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

Code Number: 145-095-E
Division Number: VI
Professional Group: Information Technology
Joint Meeting with: -
Meeting Number: 95
Simultaneous Interpretation: -

Critical technological and architectural choices for access and preservation in a digital library environment

Svein Arne Solbakk
svein.solbakk@nb.no
IT-director
National Library of Norway
Finsetvn 2, NO-8607 Mo i Rana,
Norway

Abstract:

The article describes some basic architectural choices for the access to and preservation of digital objects at the National Library of Norway. A digital repository is a core element for the handling of both access to and preservation of the digital objects. Strategies for giving access to the complete holdings include the use of a powerful search engine and the OAI protocol to harvest metadata from conventional catalogue systems to make textual or structured indexes.

August 2002

Background
The construction of a Digital National Library of Norway has already been going on for several years. Throughout this period of time the technology has evolved a lot, and various relevant standards and de facto standards has come and gone. However, there are some basic principal choices regarding the architecture of a digital library, that may be made independently of current technology.

This article discusses some of these principal questions regarding storage and preservation of digital collections, and regarding how to give access to the digital collections.

Objectives
The basic objectives for developing a digital library at the National Library are to:

1. be able to offer powerful digital library services adapted to user expectations and needs
2. be able to handle access to and preservation of large amounts of digital data with a variety of properties
3. be able to collaborate with other institutions on user services as well as the handling of access to and preservation of digital objects
4. let other service providers make value added services which include resources from the digital collections at the library

The digital repository
At the core of the digital library we have established a digital repository. Some important architectural choices for this storage are:

- ➔ The storage technology should be hidden from the applications
- ➔ Every object in the storage should be given a globally unique identifier
- ➔ Every object must have sufficient metadata for search, retrieval and preservation
- ➔ Formats and quality levels should be carefully chosen to facilitate long term preservation
- ➔ It should be possible to give all objects in the storage eternal life through migration and emulation
- ➔ The methods for handling a digital object should be the same regardless of the type of information contained in the object
- ➔ One should use off-the-shelf technology which is as open as possible

At the National Library of Norway we are currently developing a generalised input/output-service for the digital repository. Any application with a need to put a digital object into the storage, or with a need to get a digital object out from the storage, will have to use this general service. The type of technology being used to store and maintain the object will in this way be hidden from the applications. In addition, this strategy let us have a common approach towards advanced access control and the handling of copyright.

Hiding the storage technology from the applications makes it possible to change the technology whenever this is necessary without affecting the applications. One may e.g. change tape technology within a tape robot, change the tape robot itself, change the disk systems, change the storage area network, or change the servers and software administrating all the networked storage, and so on, and then migrate all data from the old to the new technology and at the same time let all applications and services be up an running without any modification.

Every object in the digital repository is given a globally unique URN (Uniform Resource Name) for identification. A URN resolution service is developed according to the guidelines given by IETF (Internet Engineering Task Force). The service is available for external institutions (giving URLs as input). The use of a common identification scheme for all the objects is essential to be able to hide the storage technology from the applications. In this way our applications always use the persistent URN to identify a given object, and the URN resolution service holds track of the physical location of the object at the moment.

The required metadata level is still under evaluation. Currently the metadata for the digital objects resides within various catalogue systems within the National Library, and a variety of cataloguing formats and levels are used. However, we are evaluating the concept of a minimum metadata core based on the Dublin Core format, which all objects should have as a minimum, regardless of what type of information the

object contains. This would make it easier to let a user search in several or all of the catalogues in a single operation.

However, it is obviously not possible for the National Library e.g. to catalogue every Norwegian web page at this level. Therefore one should try to make it as easy as possible for the publishers to include good Dublin Core records in their web pages, and encourage them to do this. Another strategy would be to try to extract as much meaningful information as possible from the web pages during the harvesting process, and then generate DC records from this.

In addition to the DC level metadata, it is possible to generate full text indexes of some information types to facilitate an "Internet search engine" kind of search. This is of course an interesting strategy for the harvested web pages, but it is also interesting for other kinds of objects containing textual information. Using digitisation and OCR-tools makes it possible to offer this kind of search methods on historical objects as e.g. manuscripts, newspapers, periodicals, and books. This may also be an interesting approach for images and sound using pattern and sound recognition to generate searchable data.

In addition to the metadata necessary to search the objects, it is necessary to have technical metadata to be able to preserve the digital information into the future. The National Library of Norway has still not made a final decision on the necessary level of technical information for the variety of information types that will be handled by the digital repository. However, there is currently a lot of work going on in the preservation community that will be considered closely.

To be able to preserve the digital objects over a long period of time, there should always be one high quality version of every digital object using a format with no loss of information. Taking into account the technological development, one should always use the highest quality level one can afford at the moment. This is still a major challenge for audiovisual information. Also, one should use a limited number of formats altogether to make it easier to maintain them over a long period of time.

For formatted textual objects and databases, using XML is a promising strategy. Converting such objects to XML makes it a lot easier to interpret the data in the future than if one would have to be able to interpret a vide variety of formats which are more or less dependent on specific hardware and software environments to work. Also, having the data in the XML format makes it possible to generate advanced full text indexes to facilitate various search strategies.

Our digital repository has only existed for a short period of time. Thus the preservation strategies has not been challenged very much yet. Our strategy is to migrate the high quality version of digital objects to new formats when this is necessary and possible. For some complex objects this may not be a possible strategy, and in this case we need to consider other strategies as emulation or maintaining a technical museum of computers and software. Unfortunately, none of these strategies give the answer to all the challenges.

The technology being used in the digital repository should be as flexible, scalable, standardised and reasonably priced as possible. Currently, Linux is an attractive platform which is getting more and more popular, and interoperability between various kinds of components from various vendors are continuously getting better. However, integrating technology from a variety of vendors is still a complex task! It may therefore be a good idea to get someone to be responsible for the support of the complete installation rather than signing separate support agreements on a variety of components.

A general search infrastructure

The average user currently expects to find an interface to the digital national library that resembles the interface of an Internet search engine. It should be possible to get a first impression on what kinds of

information the National Library has on a certain topic, using an Internet type of search on the complete holdings of the National Library. Obviously, one would get thousands of hits on a too general search, but this is exactly the same situation as on the Internet.

It should also be possible to refine a search by using structured metadata e.g. at the Dublin Core level. And at last, it should be possible to perform advanced searches within specialised systems for the more advanced information miner.

To support the advanced information miner, the National Library of Norway uses several specialised databases to catalogue various information types. Currently, most of these databases are stand-alone systems, which may not be easily integrated into common searches. However, most of the digital objects being catalogued currently reside within the digital repository in a coordinated way.

To be able to offer a better view towards the complete holdings of the National Library, experiments are currently being performed using two different but related strategies, namely simple Internet-like search and structured search.

To support an Internet search engine interface, OCR are used to generate text from image based textual information as digitised microfilm, newspapers, books, or manuscripts. In addition, metadata are exported as text from various metadatabases, and born digital textual information, like web pages, are prepared for indexing. Then a set of indexes are generated from this textual information together with information linking the text to the correct images, to the objects described by the metadata, or to the born digital information. At last, a powerful search engine is used to perform full text search on the indexes, and in the result list to direct the user to the digitised images or to the digital objects being described in the metadata records.

To support a more structured search, it is considered to implement an OAI interface on all the metadata catalogues, supporting export of metadata in the Dublin Core format. It will then be possible to harvest DC metadata in XML format at regular intervals from all the catalogues at the National Library, and then use this information as input to the indexer of a powerful search engine. Having such an index, it is possible to implement a structured search interface following the Dublin Core standard, to let the user search in a structured way in all the information with sufficient structured metadata at the National Library.

It is also possible to combine these two strategies in a variety of ways. E.g. by tagging the information with geographical information or with timestamps, one may support views into the complete holdings by giving a geographical area or a time period (or both) as input.

The architecture of the system easily supports common searches on a large number of indexes, and the indexes may be distributed throughout the Internet. Thus, services including data from several institutions are easy to implement, if the institutions uses a similar architecture.

In the same way, the architecture is well suited to let external service providers make value added services including the indexes residing at the National Library, and integrating this information with information from other sources.

A user interface example

The National Libraries in the Nordic countries have for several years worked together on issues related to capturing the web in the Nordic Web Archive collaboration. This summer an NWA project focussing on an access module towards harvested web pages was concluded. Since several of the countries already had

harvested a large number of pages, it was decided to make the access module independent of the implementation of the archive.

The objective of the project was to be able to search in the web archives in the same way as in conventional Internet search engines. But at the same time, one should be able to browse the complete harvested collection of web pages both within a given timestamp, and across time by viewing various versions of the same pages.

The solution was to define a common XML based format, which was to be used as input to the indexer of a search engine. All the web archives then were to be exported to this common format and indexed in the same way in the various countries. A powerful search engine is then used to perform conventional full text searches on the indexes.

When viewing a web page from the result list, the search engine is also used to locate sub elements within the web page, in the web archive. The search engine also searches for other versions of the page in question. If several versions exists, this information is displayed on a timeline at the top of the browser window. Also, when navigating within the web archive, either within a given timestamp or across several versions of a given web page, the search engine is used to locate the new page one is navigating to, including its sub elements.

The software being developed in the project will be made available as open source software. In this way the project partners hope that other institutions around the world can make use of the results, as well as develop them further. If several institutions around the world follows the same strategy for the harvesting of and access to web archives, the result may be a powerful toolset developed in a world wide collaboration.

The basic underlying architecture in this example is the same as sketched above. All the web pages reside within the digital repository and a powerful search engine is a vital part of the access technology.

The architecture being used makes it easy to support search into several web archives at the same time. The architecture is also well suited for other information types, and could also have been used e.g. to implement a national or a Nordic digital newspaper archive access module.

Future

The next step in the development of our digital National Library, is to establish advanced access control mechanisms integrated with a copyright management system. Also, research is needed to develop good user interfaces adaptable to the needs of various user groups.