

The Kulturarw³ Project - The Royal Swedish Web Archiw³e

Krister Persson M. Sc., Project Kulturarw³
Allan Arvidson, Ph. D., Head Project Kulturarw³
Johan Mannerheim, Ba. Sc., Head DoIT
Department of Information Technology (DOIT)
The Royal Library, The National Library of Sweden

In 1661 the Royal Library (Kungl. biblioteket, abbreviated KB) was assigned the task of collecting all Swedish printed publications. Since then KB has collected, preserved and given access to an important part of our cultural and historical heritage. In the future an increasing amount of material will be published on the Internet, and only on the Internet. If the Royal Library is going to continue to fulfil its historic role, the activities must be widened to encompass also what is published electronically. In 1996 KB inaugurated a project, entitled Kulturarw³ (The Swedish Archiw³e) to address those issues. The aim of the project is to test methods of collecting, preserving and providing access to Swedish electronic documents, which are accessible on line in such a way that they can be regarded as published.

Since the start of the project there have been made seven complete downloads of the Swedish web. Currently, the collection comprises about 65 million items. About half of them are text documents, mostly html and plain text. Through this project KB is also laying the foundations of a collection of Swedish electronic publishing for our time and for coming generations.

Collecting the web

How?

There are basically two approaches to how to collect electronic documents: First, there is the comprehensive approach. In this approach the goal is to collect everything on the Internet by means of automatic software. Second, there is the selective approach where documents deemed worthy of preservation are selected by humans.

The Kulturarw³ project has chosen the comprehensive approach for several reasons. One doesn't know what information future generations will consider important. It also requires humans to make the selection, i.e. it demands manpower. Computer storage is also getting cheaper. In fact it is probably cheaper to collect everything then to take only a selection. Also, in the legal deposit material there is no selection.

What?

The next question to ask is what to collect.

The first problem is to define what is Sweden on the Internet. There is no clear definition. In this project we define Sweden as 1) everything that has a server address ending on **.se**, 2) generic top-level domains (**com**, **org** and **net**) registered with a Swedish address or telephone number, 3) Swedish domains under **.nu** (Niue, *nu* means *now* in Swedish). There is no selection on document type, i.e. all picture, sound and other file types are collected.

It should be noted that it is very difficult, if not impossible, to be complete. There will always be webservers which are not found, Swedish material residing on servers registered under the "wrong" country code etc.

Strategy

The collection strategy is to take snapshots a couple of times a year. The collecting robot starts with an empty collection and harvests every page once and then stops. In this way a complete copy of the Swedish web is stored each time. To be a real "snapshot" the collection time should be as short as possible. In practice it takes a couple of months. The limiting factor is the big websites, which it takes a long time to harvest completely.

Problems

The strategy chosen will miss a lot of short-lived material, in particular web newspapers. For such material one should of course try to get every issue. In the future we want to use a more flexible approach: the daily papers will be harvested every day, the weekly every week and so on. There is also the possibility to have the search robot automatically check how often certain material change and adjust its strategy accordingly.

Another problem is pages that demand some form of interaction from the visitor. Such material is generally lost since a software program can't fill in key words in a data field. Extreme examples of this are sites which requires you to download a plug-in to be able to navigate the site.

When visiting the websites we respect information about what to acquire and index, i.e. robots.txt files and robots metadata. However, such data are usually made up with an indexing robot in mind. Often pictures and short-lived material are blocked for access on the grounds that you cannot index pictures and it doesn't make sense to index very short-lived pages; before it is indexed and loaded into the database it has disappeared from the server. For this project however it is important to get such material.

There are also problems with authenticity since a page usually is made up of several objects and they, by necessity, are harvested at different times. Suppose a page is gathered at some time. This page happens to have an inline picture. For some reason some macroscopic time passes before the picture is acquired. Meanwhile (between getting the main page and getting the picture) an update is made of the page in which both the text of the page is modified and the picture is exchanged for a new. This will mean the archive we harvest will associate the wrong picture with the text, i.e. we will reconstruct a page that never existed! Acquiring all inline material as soon as possible reduces this problem.

Sweden on the web

Since the start of the project seven downloads of the Swedish web has been done, the first in summer of 1997, the latest during the spring of 2000.

In the latest complete download, spring 1999, 15 million files were collected corresponding to about 7.5 million pages. The data amounts to about 300 MByte. More than 100 different MIME-types have been found. However, the four most common, text/html, text/plain, image/jpeg and image/gif, comprises about 97% of all documents.

Since the first download the number of webserver under .se has risen from 16700 to 37100. Please note that in the first download only webserver found under .se where accessed. In the latest round also 25600 non-.se webserver where processed, i.e. about 40% of the total. The increase seems to be somewhat smaller than the 18-months doubling time often quoted for various other Internet related parameters.

Accessing the Material

The aim of a national web archive must be to make the collected material accessible to as many people as possible. There are many reasons for that. Among others: Secure every citizens right to free access of information. Secure the cultural heritage, where a great deal of information is distributed exclusively on the web. In order not to loose track of our cultural development the collecting efforts have to include the web. One has also to note the difference between the intellectual content and the more technical aspects of collecting and storing web documents.

Surfable

The archive must be organised in such a way that navigation in the material (surf the historic web) is easy. The time aspect adds a further dimension to the web, which might be compared to an ordinary map in two dimensions. Collecting different instances of the web can be resembled with making further map-sheets. Time forms the third dimension making it possible to travel (taking the time-elevator) between different time levels of the web. Surf the historical web must be given the added feature of changing time frames. With this possibility it is easy to scan the historical development of a web site. A first version of such an application has been implemented in the Swedish Kulturarw³ Project.

Searchable

The task of making the archive searchable can be divided into two subtopics:

First, there is a need to keep track on every stored object/file (i.e. html, text, pictures and sound). A great deal of information can be extracted from the web documents themselves. In addition information about each object, such as size, file type and creation time, can be extracted by the collection software. There is of course a possibility to keep track of hypertext links between objects. With a relatively simple database model one can achieve things such as bi-directional links. It will, for example, be possible to answer questions such as *How many links point to this object?*

Second, not every item/file of the archive is indexable in the traditional way. But there is a great deal of text objects, where free-text indexing is possible. Objects that are not indexable in a traditional manner may be "indexed" according to context. I.e. inline pictures may be indexed regarding the web document they belong to.

Researchers may probably be more interested in traditional ways of information retrieval. Therefore free-text search or other more advanced indexing facilities have to be added to the archive. Due to the huge amount of collected items, the indexing has to be performed automatically.

Certainly, the time aspect has to be added to the search facilities. One must be able to limit the time interval wherein the search is performed. This extended free-text search is important and needed in order to get a view over the archive.

How to bring order to all the archived material?

So far the discussion has only covered the technical aspects of collecting web material. What about the intellectual content? In order to make the archive useful for future scientists and researchers there has to be an indexing mechanism classifying each and every object in the archive. Before this is done, accessing the web archive will differ from accessing a traditional library. A free-text index will only be a first step to provide access to the intellectual content. Compare this to how to find documents on the web today using a search engine! If one wants to make the web archive look more like an ordinary archive, several functions have to be added, e.g. cataloguing. Due to the huge amount of information, this has to be done automatically. A lot of development needs to be done in this area.

Future development

At present there is no public access given to the archive because a legal framework is missing. Even if such a framework has been implemented there are still topics that must be addressed, e.g. copyright control.

There are many possible future applications. Advanced analysis of the contents of the web will be possible. Language engineering will also open up new research fields with improved performance of the indexing facilities. Different kinds of web statistics may be an interesting fringe benefit of the archiving projects. The archive will of course also provide the opportunity to forecast future development of the Internet.

Preservation of digital information

Will we be able to look at the collected information in the future? In the Kulturarw³ project we have found more than a hundred different file formats. Some of them are rare others are more common. None the less there will be changes in standards. Will we be able to look at the earlier generations of web documents?

The difficult preservation problem is not the length of life of tapes and other carriers of information, as it is easy to copy the bits and bytes of which the digital information consists, and the copy is identical with the original. The difficulty is the short lifetime 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 maintenance cost for the archive, but a cost that can be controlled. One of the drawbacks of migration is that it is inevitable, that sometimes you will 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 our opinion a combination of the migration and emulation approaches is the best way to go.

How to store the collected files?

Every file has to be stored undistorted. In the archive hyperlinks that connect different web pages with each other have to be redirected. One way to achieve this is to make the redirection *on the fly*, when the object is requested/displayed. The motivation for this is to keep the archived object in its original form. Amongst other things this makes it easier to develop future implementations of the archive interface. Obviously it is also important to keep the collected items in their original form for the sake of authenticity.

A database containing information about every item in the file collection is also needed. N.B. there is different demands of such a database. There has to be stored technical information about every object (file size, type, URL, links to other URLs, info about inline objects etc.). This information is needed to make it possible to recreate old websites. This database is separated from a library database containing ordinary library records.

Where to store the collected files?

The Kulturarw³ Project has chosen a storage facility containing a big disk array connected to a tape archive robot. Today's development on mass storage devices is interesting and in future it may be possible to keep all the material on line (i.e. on disks). For safety purpose two or more complete copies of the archived file collection will be produced. Therefore, it is not necessary to back up the archive. On the other hand the database overhead and its possible changes have to be backed up regularly. A disk array connected to a tape robot is our choice of hardware.

Contact information

The Kulturarw3 Project:

Web page: <http://kulturarw3.kb.se/>

The Royal Library

Web page: <http://www.kb.se/>, <http://www.kb.se/ENG/kbstart.htm>

e-mail the authors

Krister Persson: krister.persson@kb.se, +46 8 463 4068

Allan Arvidson: allan.arvidson@kb.se

Johan Mannerheim: johan.mannerheim@kb.se