Human IT

Tidskrift för studier av IT
ur ett humanvetenskapligt perspektiv

Preserving the Digital Heritage of the World

some thoughts after having collected 30 million Swedish web pages

by Johan Mannerheim

This paper is dedicated to the tremendous task of archiving the digital heritage of the world. Much thought and many discussions and papers have been given to the question of web archiving during the last years, while most of the documents on the Internet have disappeared and cannot be retrieved again. I will discuss different possible approaches to some of the problems relating to the preservation of digital information, against a background of experiences from the Kulturarw^3 Project of the Swedish National Library as well as other ongoing projects.

Contents

1. The Kulturarw^3 Project

The Kulturarw^3 Project started in September 1996, when the Royal Library hired an engineer, Johan Palmkvist, and I was made part time project leader. It was initially financed by a government grant of 3 million SEK (Swedish crowns) to test methods of collecting, preserving and providing access to Swedish electronic documents.

The name Kulturarw^3 means Cultural Heritage in Swedish but is properly spelled with a "v" at the end. The "w" has the same sound value in our language and we have indexed it to point out that the WWW or World Wide Web not only is something new and modern, but also part of our cultural heritage.

The project has made six comprehensive harvests of the Swedish Web since January 1997: two in each year. The seventh harvest will soon be completed. A harvesting robot is used to search and retrieve Swedish web pages within the domains ".se", ".com", ".net", ".org" and ".nu". A flow chart shows how it works (fig. 1).

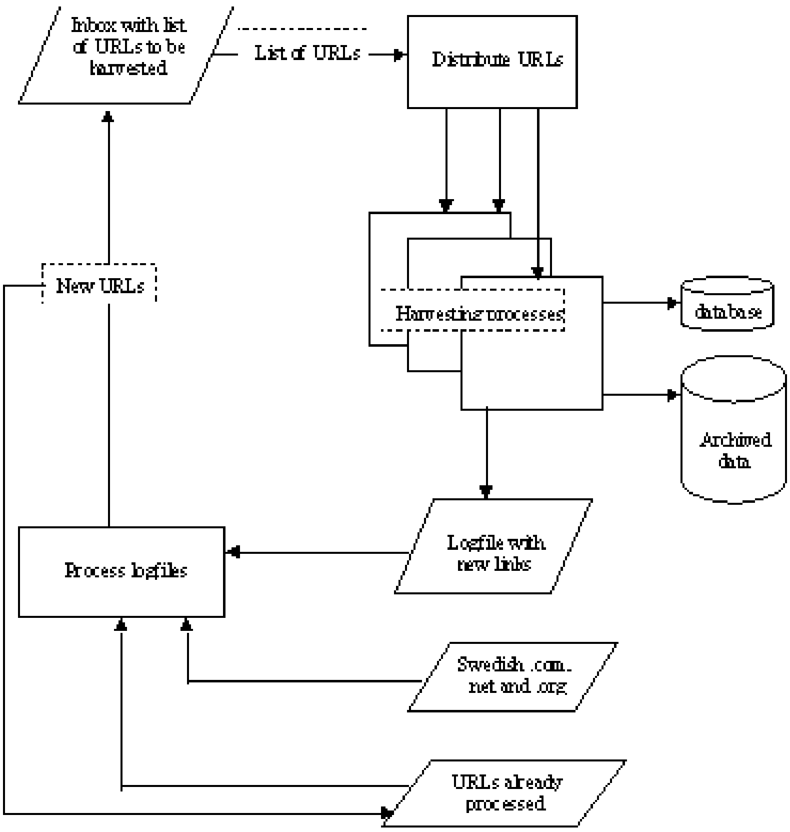

Fig 1

The current number of Swedish web pages on the Internet is about 8 million from 63 000 web sites (37 000 of which are ".se" and more than 26 000 registered in other domains). Including pictures, sound etc. there are 16 million files. The total size is under 300 GB. Our collection so far comprises about 75 million files and 1.4 terabyte or 70 DLT-tapes. So the big problem is not the actual size of the archive but rather to handle the large number of files.

The project has got a special grant of 5 million SEK from the Knut and Alice Wallenberg's Foundation for obtaining archiving equipment. An archival computer with a disk array of 1.5 terabyte is in use to test different ways of organising the web archive. A tape robot storage system will be installed in 2000; it also serves other projects dealing with digital preservation.

There are at least a hundred different file formats in the collection, some of them standardised or relatively standardised ones such as HTML and the picture formats JPEG and GIF, other ones are proprietary and probably have a shorter life span, such as different versions of MS Word, Excel and PowerPoint. HTML, JPEG, GIF and plain text together make up 97% of the files. It is among the other 3% of the files, that we will find the more immediate problems. Migration, that is conversion to readable formats, will probably be the normal method to keep old documents readable.

Access to the Kulturarw^3 documents is as a rule not allowed until a legal framework has been created, which will be accomplished mainly by an ongoing revision of the Swedish deposit law. The Swedish ministry of education has commissioned a report, which was published late in 1998. The report suggests that the Royal Library and the National Archive of Recorded Sound and Moving Images divide the respons-ibility of preserving and giving access to the historic Web. However, the report also suggests that the access will be limited to researchers from established research organisations. Such a limitation would be contrary to the democratic aim of the Swedish deposit law to guarantee free access to information. We are now waiting for the politicians to make a long-term decision on web archiving.

Kulturarw^3 co-operates with another, later, Royal Library project called Svesök (meaning something like Swesearch in English), which is creating access tools to the current Swedish web. Svesök gets text files from Kulturarw^3, adds descriptions in the Dublin Core format to a very small selection of home pages (3200 at the moment) and puts them into a subject tree structure. This selection is the electronic publication part of the Swedish National Bibliography. All the text pages which Svesök retrieves from Kulturarw^3 are automatically indexed. A search robot is provided which lists those pages which have DC descriptions first. Kulturarw^3 intends to find ways to save the efforts of the Svesök cataloguing into the historic web archive.

At present the staff of Kulturarw^3 consists of two persons, Allan Arvidson and Krister Persson.

The focus of this paper will now move from a description of one way of preserving web material, the Swedish project as it is today, to a discussion of some possible approaches to the challenges of collecting web publications, and of preserving and giving access to digital information.

(To the top)

2. Collecting

What?

The first thing to decide is what to collect. In today's projects you will find two main approaches.

The comprehensive one is represented by the Kulturarw^3 Project, by Brewster Kahle's Internet Archive and, more recently, by the Finnish EVA Project. The scope is to collect everything published on the Internet. These projects are collecting millions of documents. The selective approach is represented by the PANDORA Project of the National Library of Australia and EPPP (Electronic Publications Pilot Project) of the National Library of Canada. The scope is to collect important publications which can be made accessible at once. They are "only" collecting thousands of documents.

An argument for being selective is that you should not spend your limited resources on preserving lots of trash. However, doing an intelligent selection is difficult and researchers in the future will criticise our choices. Even if we try our very best, important digital information will get lost.

Computer storage is getting cheaper and cheaper, while the cost of personnel is not. It might seem a paradox, but it is a fact that the selec-tive projects use more staff than the comprehensive ones.

If selection is made in the indexing process, and not in the collecting process, we have at least saved the publications and the inevitable mistakes we will make when we select publications for cataloguing and indexing, can be corrected in the future.

Who?

Who should preserve the digital publications? There are at least three approaches to this problem. One is to make the publishers and other institutions directly responsible as was advocated in the USA by the Task Force on Archiving of Digital Information in 1996. The second is the national approach exemplified by Denmark and by the Australian, Canadian, Finnish and Swedish projects. The third is the international one represented by the Internet Archive.

Long-term preservation should be undertaken by long-term institutions with stable financing, that last for hundreds of years. To give the task to the national library in each country, widening its responsibility for printed publications to include digital publications, based on rewriting the deposit law, seems to be a good solution for many countries. Collection and preservation is best done at one institution with good resources, while indexing and selection might be done in co-operation with other institutions.

The institutional approach is not so stable. It also combines badly with automatic, comprehensive collecting of web publications, as each publisher and institution will find their own solution for preservation of their own publications. Links pointing to resources on other sites will not function.

The interactive character of the web pages with links to other pages, regardless of national boundaries, speaks for the international approach. But there seems to be a long road to go before it would be possible to create an international institution for web archiving with long-term stable financing. It seems more realistic to start co-operation between national web archives, not only to exchange experiences and provide each other with support, but to create a forum for raising questions of standards, exchange formats, communication between the archives, etc.

Waiting for a permanent solution, which seems close in Sweden and Finland, but so far fairly distant in most other countries, institut-ions, companies and individuals have to rely on themselves if they want to preserve their old web pages.

How?

The usual way to collect web documents is by harvesting, i.e. using a robot software, which searches for documents on the Internet and retrieves them by downloading a copy. Another way is to let the publisher deliver the material by tape, magneto-optic disk or via the Internet. The harvesting approach is, of course, the only possible one for comprehensive web collecting. Dealing with, as in Sweden's case, the owners of more than 60 000 web sites would be a nightmare. But most selective projects also seem to take the harvesting approach, as it is simple and practical. However, for some of the sites protected by user accounts and passwords, and for very large sites, delivery might be used in the future.

The short life of the pages is a special characteristic of web publications, which makes them different from printed publications and electronic publications on CD-ROM and other carriers. It is so cheap and easy to change a web page. The average life of publications on the Web (or rather of editions and issues of web publications) is only some months. One must take this into account when one decides between the snapshot approach and the continuous approach.

The snapshot approach is to take two, four, six or another number of snapshots of the Web each year and let that represent the web publications of that year. It is an attractive way to select automatically and reduce the size of the web archive. The main problem with the snapshot is that you will lose information like newspaper and journal issues and other important pages, which have a short life. This means that you have to give certain web publications special treatment, which will increase staff time and costs. This is for instance done by the Australian Pandora project.

The continuous approach is not in practical use today. The idea would be to collect as many editions and issues as possible. To do that one needs a harvesting robot that collects information about the frequency of change of each URL, and uses that information for its collecting strategy. According to expertise it would not be too difficult to construct such a robot software.

(To the top)

3. Preservation of digital information

The next set of problems concerns the long-term preservation and access of digital information in general (of which web publications constitute one subset). The amount of digital information created is increasing drastically. The time when word processors and economy systems were tools to create written or printed documents is gone. Now more and more information is primarily digital. It might be in a text format like MS Word, HTML or XML, in an image format like TIFF or JPEG, in some kind of data base or in a more specialised system. Today, not only print-outs but also printed reports should often be regarded as secondary forms which are used to spread the information or a selection of it on paper, as well as different digital formats like HTML, PDF and reports in Excel could be secondary forms to spread the information on intranets or the Internet. But for long-term preservation, most institutions and companies still stick to paper and in some cases microfilm, when they are not turning a blind eye to the problem.

I will take one example from a research library perspective: what happens to the manuscripts of today? For centuries, The Royal Library has collected personal archives of authors and other persons related to the publication and production of books. These are frequently used sources for studies in literature, art, history and other academic disciplines. Today, the corresponding material is in the author's PC till she or he buys a new computer, when most of it gets lost. Therefore, on the initiative of the author and professor Sven Lindqvist, a member of the library board, we have just started a project to find ways of preserving digital personal archives. Such an archive might include different versions of texts reflecting the creative process, as well as e-mail correspondence and research material collected by the author.

The difficult preservation problem is not the lifelength of tapes and other carriers of the information, as it is easy to copy the 1s and 0s of which the digital information consists, and the copy is identical with the original. The difficulty is the short life of the software and hardware environments. You need a new computer every third year and have problems reading documents older than ten years, because today's software can only import a limited selection of file formats.

The question is: how are our successors going to read the digits we have preserved for them? There are at least three approaches to this problem: the technological museum; the migration; and the emulation approaches. The first approach would be to create a technological museum with old computers and software. But this would be only a temporary solution as you soon will run out of spare parts and the cost to uphold the knowledge of how to run the systems and software will rise tremendously.

The migration approach means successive conversions of the files to current formats, when the old ones are outdated. That means a maintenance cost for the archive, but a cost that can be controlled. One of the drawbacks of migration is that it is inevitable that you will sometimes lose some information or functionality of a document when it is converted from one software to another. Even if the textual contents will be correct and complete, some of the authenticity of the document will get lost. So you need a good strategy, trying to use standards and as few conversions as possible.

The emulation approach means reading old files by writing new software in your current computer environment, emulating the old programs or at least the reading part of them. In a way, this is the most comfortable approach. You save the information in the original format and rely on the ability of future generations to create reading software for their use.

In my opinion a combination of the migration and emulation approaches is the best way to go.

(To the top)

4. Digitisation and preservation

Yesterday's information can be made accessible by digitisation of texts and images on paper and other traditional materials. But the digitisation and digital library projects of today are focused on availability only and are seldom taking preservation consequences into account. When the quality of the digital images is low, the projects will generate more use of the originals for study and reproduction and cause a threat to their survival.

On the other hand, preservation planning seldom includes digitisation as a means of protecting the originals from use and tear. There is a need to bring professionals from the different spheres together into co-operation and united thinking.

At the Royal Library we have started such a project, called Platform for image databases, tackling questions like the quality needed for archival copies, presentation copies and delivery copies (in terms of e.g. resolution and colour depth), standards, recommendable file formats for archival purposes, presentation and delivery copies, compression, safety, permanence of digital information and migration as well as questions concerning registration and cataloguing of images, search methods and user interfaces. A full report in Swedish will be published before summer.

(To the top)

5. Access

If you search for a tree on the Internet today, you will get the whole forest as an answer. In the long list presented, you will be lucky if you find a relevant hit on page seven. This problem will not lessen and the list will not be shorter in a historic web archive. Cataloguing, even if it is done at a minimum level, can hardly be accomplished for more than some per mille of the web pages. (8000 is one per mille in Sweden's case.) Therefore, it is important to promote the use of metadata in order to help and encourage the producers to make their own cataloguing and put that onto the page.

After years of discussion, it seems that the Internet community rallies around the metadata format Dublin Core. The Royal Library promotes metadata by meetings and by information on the Web, by having a template for Dublin Core creation at Svesök, and by encouraging other actors to also provide Dublin Core templates.

Automatic indexing and cataloguing might also be used more for digitised material in the future. A possible development for the retrieval of web publications is illustrated in figure 2.

Web retrievability

The large squares represent the Web or the Swedish Web or some part of the Web

Today In 5 years?

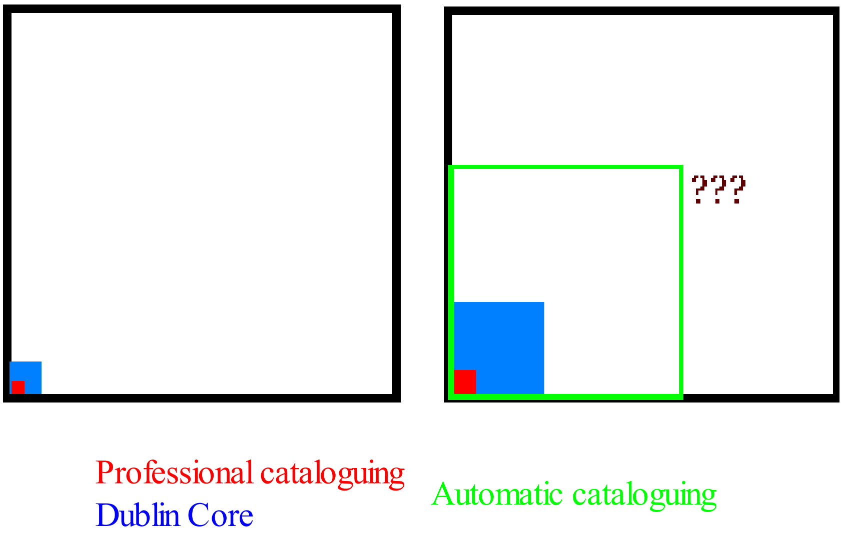

Fig 2

To give access to web publications is part of the much larger challenge to make all kinds of historic digital information easily available in the future, both documents and objects which are originally digital and those which are the results of the digitisation of older collections in archives, libraries and museums. In a project called Automatic Indexing of Newspapers at the Royal Library we try using optical character reading of old newspapers, some of them printed in gothic type. The resulting digital text is then used just as it is (without any expensive corrections made) for indexing and fuzzy search. The hits are linked to the image files of the newspaper pages, as the digital text is corrupt and almost unreadable.

Let us hope that techniques developed within different specific projects can be scaled and applied to large segments of digital infor-mation in the future. The important thing is to find as automated methods as possible for the retrieval of and access to the information, as the cost of manual handling is much too high for such a large number of objects.

(To the top)

6. Co-operation in its beginning

In 1997 Kulturarw[3] initiated an informal group of technical co-operation within the Nordic countries called the Nordic Web Archive. We have also downloaded web pages for seven Central American countries, viz. Belize, Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua and Panama. Within IFLA, the International Federation of Library Associations and Institutions, there is a growing interest in web archiving. An open session on the subject will be held this year in Jerusalem.

There is certainly need for much more co-operation in the future both on preservation of digital information as a whole and on web archiving. Then, just to take one example, it will perhaps be possible to follow an old link on a web page in one national web archive to the proper document from the same time in another.

(To the top)

About the Author

Johan Mannerheim, fil. kand. med både naturvetenskapliga och humanistiska ämnen, bibliotekarie, avdelningsdirektör, chef för Data och IT-enheten vid Kungl. biblioteket. Han har byggt upp den nationella mikrofilmningen av svensk dagspress i KB:s regi och medverkat i standardisering inom områdena mikrofilmning och "imaging". Han är boksamlare med inriktning på bokens historia och har undervisat vid Grafiska institutet i ämnet. Sedan mitten av 90-talet har han ägnat sig åt att utveckla KB:s infrastruktur på IT-området och är bl.a. ansvarig för den insamling av svenska webbsidor som KB bedriver för att bevara dem för framtiden.

(To the top)

Literature

E-plikt – att säkra det elektroniska kulturarvet, 1998, (SOU 1998:111), governmental committee report on securing the electronic cultural heritage,
http://utbildning.regeringen.se/propositionermm/sou/pdf/1998/sou98_111.pdf (in Swedish)
Internet Archive
http://www.archive.org/
Kulturarw^3
http://kulturarw3.kb.se/html/kulturarw3.eng.html (in English)
http://kulturarw3.kb.se/index.html (in Swedish)
Kungl. bibliotekets yttrande angående pliktutredningen, Pronouncement of the Royal Library on the committee report E-plikt
http://www.kb.se/BIBSAM/EPLIKT/kbsvar.htm (in Swedish)
Lagen om pliktexemplar (SFS 1993:1392), the Swedish deposit law
http://www.notisum.se/rnp/sls/lag/19931392.HTM (in Swedish)
Metadata och Dublin Core
http://www.kb.se/bus/dc/dcstart.htm (in Swedish)
National Library of Canada Electronic Collection
http://collection.nlc-bnc.ca/e-coll-e/index-e.htm
PANDORA Project
http://pandora.nla.gov.au/
Platform for image databases
http://www.kb.se/DoIT/bildbas_eng.htm
Preserving Digital Information
Report of the Task Force on Archiving of Digital Information commissioned by The Commission on Preservation and Access and The Research Libraries Group, Inc. May 1, 1996
http://www.rlg.org/ArchTF/
Project EVA
http://linnea.helsinki.fi/eva/english.html
The Royal Library
http://www.kb.se/ENG/kbstart.htm (in English)
http://www.kb.se/ (in Swedish)
Svesök, search for current Swedish web pages

http://www.svesok.kb.se/ (in Swedish)

(To the top)

Return to Human IT 1/2000