DOCUMENT RESUME

ED 450 729 IR 057 985

AUTHOR Arvidson, Allan; Persson, Krister; Mannerheim, Johan
TITLE The Kulturarw3 Project--The Royal Swedish Web Archiw3e--An Example of "Complete" Collection of Web Pages.
PUB DATE 2000-08-00
NOTE 7p.; In: IFLA Council and General Conference: Conference Proceedings (66th, Jerusalem, Israel, August 13-18, 2000); see IR 057 981.
AVAILABLE FROM For full text: http://www.ifla.org/IV/ifla66/papers/154-157e.htm.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Access to Information; *Electronic Publishing; Foreign Countries; Information Retrieval; Library Materials; *Library Services; National Libraries; Preservation; Swedish; *World Wide Web
IDENTIFIERS Legal Deposit; Sweden

ABSTRACT

This paper describes the Kulturarw3 Project, a project of the Royal Library of Sweden to test methods of collecting, preserving, and providing access to Swedish electronic documents that are accessible on the World Wide Web in such a way that they can be regarded as published. The first section discusses issues related to collecting electronic documents, including how to collect them, what to collect, the collection strategy, and problems. The second section provides data on the number of Swedish documents on the Web. The third section addresses accessing the material through "surfing" and free-text search. The fourth section covers the preservation of digital information, and the fifth section considers legal issues related to the legal deposit of electronically published material. Contact information for the Kulturarw3 Project, the Royal Library, and the authors is provided. (MES)

66th IFLA Council and General Conference

Jerusalem, Israel, 13-18 August

Conference Proceedings

Code Number: 154-157-E
Division Number: VI
Professional Group: Preservation and Conservation
Joint Meeting with:
Meeting Number: 157
Simultaneous Interpretation: *No*

The Kulturarw3 Project - The Royal Swedish Web Archiw3e - An example of "complete" collection of web pages

Allan Arvidson, Krister Persson *Project Kulturarw*^{3}
and
Johan Mannerheim
Division of Information Technology (DoIT)
The Royal Library, The National Library of Sweden

Paper

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 Kulturarw^3 (The Swedish Archiw^{3}e) 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 Kulturarw^3 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 than 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.

To collect a few times year is of course inadequate for eg web newspapers. For such material we will in the future try to get every issue. In general we need to distinguish between different material: 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 schedule accordingly.

Problems

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. An extreme example are sites which require you to download a plug-in to be able to navigate the site.

When visiting the websites we obey instructions about what to acquire and index which the servers supply, 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. In this project, however, it is important to get such material. There are many cases where we would like to ignore such instructions. We choose to obey them since at present the legal framework for our activity is unclear. Also, we believe that in the long run it will pay off to be nice to people and obey "netiquette".

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 a certain 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 one. This will mean that in the archive we 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 most recent one 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 Gbyte/sweep. 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 webservers under .se has risen from 16700 to 37100. Please note that in the first download only webservers found under .se where accessed. In the latest sweep also 25600 non-.se webservers 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

For the access to the archive we have put priorities on the usual ways to access the web; surfing and free-text search. Less priority has been put on traditional library methods; catalogueing, Z39.50 etc. There are several reasons for this. We believe the users of the archive will be familiar with normal web-related tools and methods. These tools are already available now while many standard library methods are not yet ready for the web material. E.g. catalogueing; the only ways to cataloue the web will be by automatic means, and such software is not yet ready for full scale use.

Also we have decided that no special tools should be needed for accessing the archive. A normal web browser with no plug-ins should be sufficient.

Surfable

The archive must be organised in such a way that navigation in the material (surfing 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. Surfing 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 Kulturarw^3 Project.

Searchable
Free-text search is the next access method to be added. Here a lot of commercial and non-commercial software exists. However, indexing our archive has an added complication; the time aspect has to be taken care of. This mean that it will be possible to search for something with the additional condition of a time span, i.e. everything written about The Royal Library in 1998-1999.

Future development
There are many possible future applications. One of the most obvious enhancements is the possibility to search on metadata. When and if methods for automatic generation of metadata and automatic catalogueing are available they can be added as a new way to access the archive.

The archive offer the possibility to do unique research about the web. Different kinds of web statistics may be an interesting fringe benefit. The archive will of course also provide the opportunity to forecast future development of the internet.

Preservation of digital information

Here we will discuss some of the more technical aspects of the archive. When organizing the material one has to take into account the way it is supposed to be access. Also the number and sizes of the files plays an important role. At present we have collected about 70 milj files, totaling about 1500 GByte. Soon the number of files will be several hundred millions totaling tens of thousand og GByte. In the latest sweep more then 60000 webservers where accessed. These are important parameters when building the archive.

There are also other demands that the achive must fulfill. Among the most imortant are:

- The original document must never be changed
- Several pieces of metadata must be stored about each object. E.g URL and time stamp.
- The files must be organized in such a way that access is easy.

We have chosen to store all information about an object in one single file. This file is defined as a multi-part MIME file (for a description of the MIME standard see http://www.rfc-editor.org/rfc/rfc1521.txt). The file is divided into three separate parts. The first part contains the metadata associated with the collection process: when it was collected, by what version of the software etc. The second part contains the metadata delivered by the webserver, document type etc. The third part contains the actual content of the document. The name of the file is first a 33 character long character string to which is added a time stamp. The url is not a good name for the file for two reasons. An url can contain special characters which have meaning for a certain computer system and they have different length.

In this way everything know about a certain document is contained in one single file. Nothing outside this file is needed to build the archive. We could loose all databases associated with the archive, from the original files we would be able to

rebuild everything.

We can also add other parts to the file when needed. Suppose we decide to migrate a picture from one image format to another. One possibility would be to add one more part to the file containing the migrated version and a part containing history information about the picture; in particular information about the migration process.

We plan to store files that belong together near each other. In practice this means that all files belonging to a certain webserver will be grouped together. This has the advantage that if a document demanded by the user is on magnetic tape (see below), the whole site will be retrieved. Also, considering the number of files that has to be handled it makes good sense to try to group them together.

For the physical storage we have chosen a tape archive controlled by special software, Hierarchic Storage Management (HSM). The HSM software make files that reside on tape look as if they in fact are on disk. Only when you try to access a file that is on tape the difference is noticed because it takes longer to access it since it first have to be read from the tape and stored on disk. An important concequence of this is that when designing access tools, it must be taken into account that most of the material will not be on fast disks.

Legal

At present there is no public access to the archive because a legal framework is missing. The ministry of education has published a report dealing with legal deposit of electronicly published material. In the report it is proposed that the Royal Library, together with the Archive for Sound and Moving Pictures, is given the task of collecting material published on the internet and that it should be done with the methods now used by the kulturarw^3 project. It is also proposed that a selection of databases on the web are collected.

The report proposes that access to the archive should be given to "researchers affiliated with recognised institutions". The Royal Library, however, thinks that the rules governing the access to the web archive should be same as for other legal deposit material, i.e. also accessible to the public. Which is the intention of the legal deposit law; to secure every citisens right to the free access of information.

As yet there has not been any decisions taken on the proposals.

Contact information

The Kulturarw^3 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
Allan Arvidson: allan.arvidson@kb.se
Krister Persson: krister.persson@kb.se

Johan Mannerheim: johan.mannerheim@kb.se

Latest Revision: *August 2, 2000*

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").