

THE NATIONAL LIBRARY of Finland Bulletin 2009

16 40

[Home](#)

Leena Jansson

Finland's Internet Archive

The National Library of Finland is responsible for the collection, description, preservation and accessibility of materials related to the national imprint. The National Library of Finland serves the entire nation by preserving the published cultural tradition for future generations. Thanks to legal deposit legislation, almost all of Finland's published cultural heritage is available as source materials for history and cultural researchers.

Besides printed products and sound recordings, the *act on collecting and preserving cultural materials* that went into effect in early 2008 also covers online materials as well as radio and television recordings. The law obligates the National Library of Finland to retrieve and record public online materials located on Finnish Internet servers as well as online materials located on foreign servers, but particularly intended for the Finnish public. The task of the National Audiovisual Archive is to archive radio and television programs. (*Act on Collecting and Preserving Cultural Materials 1433/2007*, and statute that will go into effect in 2009).

The Internet Archive is thus the newest expansion of the National Library of Finland's legally mandated maintenance of the National Collection. Although the mapping of the National Collection began officially with a law that went into effect in 1707, Finnish printing presses had already been obligated to furnish legal deposit copies of their products to the Library since the 1600s. Times change, as well as publication modes; in the Information Society an increasingly larger part of communication and public discussion has been transferred to data networks. For future generations and Internet researchers in particular, the archive forms a continuously expanding source of information pertaining to the Finnish online world as well as the phenomena it depicts.



Member of Parliament Jyrki Kasvi opened the Finnish Web Archive at the National Library of Finland on 2 April 2009. Tommi Jauhainen presented examples of the archive's materials.

The Internet Archive's content and collection

At periodic intervals, the National Library of Finland retrieves and stores materials available on public information networks representatively and multifacetedly. Besides web harvesting, selected so-called thematic harvesting is also carried out. The term web harvesting refers primarily to the highly automated collection of online materials performed by applications developed for that purpose.

Web harvesting processes are divided into so-called Finland harvesting and its supplementary thematic harvesting. In Finland harvesting, domestic online materials are collected with automated collection programs. The annual collections are not based on the selection of specific subject; the objective is to obtain a wide-angled snapshot of the network's content. The Library archives web pages whose domain name is ".fi" or ".ax". Efforts are also made to archive domestic websites whose domain name is ".com", ".net" and so forth. The harvesting of Finnish websites is implemented at least once a year.

Until now, the National Library of Finland has downloaded a total of 131 million files from the Internet; this includes millions of websites and image files as well as thousands of word, audio and video files. For example, there are 77 million .html files, or individual web pages, 41 million image

files and hundreds of thousands of audio and video files. Among the recorded pages are materials produced by communities as well as private citizens.

Besides the annual collections, the Internet Archive is expanding the scope of its thematic harvests; these are conducted throughout the year and often concentrate on specific subjects or timely events. The purpose of the thematic harvests is to anticipate future research needs and supplement any areas missed in the annual harvesting. The subjects of the harvests can be, for example, significant national and political events, occurrences whose materials tend to disappear quickly from the Internet, as well as unexpected situations in world politics, natural catastrophes, and other similar phenomena. Thematic collections can also be carried out jointly with memory organizations and various research institutes. These collections are also implemented with automated web crawlers, but experts are employed in content planning and the checking of results. Examples of thematic harvesting already carried out include election websites and materials related to Finns living abroad.

Materials associated with use restrictions, chargeable online publications or a reliance on databases (for example publication banks) cannot be recovered fully automatically with the currently available tools. With respect to the recording of these materials groups, the National Library of Finland cooperates case-specifically with sponsors and publishers.

Online materials have been collected and recorded since 2006 based on copyright legislation. The act on collecting and preserving cultural materials made the opening of the Internet Archive for public use possible.

Use of the Internet Archive

The National Library of Finland opened its Internet Archive on 2 April 2009. As planned, the archive of radio and television programs maintained by the National Audiovisual Archive will become publicly accessible in 2010. The National Library of Finland's customers can utilize the Internet Archive at specially equipped workstations on the Library's premises. Initially there has been only one workstation for customer use, but two more will be activated in the early spring. Subsequently, the service will also be expanded for customer use at the country's other legal deposit libraries, (the Joensuu, Jyväskylä, Oulu and Turku university libraries and Åbo Akademi Library) the National Audiovisual Archive and Library of Parliament. The radio and television archive in the University of Tampere's journalism research units will also become available.



A telephone communication that is recorded nowhere. The message itself binds people just as effectively, whether it is a question of a fixed-line telephone, such as those in a city library, or a cell phone at the steps of an art museum. People are always accessible, perhaps also controllable.

Customers working locally on the institutes' premises can access the Archive's resources only with hardware equipped for this purpose. The legal deposit workstation and Internet Archive are subject to strict security requirements. For that reason the machine has no Internet connection, nor can digital copies of the Internet Archive's materials be made with normal storage devices such as, for example, a USB memory stick. The legal deposit workstation is connected to a printer where users can print out hard copies of the Archive's documents for their own use. So-called indirect copying is allowed, and for example display screens may be photographed. Concerning the customer equipment, the intent will be to anticipate the Internet's most common file formats and install the software necessary for the viewing of the Archive's contents. The material harvested from the Internet consists primarily of websites as well as the images and other materials on the pages. Using the Internet Archive requires no recognition, nor is a register of users maintained. Access to certain auxiliary services planned for the Archive may require customer registration, but in principle the use of the Archive is fully possible without it.

This is a complex and unique project, even by global standards. The data security requirements for customer hardware are extremely high. Besides data security and data protection, other factors possibly limiting the utilization of the Internet Archive have arisen throughout the world. In Denmark, for example, data protection authorities have taken the position that online materials harvested for their Internet Archive may also contain sensitive personal information. There, for data protection reasons, the Internet Archive is closed to the general public and the use of the materials is limited to scientific purposes. In Finland, customer access to the archive is governed by Section 16 of the Copyright Act (28.12.2007/1436, Use of Works in Libraries Preserving Cultural Materials).

The Internet Archive's users



Classical music on Aleksanterinkatu. Everyday history that will only be stored in the Web Archive if a passerby at the scene downloads a video to his or her own website, and the National Library of Finland's web harvesters happen to record it.

In compliance with the *Act on Collecting and Preserving Cultural materials 1433/2007*, the National Library of Finland also archives, besides books and magazines, everyday printings (ephemera, advertisements, annual reports and posters), that have a tendency to vanish over time. The same principle has been applied to the archiving of Internet websites. At the Internet Archive, particular attention is paid to the preservation of materials exhibiting the Internet's particular characteristics.

The most frequently used materials in the National Library of Finland's National Collections are newspapers, magazines and ephemera. These are used as source materials, particularly in historical and sociological research studies. The Library expects the Internet Archive's users to be at least partially the same as those using conventional materials, but besides researchers in the humanities and social science fields, it is expected that information technology researchers, those interested in the history of graphical design, as well as game researchers, will be among those benefiting from the Internet Archive's resources. According to Researcher Jaakko Suominen, the Internet Archive is an excellent source for studying the Internet itself, but a researcher studying any phenomenon past the mid-1990s can benefit from the downloaded and archived websites. In the future, the use of websites as research sources will be inevitable. Until now, accessing radio and television programs has been fairly difficult. The opening of the radio and TV archives will bring a welcome addition to media researchers' source materials.

The Internet Archive's usability

For copyright reasons the service is not accessible over the Internet, but a directory for the Internet Archives will be opened in the spring of 2009. This is a service in which an Internet address can be used to check if a certain page has been downloaded into the Internet Archive. The directory can be accessed at: <http://verkkoarkisto.kansalliskirjasto.fi>.

There are two ways to search the Internet Archive: address and keyword search. With the website address search, pages can be searched directly on the basis of a website address. The keyword search is the well-known text search from Google.

Most of the web pages stored in the Internet Archive are not referenced in the Library's databases. The intent is to only transmit information concerning thematic collections (Internet archiving related to certain themes or timely events) to the National Bibliography or National Discography, in other words the Fennica or Viola databases.

*Leena Jansson, Planner
The National Library of Finland*

[Print this article \(PDF\)](#) [Print entire issue \(PDF\)](#)

[Home](#) | [▲ Top](#)

NettiAsema4