68th IFLA Council and General Conference
August 18-24, 2002

Code Number: 111-163-E
Division Number: I
Professional Group: National Libraries
Joint Meeting with: Information Technology
Meeting Number: 163
Simultaneous Interpretation: -

The Collection of Swedish web pages at the Royal Library - The Web Heritage of Sweden

Allan Arvidson
The Royal Library, The National Library of Sweden
Stockholm, Sweden

Abstract:

The Royal Library as since 1997 harvested the Swedish web space regulary. This paper discuss the evolution of the Swedish web since then. We also try to answer the question wether the collection gives a true picture of the swedish web, both with respect to geographical coverage and to various technical issues. We end by commenting on possible future development of web technologies and how that might influence the work.

The collection

The Swedish web has been harvested regulary since 1997. The harvesting as been done using automatic programs in order to collect as much as possible. This has yielded a number of "snapshots" of the Swedish web space from which some observations can be made.

The size of the "snapshots" has grown considerable since the start. In 1997 we harvested 6.8 million urls from 15700 web sites. The latest complete download in 2001 yielded 30 million objects from 126000 web sites. The first download occupied 140 GBytes of data, the latest 1335 GBytes.

The number of different document types has not risen very much. In 1997 there was 295 different mime types found, compared to 424 in 2001. The relative proportion of the different types however, has been remarkably constant. Html documents has remained around 50% throughout the period, with jpeg and gif pictures making up about 45%. These numbers has only varied by a few per cents during the period.

The web sites varies a lot in size, the most common size is one (1) document. The big web sites however, can have in excess of a million urls. These are universities and a few big web hotels. It should be noted however that web sites in excess of 100000 urls number only a few tens.

A true picture of Sweden?

The overall goal of the work is to acquire a collection of data which gives a true picture of the Swedish web space at the time of archiving. Has this been achived?

First we look at the "geographical" coverage. We of course harvest everything found under the Swedish top-level domain "se". But there is nothing to stop a Swedish company or person to register a domain under the international top domains "com", "org" and "net". Also, many countries allow (nearly) anybody to register a domain under their top domain. In Sweden the domain "nu" has become very popular because of its Swedish significance (*nu*, means *now* in Swedish). Special efforts has been made to identify domains registered under these top domains which can be considered Swedish in some respect. How succesful these efforts have been is hard to know. Also, there are certainly a lot of Swedish material located under other country codes. One serious problem here is the organisations registering domain names as a rule doesn't reveal the identity of the domain owner. We know that at least half of the Swedish domains are registered under non-"se" names. Nevertheless we think that our coverage is at present rather good.

Tuning to more technical problems; harvesting material consisting of static pages linked by standard html links is easy. However, an increasing number of other techiques are being used, javascript, flash etc. Scripting techniques, like javascript, cannot in principle be succesfully treated since the result of executing the script can depend on many things, e.g. type of browser. Flash is using a plug-in and is also a proprietary format, making any attempt at harvesting such pages very difficult.

Another difficult problem is webservers that tailors its pages for each individual user. Using e.g. cookies the webserver gives each user a unique page; trying to phantom the users interest etc. The problem is of course: which users web is to be saved?

A serious problem is all types of interactivity: games, databases access by searching on key words etc. Here the harvesting robots fall short. All this material is in practice lost.

To answer the question posed above; we think that our "geographical" coverage is good, as is the harvesting of pages with simple html-linked static pages. We have very little of the dynamic web, i.e. all sorts of sites where there is a dialogue between the user and the webserver, e.g. intaractive games. The static cases mentioned above is dominating the web and we don't miss an awful lot where numbers is concerned. It is rather that we miss a certain type of material more or less completely.

The Future

In the future we can expect more use of interactive pages, non-html techniques (scripts, plug-ins etc). The development of new web publishing techniques give very little, if any, thought to preservation. Also, we can expect voice to be used to navigate the web, needing a new type of harvesting software. There is also another, more serious threat. Most techniques used on the web utilize open standards published by IETF and W3C. It is the authors feeling that it is not necessarily in the interst of the major players to support the use of open standards and there is a real threat that the web will by monopolised by a few actors, using proprietary, closed standards, making this kind of work very difficult.

Conclusion

Despite all the difficulties mentioned above it is possible to get a reasonable snapshot of the web without too much difficulty. We miss a lot of e.g. interactive pages that's true, but that doesn't make what has been saved worthless. There are many aspects of human activities which are lost forever, but doesn't make the medieval manuscripts useless when we try to understand what was going on back then.