact that the address of a web site is not widely disclosed does not yet make it private.

10.1.3 Public Records on the Web

Public records are excluded from the scope of the Act. The Library will make no efforts toward collecting public records. However, it is impossible to filter such contents from large-scale automated web archiving, should they happen to be freely accessible.

The National Library feels that occasional unintentional harvesting of records is not a major problem, especially if the contents are not in any way sensitive. See also "Sensitive Data", page 15.

10.1.4 Honouring Robots Exclusion

Mandated by an Act, the National Library cannot always conform to robot exclusion rules. There are entire sites that for some reason or by mistake exclude robots altogether. In such cases the Library disregards the robots exclusions.

It cannot be a general policy, though. Many excluded areas do contain materials that are excluded for a very good reason. These include all manners of old unlinked versions of pages, unused graphic elements, etc. It is not in the interest of the Library to harvest these, and so overriding robot exclusions remains an exception.

11 Sensitive Data

11.1 Personal Data Illegitimately Published

The Data Protection Ombudsman has reviewed the practices of the Library regarding the web archive. His main concern is personal data that should not have been made available to the public in the first place. Such material should not be harvested and should be deleted from the archive.

Web publishers who have by mistake published persons' social security numbers or other sensitive data have occasionally contacted the Library. The Library has been happy to conceal such materials from the public. The Ombudsman's view is that we should rather delete them entirely. Policies will be reviewed in the near future.

The real problem is identifying such data in the first place. Given the amounts of files harvested, the Library cannot know the exact contents of the archive. Instead, the Library must trust on the awareness of those who accidentally published such material. The users of the archive should also be asked to notify library staff if they encounter materials they suspect are sensitive.

11.2 Inspection and Correction of Personal Data

Another issue is the right of a person to inspect data about her, and have it corrected.

The legislator obviously has assumed that datasets of personal data are much more orderly than the web archive. It is absolutely beyond of the Library's – or anyone's – resources and technical capabilities to exhaustively establish what, if any, data about a certain person the web archive contains.

The second step, correction of that data, goes baldly against the very purpose of the Cultural Materials Act.

This issue is still being considered.

11.3 Confidential and Secret Information

The National Library only collects materials intended for the general public. We assume that web materials that are available without qualification are not and cannot be confidential. For example, the Criminal Code defines "corporate secret" as "information that an entrepreneur keeps secret". In other words, information available on the web is not, by definition, a corporate secret.

We have not yet been in a situation where a request to *deposit* web materials would have been refused on the grounds of secrecy. The Library does receive confidential materials in printed form, and should the need arise, we assume that the procedures could be adopted to web materials, too. Such contents will be made available only after the period of concealment expires.

11.4 Web Contents that Violate Law

The National Library has recognized a number of ways web contents may act as means of offence, see page 8.

The principle of the Library is that materials that it has been illegal to publish will be concealed from the public viewing. However, it is not the Library's role to judge on these matters. We do not have the mandate, resources or expertise to determine if there is a violation or not. Consequently, the Library will take steps only based on a non-appealable judgment.

11.5 Web Contents Illegal to Hold

As mentioned on page 8, holding of recordings of actual sexual or morally offending activities involving children is a criminal offence. The National Library will delete and has once deleted such materials.

12 Using the Web Archive

12.1 Where the Web Archive is Available

The Web Archive is available to the users in the premises of eight institutions, as provided by the Copyright Act (see page 5).

These places are

(1) the National Library itself;

(2) the National Audiovisual Archive, which on the other hand may make its radio and television archive available in the rest of the institutions listed here;

(3 – 7) the libraries where additional copies of deposited printed matter are stored, i.e.
- Jyväskylä University Library,
- Turku University Library,
- Oulu University Library,
- University of Eastern Finland Library, and
- Åbo Akademi University Library; and finally

(8) the Library of Parliament that otherwise ceased to be one of the storing libraries in 2008.

Some of these libraries have branches, and it is possible to expand the network of workstations to them in the future.

12.2 Workstations

As mentioned earlier (see page 5), the Copyright Act forbids making digital copies of the digital cultural materials. For this reason, the workstations have some unusual features.

- Users only can physically have access to keyboard, display, mouse and loudspeakers and headset. The computer itself is out of reach.
- Users have no access to any data ports (USB port, etc.).
- The computers do not have Bluetooth or Wi-Fi equipment.

In the National Library, the workstations are Mac Minis that can run Mac, Windows, and Linux operating systems. They reboot automatically after a period of idleness.

It has been necessary to block access to all other websites but the ones that are intended to be used on these workstations.

Users may print materials. They may also record the sound out of speakers and take photos of the screen, as in these cases the data takes an analogue form during the transmission.

These workstations also give access to the radio and television archive of the National Audiovisual Archive. Other digital materials of restricted nature are available, too, and more will be added in the near future. These include for example more recent digitized journals that still are under copyright.

12.3 Terms of Use

The web archive is intended for research and private study. Private study is usually interpreted very liberally and covers in fact all manners of use that are not blatantly something else.

As far as it concerns the National Library, users do not need to give their personal information when using the archive. Local arrangements may vary, but to our knowledge registration is not obligatory in any of the organizations.

13 Preservation

At the moment, the National Library only preserves the web archive at bit preservation level.

The National Library is participating in the National Digital Library project. The project is building a shared long-term preservation system for Finnish libraries, archives and museums. Provided that the project and its funding do not face any setbacks, the long-term preservation service will be available in 2014.[8]

The service will provide its users with all elements needed for long-term preservation: bit-level preservation, risk management, preservation actions, and dissemination. The service will also give support at all steps of preservation planning in the organizations.

It is clear that preservation of web archives is a complicated task and requires continuing research and development. We may be able to keep data on every single file usable, but still the heterogeneity of web materials, their links, and their dependence on applications makes the preservation of the functionality of a web archive a formidable challenge.

[8] Further information: http://www.kdk2011.fi/en .

14 Future Directions

The present practice of web harvesting will be an important part of the Library's work on electronic cultural materials in the future, too. However, we need to find ways to complement our collections where a web crawler is not applicable.

That is why in 2011 the Library will put more focus on depositing web materials. We have entered into talks with e.g. e-book publishers in order to find practices of deposit that will be easy for the depositors as well as for the Library. In addition to actual web materials, the Library is interested in receiving digital surrogates of printed matter, i.e. files that are used in the printing process of newspapers.

Especially through the National Digital Library project, the Library hopes it will be able to reach out to the web publishers' community and add awareness about web preservation issues.

We also should find methods and resources for better quality control, especially in checking the completeness of harvested materials.

As described in the previous chapter, the Library is working in cooperation with the archive and museum sectors in order to create a robust and modern framework for digital long-term preservation. We must add our participation in international cooperation, too.