World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

juli 1, 2005

Code Number: 151-E
Meeting: 119 Acquisitions and Collection Development

Different approaches to collections for end-users

Our digital heritage as source material to end-users: collection of and access to net publications in The National Library of Norway

Kjersti Rustad (Mrs.)
The National Library of Norway
Mo i Rana, Norway

Abstract

The National Library of Norway is a multimedia knowledge centre that offers its users source material on any media. This presentation will deal with how the National Library according to the Norwegian Legal Deposit Act collects net publications, and how these documents might be made available as source material to end-users.

A web archive consisting of millions of URLs collected over several years gives us a challenge when it comes to access. How does the end-user get access to net publications in the Internet archive for purposes of research and documentation? The presentation discusses possibilities of accessing the Internet Archive, but also limitations set by the existing legislation.

1. Introduction

A researcher working on the Norwegian press history recently contacted the National Library for information. He was conducting a comparison between the news presentation in printed newspapers, broadcasting and in Internet newspapers, and he wanted access to the source material. This episode might illustrate a typical end-user request. The question is, can the National Library provide this kind of information?

The National Library of Norway is the nation's memory, and it is a multimedia knowledge centre, as stated in the Library's vision. The National Library shall preserve and make available the Norwegian cultural heritage and make records of Norwegian cultural and social life available as source material for current and future users.

So the answer to the question is YES – the National Library is expected to provide our end-user with source material on any medium. Our main tool to achieve this is the Legal Deposit Act.

When the current legal deposit act came into force in 1990, Norway was among the first countries in the world that got a legal deposit legislation which included digital documents. The Act is based on the principle that all generally available information, regardless of form or medium, must be preserved and made available as source material for purposes of research and documentation. Although the Internet as we know it today did not exist in 1990, documents made generally available on the Internet are subject to legal deposit according to the present Act. This is the foundation for the National Library of Norway's activity with regard to collecting net publications.

2. Collection of net publications

Trial harvesting of net publications has been carried out since the mid 90's, and the first total harvesting of the entire Norwegian web space was carried out in December 2002. It is however this year - 2005 - that the National Library has started the regular operation of harvesting the Norwegian Internet domain.

Our selection strategy is based on harvesting all generally available documents from the Norwegian Web space, that is the .no domain. Since the Legal Deposit Act applies to all generally available documents either produced by a Norwegian publisher or specially adapted to the Norwegian public, documents on other domains such as .org and .com might also be subject to legal deposit. This means that we plan to include other domains in the longer term. We also include websites with robots.txt. This means that our harvester ignores the robots.txt protocol. This decision is a result of a thorough discussion.

As you may be aware of robots.txt is a standard protocol which is used to ensure that search engines do not overload a website or download material which the Webmaster consider irrelevant to search engines. It is common etiquette on the Internet to respect robots.txt. But when it comes to the National Library and our responsibility to preserve all generally available documents on the Internet, we have to ignore the robots.txt to ensure that important documents actually are being preserved. Our experience has showed us that losing whole web sites might be the consequence if we were to follow robots.txt. This would not be in accordance with the Legal Deposit Act. It has, however, lead to discussions among net publishers, and the National Library had to argue in favour of this practice and explain that our harvesting activities are legally motivated, as distinct from any other search engine.

We execute the harvesting using *Heritrix*[1] – a web crawler developed by the Internet Archive[2] in collaboration with, among others, the Norwegian National Library. Our first round of harvesting net publication with Heritrix started in February this year, and initially our ambition was to harvest the complete .no domain 4 times a year. This was based on the calculation that each harvesting session of the Norwegian Internet domain would go on for about three weeks and

[1] For more information about Heritrix, see URL: http://crawler.archive.org/ (Accessed April 11, 2005)
[2] For more information about Internet Archive, see URL: http://www.archive.org/ (Accessed April 11, 2005)

collect approximately 10 mill. documents, that is URLs. We soon experienced, however, that these calculations cracked completely. After the planned three weeks of harvesting, the web crawler had discovered more than three times as many documents as calculated, and the amount increased by the hour.

The harvesting of the entire Internet domain ensures the preservation of snapshots of how the Norwegian Internet domain looked liked at a specific moment in time. This is important source material for end-users. But it is also important to be able to preserve the entire history of one specific document. For instance, all issues of a magazine, all report in one series, and updates on a daily basis of a newspapers. The selective approach is an important supplement to the total harvesting of the all the Norwegian web sites. This selective approach is step two and our second goal regarding harvesting of net publications. The harvesting of all Norwegian Internet newspapers on a daily basis will be initiated in 2005. We also intend to start frequently harvesting of all Norwegian net periodicals in 2005.

3. International cooperation

The National Library of Norway participates in various projects on an international level in the matter of archiving the Internet. The Library is part of the International Preservation Consortium (IIPC), where we cooperate with other national libraries and the Internet Archive in USA.[3]

In addition, The Library is part of the Nordic Web Archive Project (NWA), which is the Nordic national libraries' forum for coordination of activities regarding harvesting and archiving of documents on the Internet.[4] Although the name of the forum might indicate that we have a common web archive, this is not the case. The Nordic national libraries do separate harvests and have separate web archives. However, the NWA Project has developed a tool for accessing web archives, called the NWA Toolset.

4. Preservation of net publications

The National Library has a perspective of 1000 years regarding preservation. That is, our aim is that the information on documents in any format shall be available as source material a 1000 years from now. Whether this is a realistic aim only our future end-users will know, but anyway this is our goal and we plan the preservation activities accordingly. The net publications harvested by Heritrix are stored in the National Library's Digital Long Term Preservation Repository.

5. Access to net publications

How does the end-user get access to net publications in the Internet archive for purposes of research and documentation? There are at least two aspects that should be taken under consideration regarding this question: metadata and legislation.

5.1. Metadata

The National Library will during this year decide to what extent documents in the net archive should be catalogued and included in the national bibliography. The need for bibliographic control of digital documents, how the documents should be made available, and the amount of documents in the Internet Archive are factors that have to be taken into consideration. Anyway, cataloguing the net archive will probably only be the case for a tiny fraction of documents.

Realizing the fact that probably far less than 1% of the documents collected from the Norwegian Internet domain might be subject to bibliographic registration, we have to have other tools for

[3] For more information about the IIPC, see URL: http://netpreserve.org/about/index.php (Accessed April 11, 2005)

[4] Fore more information about the NWA Project, see URL: http://nwa.nb.no/ (Accessed April 11, 2005)

gaining access to the web archive. The solution is this: All documents in the Internet Archive will be fully indexed using FAST indexing software. This will enable users to search the documents via free text. The documents will be accessed via the NWA Toolset. This software will provide the only way of access to more than 99 % of the material in the archive.

In order to illustrate how the net archive will appear to the user, the slides below demonstrate the NWA Toolset, which is the primary software to access the Norwegian Internet Archive.

Fig. 1

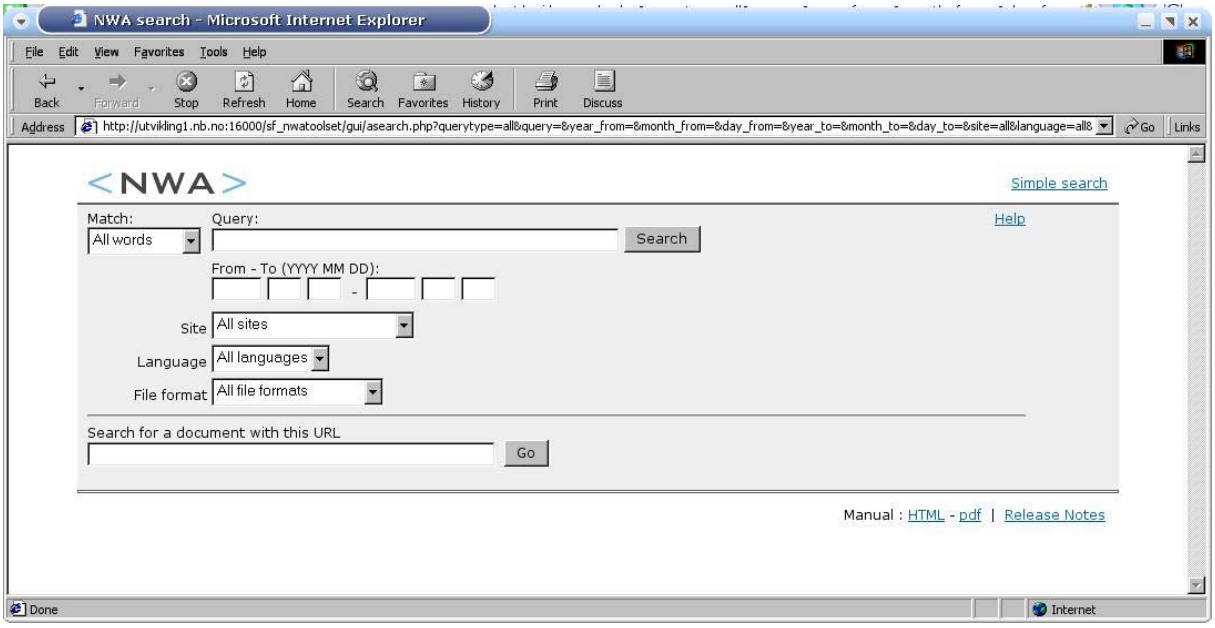

Figure 1 shows the user interface for the advanced search in NWA Toolset. The search facilities include free text search with Boolean operators and search for certain URLs. The searches might be limited by date, language and file formats, and by sites.

Fig. 2

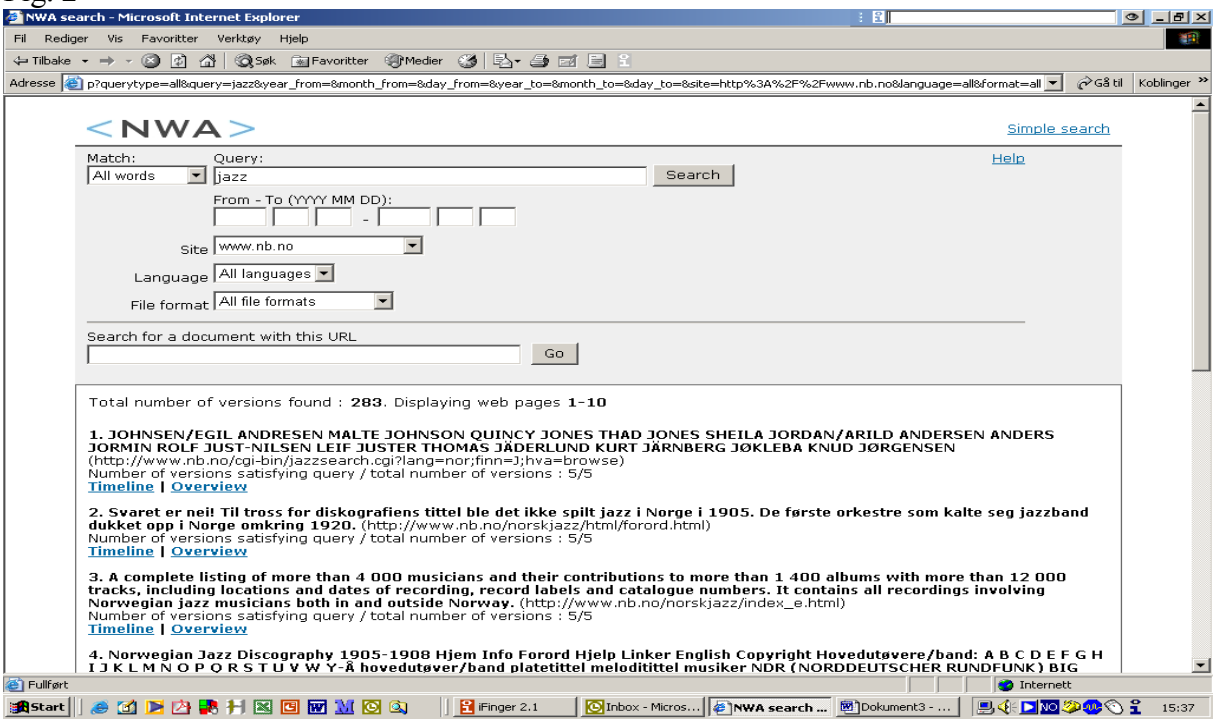

Figure 2 shows an example where we have conducted a search for the phrase *jazz*, limited to the site of the National Library of Norway. The hit list gives us two options for the presentation of each hit: Timeline or Overview.

Fig. 3

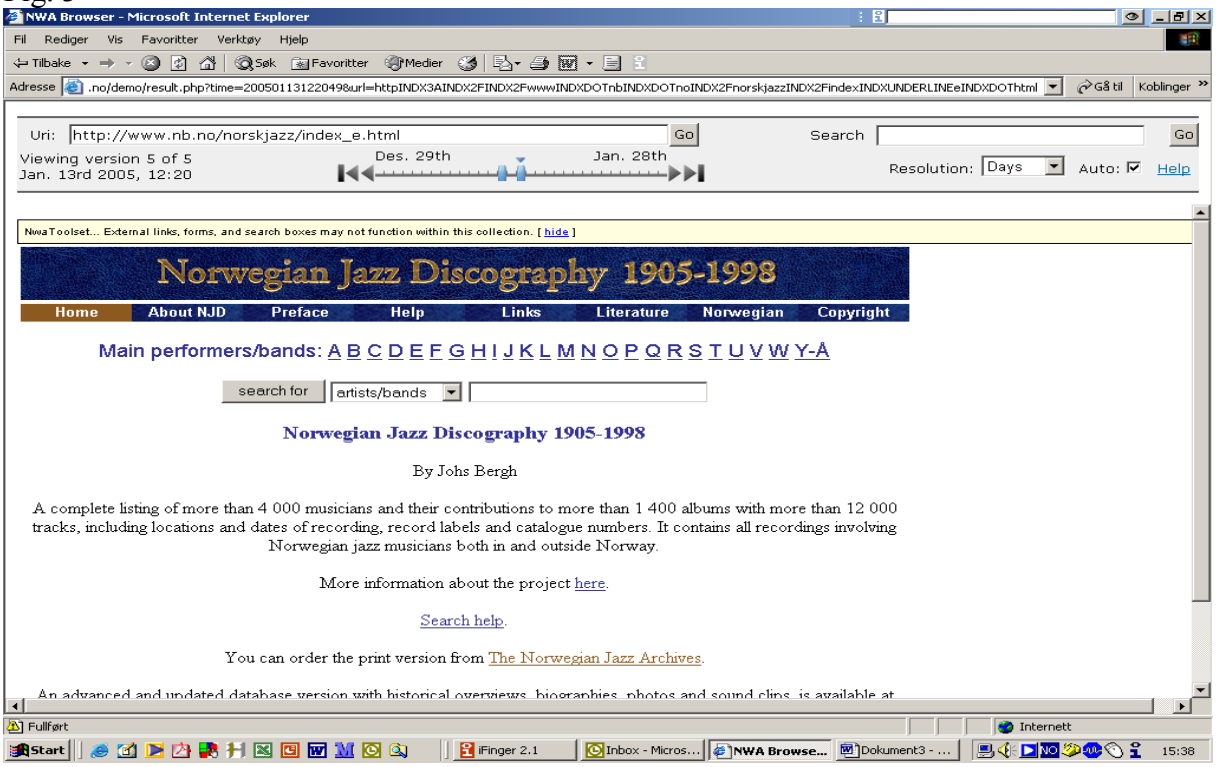

Figure 3 shows us the presentation of the document history via a timeline. As you can see this is version 5 of 5 downloads of this specific document, conducted January 13th 2005 at 20 minutes

past 12 a.m. The user might follow the history of the documents via the timeline. Options for resolutions of the timeline are years, months, days, hours, and minutes. This gives the user the possibility of a virtual voyage in time!

Fig. 4

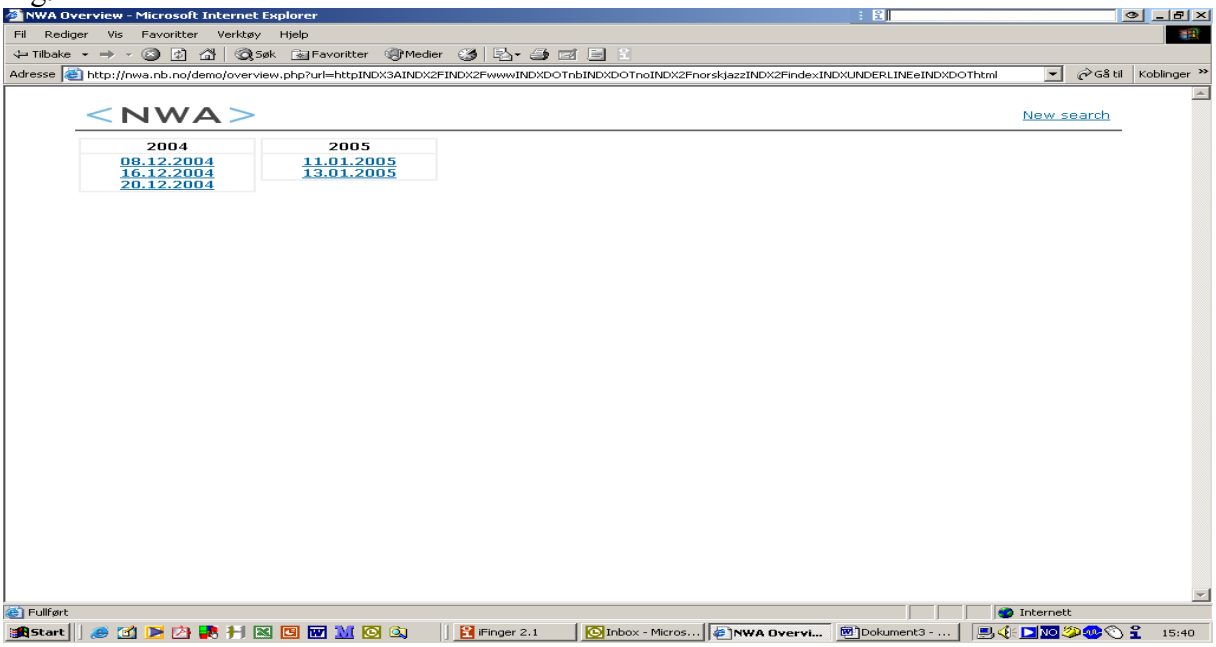

Figure 4 shows us the presentation of the document history via an overview of the number of downloads.

5.2 Legislation

Harvesting the Norwegian Internet domain enables the National Library to preserve our digital heritage. However, the existing Norwegian legislation limits the possibility of the public to access to the Internet archive.

The National Library has to take regulations both in the Legal Deposit Act, the Personal Data Act and the Copyright Act under consideration when it comes to giving access to the Internet archive to end-users.

The Legal Deposit Act states that documents deposited in accordance with the Act shall be made available for purposes of research and documentation. This means that the Act gives clear restrictions when it comes to making net publications available to the end-users. The purpose of the use of the documents decides who can get access to the publications. Possible users that might have access to the net archive are for instance researchers, scholars, students, teachers and other specialized users, for instance users with a specific hobby, such as genealogy.

Net publications, which are made available to the users, might contain both personal data and sensitive personal data. The handling of personal data is regulated in the Personal Data Act. The Act shall help to ensure that personal data are processed in accordance with fundamental respect for the right to privacy, including the need to protect personal integrity and private life. The possibility of processing personal data through search facilities in the Internet archive makes the information in net publications far more accessible than the traditionally published documents. This implies that the National Library has to have a licence from the Norwegian Data Inspectorate to conduct the harvesting of net publications. In addition, a similar licence for access is required before the Library can give access to the net publications for our end-users.

According to the Copyright Act the National Library cannot make the deposited net publications available to the public if this access is in conflict with the copyright owners' economic and ideal rights. According to the current legislation, the National Library might give access to the net publications only on on-site computers, that is, computers within the Library's premises.

To summarize, these are the demands that have to be fulfilled according to the current legislation before the National Library might give access to end-users to the net publications in the Internet Archive:

- The use of the documents must be for purposes of research and documentation
- A general licence agreement from the Norwegian Data Inspectorate according to the Personal Data Act must be signed
- The user must seek out the National Library's premises

This leads us to the recognition of the fact that the existing legislation gives the end-users less, or at least more restricted, access to digital legally deposited documents than to printed legally deposited documents.

The proposed changes in the Norwegian Copyright Act include the possibility of interlending digital documents on specific request. This means that the National Library may send a copy of a digital document upon the request from end-users situated on another library. The latter is responsible for deleting the copy after use.

Considering the fact that Norway is a sparsely populated country with large geographical distances, access to the Internet Archive only at the National Library premises in Oslo and Rana is not an optimal solution.

We can identify four levels of access for end-users:

1. On-site access for users via the National Library's computers. According to the proposed changes in the Norwegian Copyright Act this will be permitted.
2. On-site access for users via computers in universities, colleges and public libraries. This is not in accordance with existing legislation, but would in many ways offer a good solution for access.
3. Access for users from private computers.
4. Digital documents made available via for instance exhibitions on the National Library's web site. Licence agreement will be required for this.

The access to the Internet archive as stated in point 1 to 3 above must be limited to purposes of research and documentation. This means that open access to the public in general cannot be permitted according to existing legislation. However, there are ways of giving access to net publications in the Internet archive that might be the solution for specific parts of the archive. The National Library might sign agreements or licence agreements with the copyright owners which enable the Library to make net publications available to the public, for instance through exhibitions on the net via the Library's home site.

The proposed changes in the Norwegian Copyright Act include the possibility of digital licence agreements between libraries and copyright owner's organizations. This may give the National Library the possibility of making digital documents available to end-users, but it depends on payment.

The National Library's primary goal is to offer end-users on-site access via the National Library's computers. This gives, however, the end-users a geographically limited access. The matter of future access to the web archive is now under consideration by the National Library.

5.3 Which documents should the end-user have access to?
Basically, all legally deposited net publications should be made available as source material to end-users for purposes of research and documentation. In the matter of copyright protected documentation, the access of one specific user will be limited to the part of the net archive which is relevant for his or hers needs. When it comes to documents that are not protected by the Copyright Act, such as legal statutes, administrative regulations and other decisions by public authorities, there should be no need to limit access.

6. Conclusion

Over time the Norwegian Internet Archive will consist of billions of URLs which the National Library must enable the end-user to navigate among. The challenges regarding harvesting, preservation and access to web documents are huge, but the National Library believes that we have the infrastructure, the services and the organisation to accomplish this.

Net publications can from a technological point of view easily be made available to the public. Access is however limited by existing legislation, and it is vital to the National Library to ensure that all use of the Internet archive is in accordance with the Legal Deposit Act, the Copyright Act and the Personal Data Act.

This must however not prevent net publications, as records of Norwegian cultural and social life, from being available to end-users as source material for purposes of research and documentation.