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

Code Number: 090-163-E
Division Number: I
Professional Group: National Libraries
Joint Meeting with: Information Technology
Meeting Number: 163
Simultaneous Interpretation: -

Access to web archives: the Nordic Web Archive Access Project

Svein Arne Brygfjeld
National Library of Norway
Oslo, Norway

Abstract:

The national libraries of the five Nordic countries have carried out a project to find principles, methods and tools for access to archives of web documents. This project has resulted in a prototype of an access system where the user can browse, navigate and search in time and space. The access system is independent of the archive, and may easily be adapted to new running environments. Search engine technology and java-based user interfaces are essential to be able to give the wanted functionality to the user.

1. Introduction

One of the challenges for the international information preservation community is the archiving and long-term preservation of documents published on the World Wide Web (WWW). Related to the area of long-term preservation, we also find areas like harvesting the web and accessing web archives. The national libraries of the five Nordic countries (Denmark, Finland, Iceland, Norway and Sweden) are all highly engaged in finding solutions on harvesting, archiving and accessing archives. These libraries have joined forces to investigate technology and methods on these areas, an initiative named “the Nordic Web Archive”. For the last 18 months, most of the effort has been spent on finding ways of accessing web

archives. Nordunet2 has supported this work, making it possible to run a focused project called “The Nordic Web Archive access project” (NWA).

2. Aims

The core aim of the NWA project is to implement an access system for large-scale web archives. This system shall support well known access methods like search, navigation and browsing. In addition, one wants it to be possible to browse and navigate through space and time. The NWA project is based on the assumption that there are several archives, and that each archive holds possibly several versions of a significant number of web documents. Such archives are likely to be constructed in various ways, and the design and implementation of the access system should be independent of the internal structure and architecture of archives.

3. Architectural choices

Starting the work on system architecture, one wanted to split the system into significant modules easy to identify an limit. In this way, one could enable distributed collaborative development work. The modules identified are listed below, and also shown in figure 1.

- Document input: The access system will receive XML-based documents from the archive for indexing. Those documents will also contain metadata on a Dublin Core format, foreseeing future support for OAI support. There are also some archival metadata available, like time of harvesting.
- Indexing: Organising the documents for searching
- Search engine: Component to support search in indexed documents and metadata
- Web interface: Component to support user interface based on using WWW
- Archive access: Component to support document delivery from the respective web archive

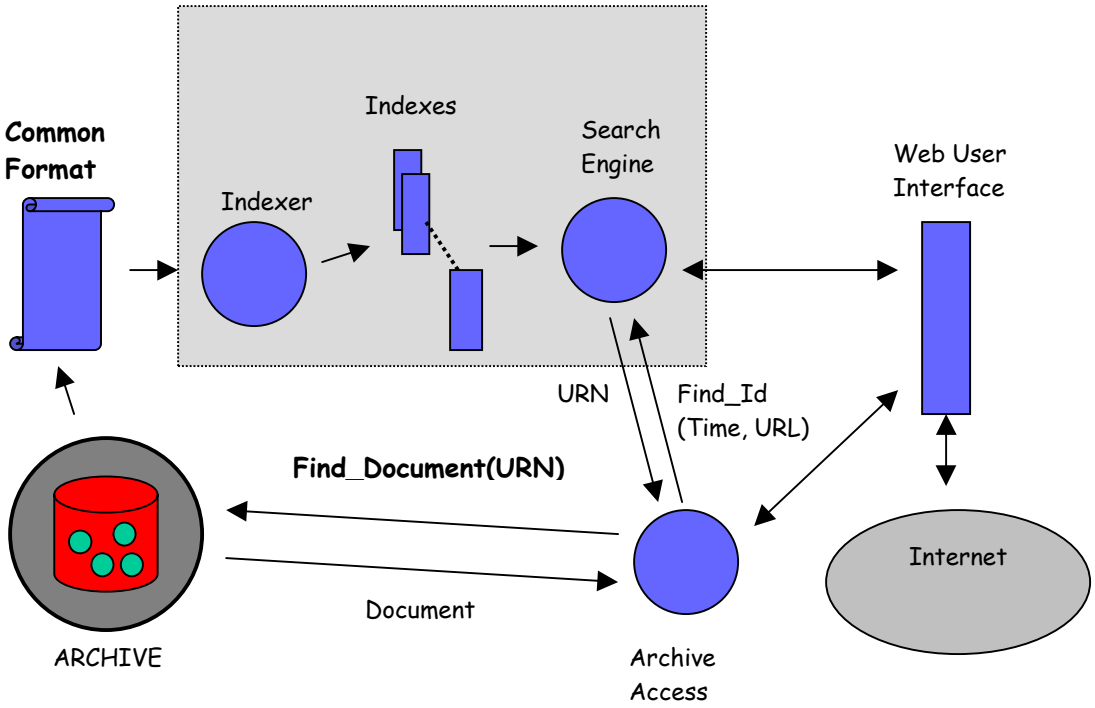

Figure 1.

4. Implementation strategies and solutions

First of all, it was wanted that the speed of the service should be as good as search engines available on the Internet. Thus, one realized it was a necessity to use a well-performing search engine to support search, navigation and browsing. Introducing such a software component has shown to simplify the development process significantly as well as giving the expected performance.
Furthermore, the implementation is highly modularised. Every module is relatively small, making it possible to do a re-implementation when needed. The approach also invites to modification of the functionality of the through introduction of new modules. Much effort has also been put into making a clean well define interface against the search engine. Porting to other search engines should therefore be also an overcoming task.

5. Archive interface

As pointed at already, access to web archives should be based on the same technology as access to the web. A web browser providing the same impression of the content as the original would be optimal.

6. Functionality

The user interface should support the well-known search, browse and navigate functions which users meet every day on the Internet. And in addition, it should also support those functions on several versions of documents, thus introducing the concept "search, browse and navigate through time and space".

6.1. Search

Search in the archive is based on the use of a dedicated search engine. The motivation to introduce a search engine, is to be able to offer a satisfying speed of operation for the user. Searching should be performed both on available metadata as well as the content of the documents. One of the challenges with respect to result presentation is the fact that one might get hits in several versions of many documents, expanding the already existing problem of large result sets.

6.2. Browse

As on the web, it should be possible for the user to browse the archive just by following links in the displayed documents. The user might also browse versions of a given document by use of the java-based time-axis shown in figure 2. Clicking on the arrow-heads will provide the previous or next version of the document from the archive.

6.3. Navigate

By the use of known locators (URL's), the user should be able to locate and navigate in the archive. So given a point of time, the archive should provide a document when a user types the original URL of the document. On the other hand, given an original URL, the archive should provide a document related to a point of time when the user gives a point of time. These functions are made available through the *locator* and *time* fields in figure 3.

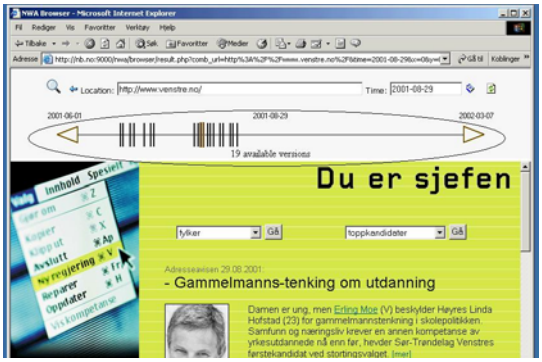

Figure 2.

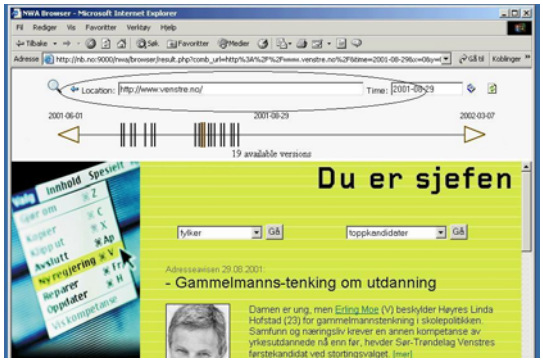

Figure 3.

7. Further reading and links

[1] The Royal Library of Denmark, http://www.kb.dk
[2] The University and National Library, Helsinki, Finland, http://
[3] The University and National Library of Island, http://
[4] The National Library of Norway, http://www.nb.no
[5] The Royal Library of Sweden, http://www.kb.se
[6] Nordic Web Archive, http://nwa.nb.no
[7] Nordunet2, http://www.nordunet2.org
[8] Kulturarw3, The Royal Library of Sweden, http://kulturarw.kb.se/html/kulturarw3.eng.html
[9] The Wayback Machine, http://www.archive.org
[10] Zeitschrift für Bibliothekswesen und Bibliographie, Issue 3/4 2001
[11] NEDLIB project, http://www.kb.nl/coop/nedlib
[12] OAIS *Reference Model for an Open Archival Information System* http://ssdoo.gsfc.nasa.gov/nost/isoas/
[13] Extensible Markup Language, http://www.w3.org

[14] Open Archives Initiative, http://www.openarchives.org
[15] Fast Search and Transfer, http://www.fast.no
[16] Open Source Initiative, http://www.opensource.org
[17] RLG DigiNews 2001, Volume 5, Number 2, http://www.rlg.org/preserv/diginews/diginews5-2.htm