Nordic Web Archive

Þorsteinn Hallgrímsson[1] and Sverre Bang[2]

[1] National and University library of Iceland
thh@bok.hi.is
[2] National Library of Norway
sverre.bang@nb.no

http://nwa.nb.no

Abstract. The National Libraries of Denmark, Finland, Iceland, Norway and Sweden are co-operating in developing a project called the Nordic Web Archive (NWA). The overall purpose of the NWA project is to develop software tools, named "the NWA toolset", for accessing historical Web Archive collections, including archive interoperability. The strategy for achieving this is to export the contents of the Web Archive to a common format, and index this output using a search engine, enabling fielded free-text search. A Web interface for search and navigation, including navigation between different versions of URL's will be developed. The toolset will be made available as Open Source software. The paper describes the background, the concepts and the technical implementation of the project.

Introduction

Before discussing the Nordic Web Archive it is important to explain the background for establishing this type of an Archive. In the vision statement that accompanied the Nordic Web Archive (NWA) project at its conception the following was stated:

"It is obvious that currently, and increasingly in the future, a large and significant part of our culture will exist on the Internet only. If the traditional axiom of the Legal Deposit laws and other collection activity hold true it is therefore an absolute necessity to extend this concept to the Web of the Internet. If this is not addressed now, an important part of our culture, together with most documentation of the cultural change involved, will be lost. It is therefore proposed that through the Nordic National Libraries' joint efforts on technology development, techniques and methods, the Web space of the Nordic countries shall be preserved for the future in an Archive to allow research and public access both today and for the generations to come. Access shall be based on technology wide spread among the users at the time of access."

For the Nordic National Libraries the biggest challenge in implementing this vision is on one hand how to collect and preserve for the future the electronic/digital publications and works, and on the other hand how to provide access to the collection. The Nordic National Libraries agree that it is virtually impossible to enforce a law that would require all who publish a work on the Internet to send a copy to the Legal De-

posit library. The only practical solution is to use Internet- or Web-harvesting and accordingly the legal deposit laws in each country must be changed allowing the national libraries to do this.

The process that led to the vision statement and the creation of the NWA has its origin in the Kulturarw[3][1] project that the Royal Library in Stockholm started in 1996. The NWA started as a Nordic forum for co-ordination and exchange of experience in the fields of harvesting and archiving Web documents. The Nordic countries realized that in addition to harvesting Web-documents and storing those in a Web Archive, it was paramount to provide software and tools that would enable developers, researchers and the general public to use the Web Archive and allow the developers to verify that the harvester retrieves the intended set of documents from the Internet. This information can be used both to improve the harvester software and to control the harvesting. Researchers and other users must be able to access the Web archive and retrieve the documents they need. For this reason and because there would be two harvesters available, the focus of the Nordic effort was shifted to access.

In August 2000 the Nordic National Libraries applied for, and received a grant from Nordunet2[2] to develop an Access Module to the Web Archive. The overall purpose of the NWA project was to develop tools for accessing historical Web archive collections, including archive interoperability to enable users to search and navigate the archived Web documents of all the Nordic countries. The strategy for achieving this was in short to export the contents of the Web archive to a common format, and index this output using a search engine, enabling fielded free-text search. The search engine would provide the possibility to develop Web interfaces for search and navigation, including navigation between different versions of URL's.

The project was started in November 2000 and finished in July 2002. This was a cooperative project with software developers at every Nordic National Library and a Project Manager at the National Library of Norway (NB). NB supplied the necessary infrastructure for the development. The outcome of the project was a software package named "**the NWA toolset**", a set of tools for searching and navigating archived web document collections. When nearing the end of NWA's formal project period the developers realised the need for continued development of the NWA Toolset. The reasons for this were some unsolved problems with the Toolset, input from end-users strongly indicated that additional functionality should be added to the Toolset and the developers wanted to make the Toolset available as Open Source software. It was also clear that it was far from stable and that the performance had to be substantially improved for large-scale implementation.

The Nordic National Libraries decided that a follow up project was needed and in late 2002 NORDINFO[3] approved funding of the project together with the libraries. The project is called NWAII and the primary goal of the project is to ensure that the NWA application software and tools will serve the Nordic National Libraries well in

---

[1] Kulturarw[3], http://www.kb.se/kw3/ENG/Default.htm, the Swedish Web Archive

[2] Nordunet2 , http://www.nordunet2.org, is a research programme financed by the Nordic Council of Ministers and by the Nordic Governments

[3] NORDINFO, http://www.nordinfo.helsinki.fi/index.htm, the Nordic Council for Scientific Information is the Nordic organization for co-operation within the Nordic research libraries.

giving researchers and scholars the access they need in order to work with the national Web archives. The following tasks were identified for accomplishing this:

- Complete unresolved tasks and correct outstanding problems with the toolset.
- Make it possible to integrate different indexing software with the NWA Toolset.
- Make the NWA Toolset available as Open Source Software. This includes making the NWA Toolset independent of the commercial software packages needed for using the current version of the NWA Toolset
- Implement additional functionality for researchers and other end-users. In addition tools for analysing the archive are needed.
- Re-evaluate the user interface by involving scholars and researchers.

The project started in March 2003 and is scheduled to finish in April 2004.

The Nordic Web Archive Toolset

Archiving the Internet for the future and providing access to it requires harvesting the Web pages in a country either periodically or continually. The Harvester delivers its output to a Web Archive and the Archival function requires a Storage module, Preservation routines for ensuring long-time preservation of the Archive, and an Access module for interaction with the Archive. Additional issues like Analytical Tools, Sustainability, Standards (metadata, architecture, etc.) and System Architecture must be considered. The NWA does not address all of those.

Making the NWA toolset available as Open Source Software will secure future development of it while the Nordic National Libraries as well as other parties can make use of the software and contribute to future development of it.

Two **Web-Harvesters** were available and in use in the Nordic countries and although neither of those applications is considered optimal for use in the NWA project in its present form they satisfy the requirements in the first phase of the project.

The **Archive** is the central module in the NWA. As a result of past digitisation projects, collecting of electronic documents and harvesting of Web pages, digital archives have been created in all the Nordic countries, each with its own organisation. Harvested documents and metadata related to harvesting are passed into the Archives. An Archive may contain several full generations of the Web space, or it may be incremental. In the latter case, the Archive does in principle not contain any duplicates.

The amount of storage needed to hold the Archive of each country is different. Sweden will need 15 to 20. Although the Web is growing very fast, storage technology and IT in general are developing at equal pace.

The Long Time Preservation and **Sustainability** of the NWA (operation, future development and maintenance) will be the responsibility of each Nordic country.

Access to the Web Archive is as already mentioned paramount for researchers and developers. Thus there is a need to build a separate access module based on efficient software and tools that are capable of indexing a very large number of full text documents. An index based on a Web archive will be cumulative; that is, depending on the harvesting policy it will contain every document that exists in the Web space now,

and has existed there since the harvesting begun. In records built by the indexing software there will be URL links to the archived documents. The chosen indexing tool has to be able to deal with a Web Archive. As a part of the NWA project an existing indexing tool (from FAST Search & Transfer ASA[4]) was chosen and configured to cope with both archived Web documents and archival related metadata. A specific user interface must be developed in order to introduce the element of different time periods and to assure that URLs pointing to the Archive link to the archived documents.

Analytical tools are needed in order to analyse the Web Archive in order to understand the nature and development of the Internet, for research and for better controlling and scheduling the harvesting of the Internet. This had to be left outside the project.

Of the above-mentioned elements the NWA primarily addresses the System Architecture, the Archival function and Access to the Web Archive.

System Architecture is the most critical part and the design criteria for the system is to make the NWA as implementation independent and modular as possible. This is a must in order to implement the NWA in many libraries, each with its own information technology infrastructure. Therefore harvesting the Internet, storing the harvested data and creating and maintaining the Web Archive is the responsibility of each library. The NWA consists of the NWA toolset and additional components, each with a clear interface. The toolset includes a Document Retriever, an Exporter and an Access Module. The additional components are a Search Engine and a Search Engine Abstraction Layer. A necessary prerequisite is a Web Archive, i.e. a historical collection of Web documents, stored along with metadata. A key requirement for the archive is that the objects are stored unaltered and that a set of metadata consisting of at least the original URL and timestamp of the objects is available.

The Document Retriever serves as the interface to the Web Archive. It delivers archive objects and associated metadata from the Web Archive to the Exporter and the Access Module upon request and therefore it must be adapted to the information technology infrastructure at the library where it is installed.

The Exporter fetches archived objects and associated metadata from the Web Archive via the Document Retriever and prepares them for indexing. The input to the Exporter is a list of archive URI's defining which archive objects the Exporter should process. The list will have to be generated at the archive side as a preparation for export. Automated tasks for creating the list will have to be tailored to suit the specific web archive architecture. The following takes place during export (see figure 1)

1. Fetch metadata for the given archive object from the Document Retriever.
2. If the format (content-type) of the archive object is not html or a convertible format output metadata and go back to 1 for next id in list.
3. Fetch the given archive object from the Document Retriever.
4. If object is html go on, else convert the object to html.
5. Extract relevant data from the html content.
6. Determine the language used in the textual contents of the object.

[4] FAST Search & Transfer ASA, http://www.fastsearch.com provides Internet search solutions

7. Output metadata received from the Document Retriever along with data extracted from the html content and the language.
8. Back to 1 for next id in list

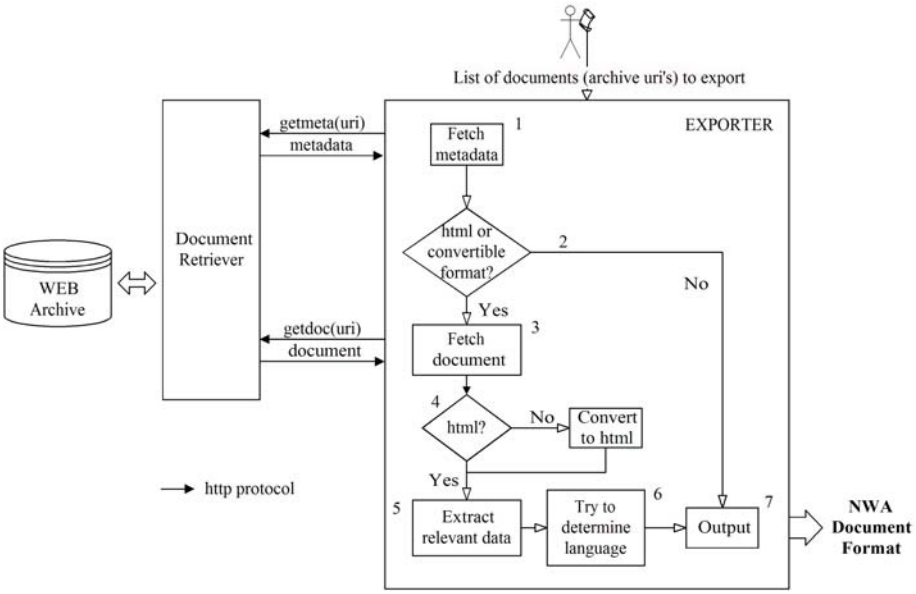

Fig. 1. Exporting Web archive objects to the NWA Document Format

Comments to some of the steps above:

1. Below is shown a typical metadata record for a given archive URI. The URL element contains the original URL of the object and the time element contains e.g. the time of harvest. Choosing the timestamp to use is a matter of Document Retriever design. If the archive holds any other timestamp (e.g. last modified) and this is considered trustworthy, the Document Retriever may be set up to return this timestamp instead.

```
<?xml version="1.0" ?>
<metadata>
  url><![CDATA[http://www.nb.no:80/katkom/frbr/4nbmkap1.htm]]></url>
  <time>20010808082501</time>
  <http-header><![CDATA[
    HTTP/1.1 200 OK
    Date: Thu, 05 Sep 2002 07:35:53 GMT
    Server: Apache/1.3.26 (Unix) PHP/4.1.2
    Last-Modified: Wed, 08 Aug 2001 08:25:01 GMT
    ETag: "17fb4-b38a-3b70f75d"
    Accept-Ranges: bytes
    Content-Length: 45962
    Connection: close
    Content-Type: text/html]]>
</http-header>

</metadata>
```

2. The Exporter keeps a list of the formats (mime-types) that are not considered convertible. E.g. a gif-image is not likely to give any useful textual information after being fed through the to-html converter so the only data exported for such an object will be the available metadata like its archive id, the original URL of the object, its mime type and its timestamp.

4/6. The to-html converter transforms non-html text objects like pdf, msword etc. into html, thus enabling extraction of data from these objects as well. The Language detection makes it possible to identify the language used in the archived object and enables the user to narrow a search to text-content objects written in a specific language. The tools used for to-html conversion and language detection in the Nordic National Libraries are third party products licensed by FAST Search & Transfer ASA. The NWAII project will however implement support for Open Source tools for this. The functionality, the number of languages supported and the number of formats supported by these are however likely to be limited compared to the commercial versions.

7. The output from the Exporter is stored in the NWA Document Format (XML). All the data in the elements on the level below the document element (see example below) will end up in a corresponding searchable field in the index. Note that the NWA Document Format schema contains additional elements as well. Only the elements that contain data are written to the output.

```
<?xml version="1.0" encoding="UTF-8" ?>

<nwaDocumentCollection xmlns="http://www.nb.no/nwa/export/1.0/"
xmlns:dc="http://nwaDocumentCollection.xsd">
<document>
  <collection>no</collection>
  <archiveidentifier>/var/hepp/2001-9-
25_1_539210b1c4b0ce5e9b9fd76d296b91ba</archiveidentifier>
  <multimedialinks>
    <link>http://www.su.no:80/oslo/sulogo.gif</link>
  </multimedialinks>
  <links>
    <link>http://www.su.no:80/oslo/default.htm</link>
    <link>http://www.su.no:80/oslo/aktivist.htm</link>
  </links>
  <teaser>Oslo SU støtter Streik på Aker</teaser>
  <dc:title>Streik på D/S Louise</dc:title>
  <dc:format>text/html</dc:format>
  <dc:identifier>http://www.su.no/oslo/dslouise.htm</dc:identifier>
  <dc:date>20010924133519</dc:date>
  <text><![CDATA[Oslo SU støtter Streik på Aker Brygge Hei alle dere
der ute Selv i disse krigs og postvalgstider går klassekampen videre.
Som mange av dere kanskje vet er det for tida streik på Aker Brygge.
Arbeidera på D/S Louise streiker. Dette har'em holdt på med i tre
uker nå, og dem b'ynner å bli slitne.]]>
  </text>
</document>
..
<document>
.
</document>
</nwaDocumentCollection>
```

The Access Module provides the user with interfaces for searching, browsing and navigating the archived Web pages. The current interfaces are not in any way final; they will undergo a redesign during the ongoing NWAII project. Figure 2 shows a

simple search interface used to search a very limited collection of Web objects harvested from the National Library of Norway's Web site.

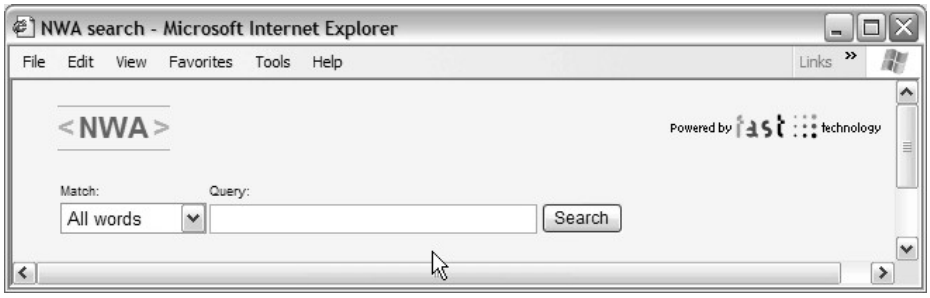

Fig. 2. Simple search interface

When the user submits a query the Access Module uses the search engine to find the objects containing the text(s) satisfying the query. The results from the query are presented in a URL oriented way as shown in figure 3. The user may point and click the links in the result list to view a specific version of a Web page.

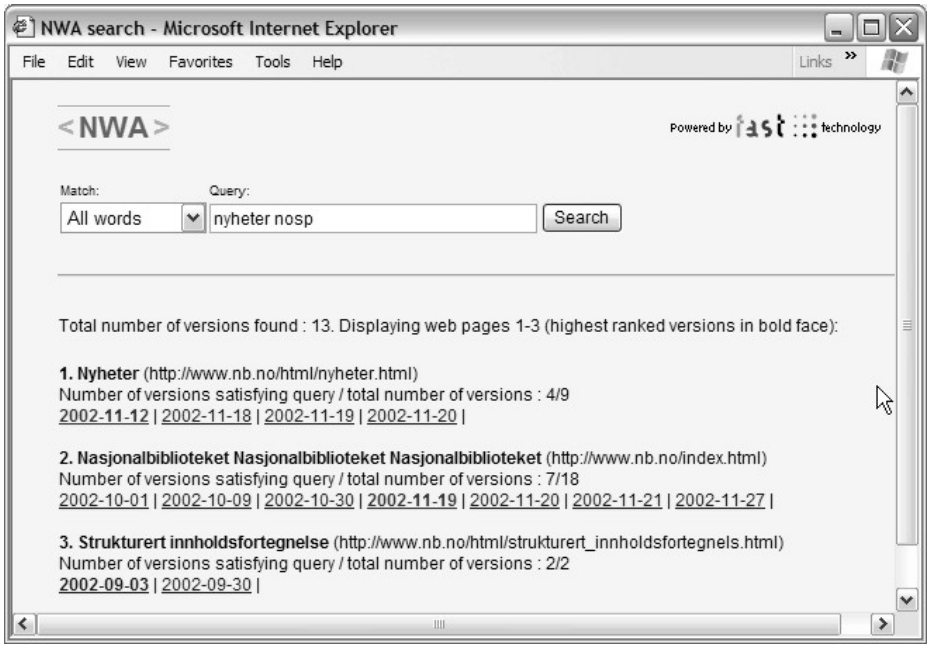

Fig. 3. Search result list

In the example given in figure 3, the query text was found in a total of 13 objects. Only 3 of those were distinct URL's. If we look at the first hit (URL) we see that there were 4 versions of this specific URL that contained the text queried for. There were a total of 9 objects in the archive with this URL, which means that 5 of the objects with that specific URL did not contain the query text submitted.

Figure 4 shows a simplified overview of what happens when a user asks for a specific Web page from the Access Module. The Access Module will retrieve the object from the archive. Before the object is delivered to the user's browser the object is parsed and all the inline links and references are altered to point back to the Access Module rather than out to the www. When the browser encounters references to inline objects, the browser will ask the Access Module to return those objects as well in order to present them as part of the Web page.

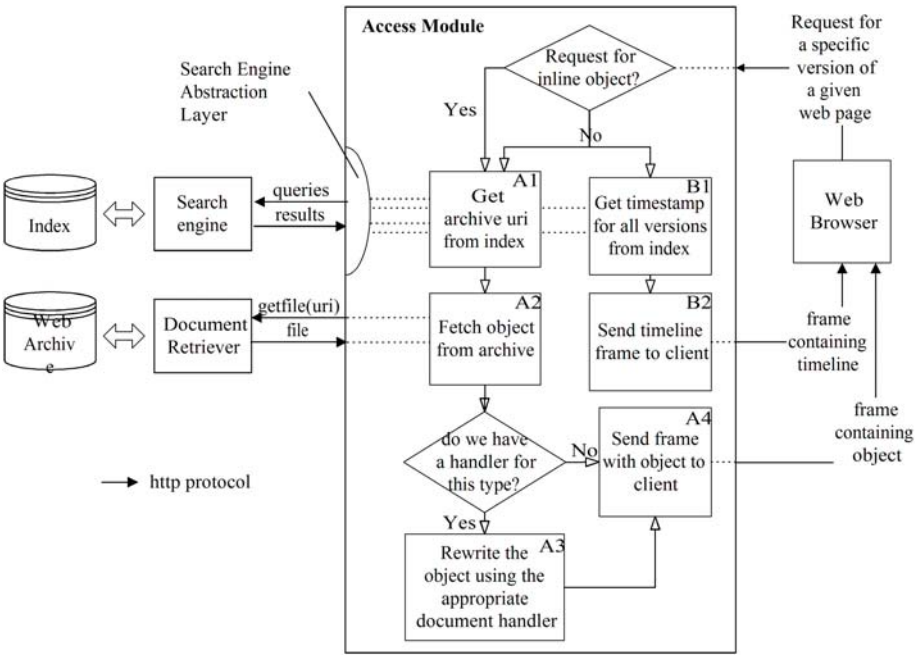

Fig. 4. Access Module

When the browser requests a specific version of a given object, two separate processes may commence (depending on whether it is an inline object or not):

A. Prepare object
1. The requested objects archive URI, and mime-type are fetched from the index.
2. The object with the archive URI mentioned in 1 is fetched from the archive.
3. If there exists a specific document handler for the mime-type of the fetched object the object is passed through this document handler. The document handler will rewrite the objects contents so that its inline references and links points to the access module rather than to the Internet.
4. The object is sent to the client.

B. Prepare timeline (not executed if the object is inline)
1. The Access Module queries the index for all the different versions of the given document.
2. The timeline is prepared and sent to the client.

The resulting Web page will contain a timeline at the top and the archive object(s) below. An example is shown in figure 5. The resolution of the timeline may be set manually or one can use the default setting, auto, which makes the timeline drill down to the appropriate resolution for display. The user may point and click on a point in the timeline to view how the Web page looked like at a specific point in time. The user may also point and click the arrows below the timeline to navigate to the next, last, previous and last version of the Web page. Below is shown the timeline and a specific version of a Web page.

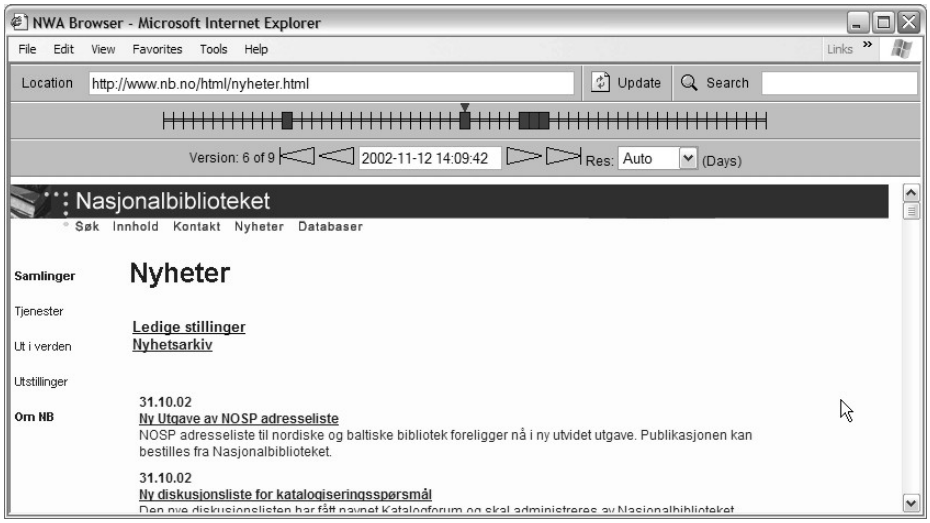

Fig. 5. Navigating between different versions of a Web page

If the user points and clicks one of the links in the displayed Web page, the browser will request the new page, including a timeline from the Access module. Choosing the link “Ledige stillinger” in the Web page above produces the output shown in figure 6.

Fig. 6. Navigating between different versions of a Web page

Pan Nordic Access

The FAST Search Engine provides the NWA cooperation with a scalable and distributed high-performance architecture for search. This enables us to use the Access Module to search and navigate all the Web Archives of the Nordic National Libraries as one Web Archive and figure 7 shows an overview the Access Module interacting with the Pan Nordic Index and Web Archive.

A **Search Node** is a node that holds and searches a part of the index. When starting to index a Web archive one would start with one search node and keep on indexing until the node is "full". A new search node is set up to take over indexing from where the first search node ended.

A Dispatch Node serves as a front-end to the search nodes. The dispatch node receives a query from a client application and sends the query to all the search nodes. Each search node will execute the query on their respective indexes and return the results to the dispatch node. The dispatch node will merge the results and pass the merged result on to the client application.

In order to query all the Nordic Web Archives simultaneously, an extra dispatch node is needed, acting as a front-end to the national dispatch nodes. Imagine a user entering the Access Module (through a Web browser) at the National Library of Iceland and submits a query (see figure 7). The Icelandic Access Module sends the query to the front-end dispatch node (which may reside locally in Iceland or at one of the other National Libraries). The front-end dispatch node queries all the Nordic indexes, merges the results and sends the merged result back to the Icelandic Access Module. The Access Module formats the result and sends it to the browser.

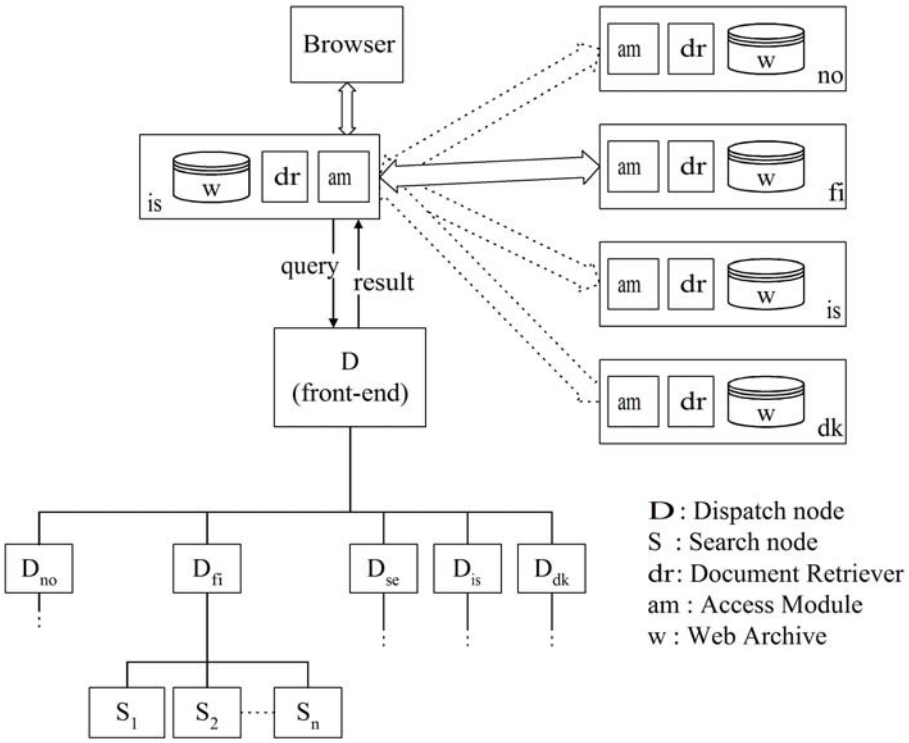

Fig. 7. Pan Nordic Access

If the user now requests a given Web page, the Access Module will query the front-end dispatcher for the object's archive URI as well as information on where the object is stored (the element *collection* in the NWA Document Format). If we assume that the requested object is located in the Finnish Web Archive the Icelandic Access Module will ask the Finnish Access Module for the object. The Finnish Access Module will retrieve the object from it's Web Archive and parse the object, changing all links and references to point to the **Icelandic** Access Module rather than out to the internet. The Icelandic Access Module receives the object and passes it through to the browser. Preparing the timeline for the Web page is also done at the Icelandic Access Module. Of course, if the object had been located locally in the Icelandic archive the parsing and the exchanging of links would have taken place locally as well.

Conclusion

Between December 12, 2002 and January 20, 2003 a survey on Web Archiving in Europe was conducted at the National Libraries of Europe. Altogether 39 countries received the questionnaire and 25 replied, or 65%, including all the countries that are known to be working on Web Archiving. According to the survey 15 countries have started a Web Archiving activity or are seriously testing the activity, and 5 can base

this on their Legal Deposit Law (Denmark, Iceland, Lithuania, Norway and Sweden). Of those countries that have seriously started Web Archiving, Sweden has the longest experience and by far the largest amount of data 5,5 Terabyte.

The pioneer in Web Archiving is the Internet Archive of San Francisco, and it has collected the largest Archive of Web documents (about 200 Tb) by harvesting all over the globe. Outside Europe Australia has for some years been active in Web Archiving by selecting certain Web sites for collection and preservation and recently the Library of Congress has followed suit.

All those projects and the related work represent the first phase in the task of preserving the intellectual contents of works (documents) that are available (published) on the Internet. The developers of the NWA and many other institutions are keenly aware of this and progress is being made towards international cooperation in defining and developing both the necessary technology components and in establishing procedures and standards that will enable the building of national and even global Web Archives with the necessary access.

From the point of view of collecting and preserving the Internet our understanding of the medium and its contents leaves much to be desired. Currently access (navigation, indexing, searching) to the documents in the Web Archives is secured for research purposes, but general access to the documents depends on the Copyright and Legal Deposit legislation in each country and is often very limited. However, information of what the National Libraries have collected into the Web-archive will be freely available. This must change in the future.

The Web Archives contain multimedia documents and currently it is only possible to index text and this can be enhanced using linguistic methods. For other types of documents like for instance images, indexing is limited to harvester generated metadata plus textual information in the file header. However a lot of research is being done with the purpose to use computer programming to index sound, images and video/movies, and when this becomes practical these parts of the Web-Archives can be indexed.

Every current tool needs to be improved and new tools must be developed. Hopefully increased awareness and activity in Web Archiving coupled with advances in technology and international cooperation will in the near future enable the National Libraries of the world to collect and preserve the Internet like the printed collections of today.

The National Libraries should consider Web-archiving and indexing as a tremendous, albeit challenging, opportunity instead of looking at it as a problem or liability.