To: nordunet2@usit.uio.no

File number: "00/1754"

NORDIC WEB ARCHIVE

Introduction

This is an application to Nordunet2 in the key-area of **Digital libraries** by the Nordic National Libraries represented by the national directors.[1]

The group of 5 national directors is the legal body responsible for the project . They will be represented by Erland Kolding Nielsen, Director of The Royal Library in Copenhagen who also will head the administration of the project.

The project for which funds are applied is entitled the **NORDIC WEB ARCHIVE**. In this project the Nordic National Libraries will through joint efforts on technology development, techniques and methods, preserve for the future, the Web space of the Nordic countries 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.

Funds are applied for the development of one of the modules, the Access module, for this project. Nordic national libraries already have developed tools for harvesting and storing the documents. Further development of harvester robot application will be financed by the national libraries.

Principal investigators are: Allan Arvidson at the Royal Library in Stockholm, Birgit Henriksen at the Royal Library in Copenhagen, Petri Heliniemi at the Helsinki University Library of Finland, Svein Arne Brygfjeld at the National Library of Norway and Þorsteinn Hallgrímsson at the National and University Library of Iceland,.

It is anticipated that total direct costs for producing the Access Module for the NORDIC WEB ARCHIVE will be in the neighbourhood of DKK 2.000.000. The applicants seek DKK 1.400.000 from Nordunet2 to be used mainly for personnel costs. The remainder will be provided by The Nordic National Libraries.

[1] The Royal Library in Copenhagen; Director Erland Kolding Nielsen, The Helsinki University Library; Director Esko Häkli, The National and University Library of Iceland; Director Einar Sigurdsson, The National Library of Norway; Director Bendik Rugås and The Royal Library in Stockholm; Director Thomas Lidman

Table of Contents

NWA Vision Statement

Most countries strive to preserve and provide access to their cultural and intellectual heritage by collecting it and storing in museums, archives and libraries. In the Nordic countries the national libraries have traditionally fulfilled this role by collecting manuscripts and published printed material. As publishing technology has progressed the libraries extended their collection activity by including physical electronic media like CD-ROM'S and some electronic publications like electronic journals.

The main instrument enabling the national libraries to do this collection activity economically and comprehensively has been the Legal Deposit law of each country, and the purpose is to preserve and provide access to the intellectual and cultural heritage of a nation. Until now this has served the libraries very well. The emergence of the Internet with its widespread connectivity and the Web interface has changed this.

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.

In the Nordic countries the legal deposit law has either been changed or will be changed to reflect this, thereby enabling the collection of material published on the Web. Access to this material will depend on the copyright laws in each country. Collecting Web-documents selectively is difficult and costly and this material would cover only a fraction of what is published on the Internet. The best solution is to collect all Web-documents of each nation, store them in an Archive and preserve it for the future.

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.

Access to the documents in the Archive will be secured for research purposes, but access to the machine generated databases containing information about archived documents will as a rule be free. The users who access the databases from non-privileged work stations will not see URLs which link the bibliographic information into the archived document. However, information of what the national libraries have collected into the Web archives will be freely available.

Project Background

The Nordic countries have for a long time co-operated on solutions to enhance access to the Web. The most visible of those is the NORDINFO-project, NWI (Nordic Web

Index), which was developed by NetLab in Lund. In the following project, Nordic Web Index II, NetLab worked with DTV.

NetLab developed an application which enabled harvesting of documents. The harvester, called Combine, is given a number of starting pages. These documents are parsed, and URL links found are fed into an input file. The contents of this file are checked against the harvesting parameters and if so, passed on to the harvester module, which attempts to retrieve the new documents. These documents are again parsed for new URLs. This goes on until every document that fits the given parameters (e.g. *.*.no) have been found.

In both NWI projects access software was Zebra, developed by Danish company Index Data. Zebra was used to index text data from harvested documents into national Web indexes. These were then united into a Nordic virtual union database via intelligent usage of Z39.50 Information retrieval protocol, which is supported by Zebra. End user access was based on Web browsers, which access the database via HTTP - Z39.50 gateway, also developed by Index Data.

In 1996 The Royal Library in Stockholm started a project called Kulturarw^3 (http://kulturarw3.kb.se/), with a view to the long-term preservation of electronic documents. The aim of this 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. Through this project KB in Stockholm is also laying the foundations of a collection of Swedish electronic publishing for our time and for coming generations.

Kulturarw^3 worked together with NetLab to modify the Combine software so that it is suitable for archiving purposes. This is no trivial task. The main difference between "normal" and "archive" harvester is that a normal harvester does not keep the documents once they have been indexed by the access module. A harvester built for archival purposes has to store all retrieved documents either into a database or into a file system. The archive harvester must also keep track on what it has already stored via for instance calculating MD5 checksums and storing them into a metadata database. This database must also contain information about when the document was retrieved.

KB in Stockholm has harvested the Swedish Web space six times since 1997, using the modified version of Combine. In Finland, a different Combine variant was developed and the Finnish Web space was harvested in Autumn 1998.

In 1998 an EU project called NEDLIB (Networked European Deposit Library) (http://www.kb.nl/nedlib) was started. The national libraries of Finland and Norway participate in this initiative, alongside national libraries of France, Germany, Italy, the Netherlands and Portugal. In this project one of the sub tasks is to collect and archive Web documents in the same manner as in the Kulturarw^3 project.

The project tested a number of existing harvesters including Combine, but decided in the end to develop a harvester dedicated for archival purposes. In many respects this application is quite different from a normal harvester. For instance, harvesting policy has been optimised so that all inline images are retrieved as soon as possible to guarantee integrity of harvested documents. Also, parameters that limit harvesting do

not apply for inline materials; the harvester run by the national library of Finland will retrieve an image from Iceland if it is used as an inline image in Finnish HTML page.

The Nedlib harvester module has been used in Finland to harvest large parts of the Finnish Web space. It has been tested by all NEDLIB libraries and the national library of Estonia, with good results. Development of the NEDLIB harvester software continues in Finland as a part of a domestic project EVA III, funded by the Ministry of Education. All of the Finnish Web space will be harvested by the application in Autumn 2000. The national library of the Netherlands plans to do the same in the Netherlands.

The Royal Library in Stockholm was active in starting the NWA co-operation in 1997. NWA started as a forum for co-ordination and exchange of experience between the different national project that were in place or about to start. The project group has met regularly and the work progressed to specifying and defining the requirements specifications for a Nordic project along the lines of the Kulturarw^3 project (see Appendix II), culminating in this application to Nordunet2. In October 1998 the Nordic National Libraries decided to take over the responsibility for the development of NWA by establishing a steering group and to fund the work of a project co-ordinator when this was deemed appropriate. His main tasks were to:

- Develop with the steering group the functional and technical requirements specifications for archival and access .
- Co-ordinating development work within the national projects
- Co-ordinating the Nordic efforts in an eventual European co-operation.

This position has not been filled yet but this will be done during late 2000. If this project proposal is successful, this person will also function as the project leader.

Present status of the NWA Project

The purpose of Nordic co-operation in this project has been to co-ordinate and share experiences. This means that all development work must take into account and build on what already exists within the Nordic countries and allow for local variations.

The NWA basically consists of two major functions, i.e. **Harvesting and Archival**. In the harvesting function the Web pages in a country e.g. in Sweden (**.se** and Swedish servers from other domains such as **.com**) are harvested until “all” pages have been retrieved. The Harvester delivers its output to the Archive. The Archival function requires a Storage module, an Access module for interaction with the Archive and Preservation routines for ensuring long-time preservation of the Archive.

Harvesting

As already mentioned there are two operational harvester modules used by respectively Sweden (Combine) and Finland (Nedlib harvester). In the KB in Stockholm there are two full-time employees involved with this work and in the Helsinki University Library one. Denmark and Norway are testing both modules and have plans to start Web archiving in the future.

Neither of these applications is considered optimal for use in the NWA project in its present form. Therefore the project group has defined a set of functional specifications for an NWA Harvesting module (see Appendix II) that will serve either as guidelines for changes and improvements to the Combine and Nedlib applications, or the development of a new harvester.

Harvester development will be continued with the staff and funding the national libraries already have.

Archival

The central module in the NWA is the Archive. As a result of past digitisation projects, collecting of electronic documents and harvesting of Web pages, digital Archives have been created in in all the Nordic countries. The Harvester modules must deliver the input to those in the proper format for each.

Harvested documents and metadata related to harvesting are passed into an archive. Metadata may be embedded into the documents or collected into the harvester database, which directs the operation of the harvester.

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. When the archival module of the harvester is storing the document, it calculates MD5 checksum of the incoming document and discards the resource if there is already a document with similar MD5 in the archive. New URL for the resource is added to the metadata database. Combine requires the user to harvest everything every time, while Nedlib harvester also supports incremental harvesting.

Both harvesters store time stamp each time a document is harvester and archived. A combination of MD5 and time stamp information in the Web archive of course enables usage of the archive for authentication purposes. It is possible to prove that a resource did exist in the Web at a certain point of time in a certain form. This functionality has potentially very high value for solving claims about who published a scientific finding or some other useful resource first. It is also technically possible to use the archive for checking scientific frauds or copyright infringements.

It will never be possible to deal with all kinds of documents or all protocols properly. For instance image maps will not work, and dynamically created documents are likewise an impossible problem. However, large portion of the Web can be catered for with tools already in existence. The fact that not everything can be done must not be used as an excuse for doing nothing.

On protocol level, the plan is to deal with HTTP and FTP only. There has been quite a lot of discussions about NNTP, but it turned out that implementing news gathering is technically challenging; a much easier solution is to set up a NNTP server in the national library for collecting the data. This has in fact already been done in Norway.

Access

Metadata database generated during harvesting is not suitable for end user access for two reasons. First, the main task of this database is to steer harvesting process, and serving the needs of human end users is a very different job. Second, relational databases such as MySQL used in Nedlib harvester do not yet cope well with full text indexing. Thus there is a need to build a separate access module based on an efficient text engine, a tool capable of indexing efficiently full text.

Normal Web indexes are not cumulative; when Web documents disappear they are also removed from the index. Of course there may be a very long delay before the index is updated since harvesters trying to gather every Web document can not visit every site too often. In Nordic countries it is still possible and feasible to check every Web document quite frequently.

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. If incremental harvesting is applied, harvesting can be a continuous process; then the Web archive and Web index may in principle be one and the same service. In records built by the indexing software there will be URL links to the archived documents.

The Access module can be largely based on existing applications such as Excalibur, FAST or Z'mbol, a modernised version of Zebra. This project will not build an indexing tool from scratch, but will co-operate with a chosen company to modify its tool.

The chosen indexing tool has to be able to deal with an Archive built by either the Combine, the Nedlib harvester or any other future harvester used for Web archiving. As a part of this project, one existing indexing tool will be chosen and enhanced in such a way that it can cope with both archived Web documents and archival related metadata.

In addition to indexing there is a need to modify user interface as well. For instance, users must be able to limit their searches by time; for instance, a user must be able to require homepages of the Finnish parliament from 2002 to 2010. Likewise it must be possible to use identifier (MD5) for searching.

There is also a need to build a proxy that enables linking of the URLs pointing to the Archive to the archived documents. The KB in Stockholm has already created this kind of tool, and it can be used as a basis for development work in this initiative.

Storage

The amount of storage needed to hold the Archive of each county is different. Sweden e.g. currently needs > 400 GB for one complete harvest. Finland will need approximately 200 GB and Iceland needs about 10 - 12 GB. The size of the index will depend on how comprehensively the text documents are indexed. Nevertheless, size of the databases will be relatively small compared with what the archived documents

need. As an aside, harvesting can be done with a work station, provided that it has enough work space for storing documents harvested during the day, and the metadata database.

Depending on the funding and technical requirements, Web documents can be stored either on traditional disk, into a hierarchical file system or on tape robot. The important factor is that the amount of data in the Web is definitely manageable with current technology. Moreover, although the Web is growing very fast, storage technology and IT in general are developing with equal pace.

The storage needs also depend on harvesting policy. Incremental harvesting requires much less disk than harvesting everything repeatedly.

National libraries will use their own funding to acquire hardware needed for creating the Web archive.

Long Time Preservation

The long time preservation of the NWA will be the responsibility of each Nordic country. This is a real challenge to everyone who is storing digital data and the preservation of the NWA will have to form an integral part of the policy each country will select for ensuring the viability of its National Digital Library. This policy must take into account both the security of the data itself and the metadata that must accompany that data.

According to the statistics generated during harvesting, more than 97 % of Web documents belong to handful of formats (HTML, JPEG, GIF). This means that although there is a huge amount of documents, it will be easy to preserve most of them. There will be problems (software applications, weird formats) but these will never constitute more than a fraction of the archive.

NWA Nordunet2 Project Description

The project partners have chosen differing tools for harvesting Web-pages and establishing National Digital Archives, resulting in different storage organisations. Because of this and combined with the need to provide flexibility in Harvesting, it was decided that each country should be responsible for the modules needed for Harvesting and Archival of its Web- pages.

The application to Nordunet2 therefore applies to defining and implementing the Access module and the interface between the Archive and the Access module. The goal can be stated as follows:

The Access module for the NWA (Nordic Web Archive) will enable the users to search and retrieve documents in the Archive. The Access module will be implemented based on the functional requirements specified by the NWA steering group (see Appendix I) and on a standard interface between the Access module and the individual Archives.

The development of sophisticated and effective software for accessing the NWA is complex. Commercial vendors like FAST, Alta Vista, Excalibur and others who provide indexing and search applications for the Internet use huge resources in manpower, technology and funds to provide effective service that is updated in synchronisation with the development of the Internet. Therefore it is important to seek alliances and co-operation with commercial vendors in developing a suitable access module for the NWA. The project group has approached two vendors, FAST (http://www.fast.no/fast.php3) and Index Data (http://www.indexdata.dk/), both of whom have shown interest in co-operation. Alternatively the use of indexing and retrieval software such as EXCALIBUR (http://www.excalib.com/products/index.shtml) should be evaluated.

Even if the laws of copyright in the Nordic countries will perhaps in the beginning restrict access to the archived documents it is important to define a full set of the functional specifications for access to, and retrieval from the Archive (see Appendix I). The main activities within the project are:

1) The Project Group will initially study and evaluate existing (commercial and non-commercial) indexing software and components and decide what software to use as a base for development trying to achieve the highest compliance to the functional specifications in Appendix I. If a commercial vendor is chosen the terms and conditions for use will be an important consideration. This activity must conclude with in a decision on what the development platform for the Access module shall be.

2) Because the Archives will be different in organisation and structure it is necessary to define a standard interface between the Access module and the Archives based on the development platform selected in 1).

3) The Access module containing at least the following functions, that are specified in more detail Appendix I, must be developed, installed and tested:

- Indexing.
- Search.
- Access.
- Access Control.
- Web User Interface.

Every effort should be made to use already available software. However because of the differences between the Archive and normal Web indexing, some modifications will be needed.

4) The result of the work done in 2) and 3) should be a software package that can be installed and demonstrated using one of the existing Archives.

It is important to note that when the NWA has been implemented, its operation, future development and maintenance will be the responsibility of each national library. The Nordic National Libraries are aware of this and accept that the preservation of the Web Archive whether national or Nordic is an expensive long time commitment and funds must be secured for the maintenance and continuous development of the Archive, ensuring that it fulfils its function in concordance to developments in Internet technology and functionality.

Plan of Work

It is estimated that the project will start in September 2000 and be finished at year end 2001. The following time schedule and milestones are planned:

1/9 2000 Project start
1/1 2001 End phase 1: Evaluating available solutions (commercial and others) plus definition of the Standard Interface to the Archive.
1/4 2001 End phase 2: Design of the Access Module.
1/10 2001 End phase 3: Implementation of Access and the Standard Interface.
1/12 2001 End phase 4: Test of Access and Standard Interface against an existing Archive.
31/12 2001 Project end

The 4 phases cannot be in parallel.

Relevance to Nordunet2

1. Nordic Dimension of the Project.

The NWA will establish in each Nordic country an archive of Web pages accumulated through time. It will be implemented in each country using its native language for the interface. It will be possible to create a Nordic union catalogue from national databases.

The tools developed as part of the NWA initiative will be used in other national libraries in Europe and possibly also elsewhere. This will enable the Nordic National Libraries to co-operate more closely with their peers abroad.

2. Connection to publicly supported research activities.

Results of publicly funded research are increasingly being published in electronic form and made available in the Web. The project will guarantee permanent availability of research results published and made available in the Internet.

All the applicants are public research libraries of great importance. The KB in Copenhagen, the Helsinki University Library and the National and University Library of Iceland are university libraries as well.

3. Network-relevance.

The Internet will be used for harvesting the documents and accessing the Web archives. Harvesting as such does not pose any significant load to the Nordic research networks, but a fast network makes the process easier.

The project aims at storing significant part of the information available in the Web for future generations. Therefore the initiative is very relevant to the Nordunet community.

4. Publication of results.

The project results will be available through the Web pages and the Nordic National Libraries and will be published at national and international forums.

Some of the tools will be available in the public domain for other national libraries who wish to harvest their domestic Web spaces. The tools may also be used by other organisations for archiving their own network resources.

On behalf of the Nordic National Libraries

Bendik Rugås Director ,
The National Library of Norway, Postboks 2674 Solli,0203 Oslo
Telephone: + 47 23 27 60 00, Fax: + 47 75 12 12 22; E-mail: Bendik.Rugaas@nb.no

Einar Sigurdsson Director ,
National and University Library of Iceland, Arngrímsgötu 3, 107 Reykjavík,
Telephone: + 354 525 5600; Fax: + 354 525 5615; E-mail: einsig@bok.hi.is

Erland Kolding Nielsen Director ,

The Royal Library in Copenhagen, Postboks 2149, 1016 København K,
Telephone:. + 45 33 47 47 47; E-mail; ekn@kb.dk

Esko Häkli Director,
The National and University Library of Finland, PB 15 (Unioninkatu 36)
Telephone + 358-9-191 23196, Fax + 358-9-191 23191,
E-mail: esko.hakli@helsinki.fi

Thomas Lidman Director,
The Royal Library in Stockholm, Box 5039, S-102 41 Stockholm,
Telephone + 46 8-463 40 00; Fax + 46 8-463 40 04; E-mail: tomas.lidman@kb.se

APPENDIX I - Functional requirements for the NWA Archive

The functional requirements for the Archive must reflect the present state of computer technology and what can realistically be achieved. It is however both wise and worth while to identify as many future requirement as possible. The major elements to be considered are:

1. **Authenticity.**

   Documents must be stored unchanged and there has to be means for proving that the document has not changed. Calculating and storing MD5 checksums and generating time stamps at the times the document is harvested will make authentication possible.

2. **Format.**

   All MIME formats should be supported. This is no problem for archiving, but indexing will be limited by the tool chosen by the project.

3. **Identification.**

Any identifier embedded in the document must be indexed. These identifiers should however not be used as archive ID for two reasons. First, there may be several versions of the same resource with the same identification. For instance, every article from one journal may contain the same ISSN. Second, an identifier mentioned in a document is not necessarily identifier of the document, but may belong to an another resource.

Every resource harvested must be assigned a unique archive ID. A decision has been made to use MD5 for this purpose. MD5 can be extended into a national bibliography number, which in turn can be used as Uniform Resource Name. This

will allow persistent linking of references and archived resources.

4. Access (navigation, indexing, searching).

Although bibliographic databases (Web indexes) created by indexing the archived documents will be freely available, access to the archived documents will depend on the Copyright and Legal Deposit legislation in each country and is very likely to be limited. Therefore provisions must be made for controlling the access to the documents according to those laws.

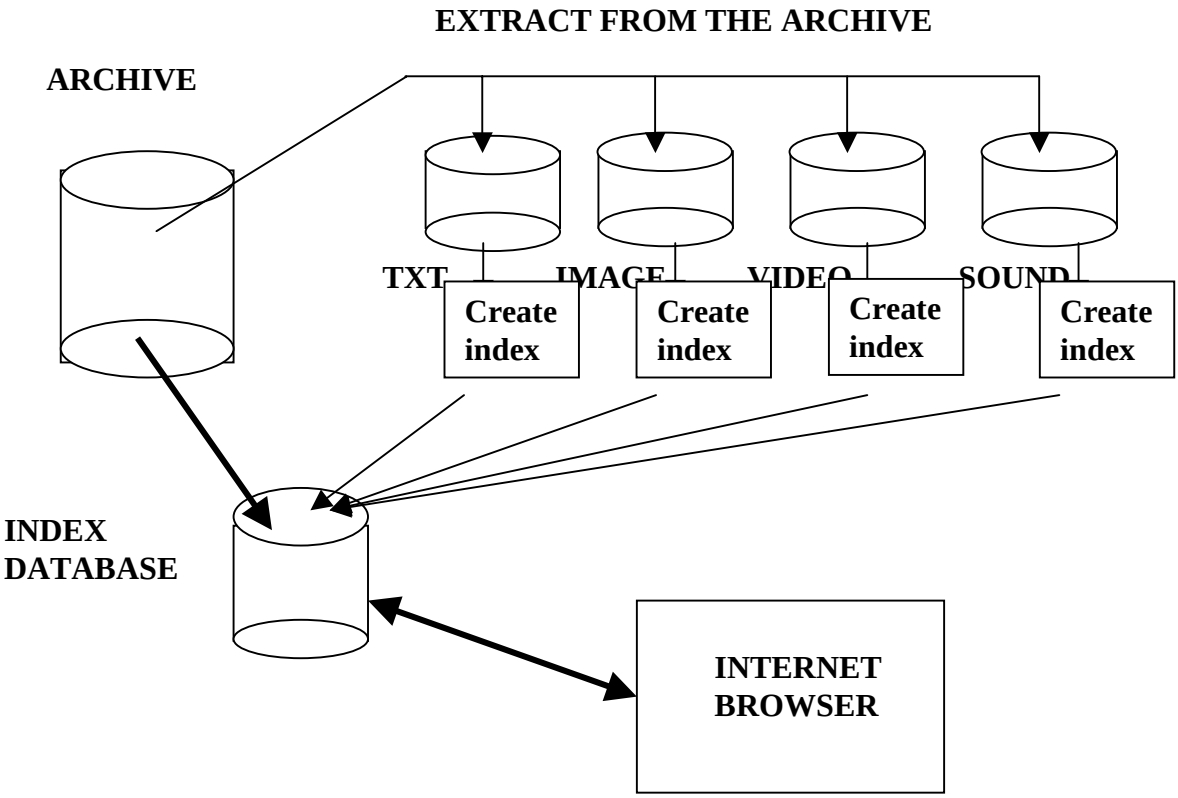
A. Indexing

Indexing will start by analysis of file formats. Subsequent processing in the indexing application will depend on format. a XML document can be processed more efficiently than an ASCII one. For some resources - images, for instance - indexing will be limited to harvester generated metadata plus textual information in the file header.

For textual documents linguistic methods should be used to enhance the indexing process. .

All indexing tools that have been considered can handle those file formats which are the most common ones in the Web. Some of the tools can even index image and sound files. In the future it will be possible to search also non-textual information such as sound and images.

In the following illustration the concept is described.



B. Search.

Search functionality will be based on the chosen product and improvements made to it. A search will produce a bibliographic reference compiled by the indexing software. References will contain links to the archive.

It should be possible to search:

- in the full text (the TXT index) using both specific and general search arguments
- by using harvester generated metadata such as archive ID and time stamp
- by domain/geography
- by organisation
- by arguments (popular expressions)
- search methods that would look to new ways of identifying documents of some particular format or type like images or sound (see indexing).

C. Navigation

If a URL is known the Archive can be accessed by using it to retrieve documents and navigation within the Archive can start from there (eg. by following links, moving to later/earlier versions of a document etc..). Some of the navigation capabilities needed are:

- by surfing
- by links as in the original
- by bi-directional/reversed linking.
- By following: "What documents refers to this page?" or "From where in the web can I get here?".
- by time i.e. move to later/earlier version of a document or site. The purpose is to investigate the history of documents, sites, etc.

5. User interface to the archive.

The interface to the archive should be through the Web and standard browsers like Netscape and Internet Explorer. Because of the variety of file formats that exist in the Web miscellaneous plug-ins will be needed.

APPENDIX II - Specification for an Archiving Robot.

Background

NWA (Nordic Web Archive) is a co-operation between the Nordic countries in the field of archiving electronic documents, in particular documents published on the World-Wide-Web. The first such project started in 1996 with a prototype robot software for searching and saving Web-documents. This robot was originally developed for indexing purpose within the Nordic Web Index project and has been modified for the special needs of an archiving robot. Much has been learned from this initial phase and it's now time to acquire a new robot which better meets our requirements.

Specification

General

- The implementation should be object-oriented and as platform independent as possible by using standards.
- The robot should, by multi-threading or otherwise, take advantage of multiprocessor/multi-port architectures.
- The robot must understand all versions of the following protocols: http, ftp, gopher and nntp. It should be possible to select/deselect protocols. It should be easy to implement new protocols.
- The robot must understand all versions of html and XML.
- It should include well known SGML DTD's, in particular TEI.
- The robot should have a nomadic clients that can be installed at certain large web sites or at web sites with low bandwidth lines.
- The robot should have a mechanism to check for and suggest web site names that are aliases.

The robot should maintain a database about all web sites and documents visited. This database should be used to obtain default values for parameters used to control the robot. The database should have a graphical GUI for administration. It should be easy to add or delete fields from the database (see also below concerning the GUI).

The database should at least contain the following fields.

- Name of web site
- IP number
- How often it is allowed to access the site and parameters used to calculate this.
- Disallow list for parts of the site that should not be visited by the robot, preferably given as a list of regular expressions.

Whether to respect robots.txt or not. There should be a general switch to select/deselect this option for all web sites.

- Statistics on failures and successes by time.
- Field to label a site as dead.
- Field for user-id and password. N.B. it is necessary to be able to have several userid-password pairs because a web-hotel may have different parts with
- different password. Needs also userid-directory link.
- Number of URL's on this site.
- Server type
- If a website is an alias for another, the name of the "real" site.
- List of websites which are aliases for this one
- Comment field where the operator may enter text.

Gathering

Here we introduce the concept of a collection. Typically one wants to give the robot the task of harvesting a selection of documents. The selection can be e.g. all documents found on a list of web-magazines. It can also be all documents found on all websites under a certain country code. These harvested documents we call a collection. The following entities should be possible to use to define a collection:

- List of websites. Optionally restricting to URL's which path's satisfy a regular expression
- Website names satisfying a regular expression. E.g. everything ending on ".se"
- MIME-type
- Language

It should be possible to select any combination of the parameters. It should be possible to use all parameters for a certain site. I.e. select all documents on a site for which: the URL's satisfies a regular expression, are of type text/html and are in Swedish.

It should be possible to decide, for each collection, how often it should be harvested. It should be possible to override the values obtained from the basic database discussed above. To each collection there should be an associated storage root where the harvested data should be written.

The robot should:

- be able to access web sites that demand user-id and password.
- handle frame-based pages
- handle maps
- collect all http-headers
- collect inline material also if it's not within the wanted domains. I.e., if a Swedish page has an inline picture which resides on a server in England, the picture
- should be collected.
- decide when to give up on a URL, after a certain time and certain number of tries.
- collect a document fully, including inline objects, as soon as possible.

- have several options for scheduling
- have the option to use different scheduling during the day then during the night

Storage

The functional specifications for the storage depends both on the harvesting methods and frequency and on requirements for accessing the Archive. From the point of view of the Harvester the storage should fulfil the following requirement:

- A collection should be stored under the storage root defined above
- The storage should be in a way that makes it easy to rebuild an original web site.
- For all objects both the content and the headers should be stored.
- The objects should be stored with a unique identifier as filename, not using "unsafe" characters.
- Together with the data there should also be a mapping URL vs. filename stored
- A time stamp should be included
- The parameters defining the collection should be stored. Here it is important that also information about eventual web site aliases are stored.
- The robot should handle URL's pointing to the same data in an intelligent way, i.e. only store the data once

Administration/Monitoring

- The robot should have a suitable administration interface, preferably a GUI. In the GUI you should be able to configure the robots configurable parameters,
- work with database administration and be able to monitor the robot.
- In the administration GUI you should also be able to handle erratic URLs (error 404 etc.) in a good fashion. The GUI should provide ways to configure
- automatic handling of these as well.
- Notify if the robots rule has changed for a web site
- It should be possible to inspect which URLs are waiting to be fetched.
- It should be possible to remove URLs from the queue, both on individual basis and on the basis of a regular expression.
- It should be possible to start and stop the robot gracefully, no 'kill -9'.
- There should be a standby-mode which suspends operation.

Define datastreams

An configurable interface/API should be provided to the robot for "tapping" data from a collection for specific use, like providing a full text search engine with material. It should be possible to make further selection on all parameters.

Statistics

To aid in monitoring it is important that enough information about the progress is made available. Therefore the robot should have a statistics module which at least reports the following

- Number and rate of URLs processed
- Number and rate of each return code
- Number and rate of new URLs extracted from the data
- New URLs sorted per web site
- Bandwidth usage
- Statistics on the occurrence of metadata.