Date : 14/05/2007

Webarchiving Internationally: Interoperability in the Future?

Grethe Jacobsen, dr. phil.
Head of Legal Deposit and Maps,
Prints and Photographs,
The Royal Library - The National Library of Denmark,
P.O. Box 2149, DK-1016 Copenhagen K,
Denmark
www.kb.dk

Meeting: 73 National Libraries **73 National Libraries**
Simultaneous Interpretation: No

WORLD LIBRARY AND INFORMATION CONGRESS: 73RD IFLA GENERAL CONFERENCE AND COUNCIL
19-23 August 2007, Durban, South Africa
http://www.ifla.org/iv/ifla73/index.htm

Abstract
Several national libraries are collecting parts of the Internet or planning to do so, but in order to render a complete impression of the Internet, web archives must be interoperable, enabling a user to make seamless searches. A questionnaire on this issue was sent to 95 national libraries. The answers show agreement with this goal and that web archiving is becoming more common. Partnering is a key ingredient in moving forward and a useful distinction is suggested in the labels curatorial partners (archives, museums) and technical partners (private companies, universities, other research institutions). Working with private, for-profit companies may also force national libraries to leave room for unorthodox thinking and experimenting. The biggest challenge right now is to make legal deposit, copyright and other legislation adapt to an Internet world, so we can preserve it and make it available to present and future generation.

Background and questionnaire

Several national libraries have begun collecting sections of the Internet, now generally considered part of the national cultural heritage. However, given the nature of the Internet, a national net archive that holds only parts of the Internet will never be able to render a full impression of the Internet, as it was available and used by its citizens at a given time. The long-range goal should, therefore, be, that all national net archives are interoperable, in order that any user from any national (or other) library will be able to make seamless searches in all net archives.

Implementing this goal involves not only technical and financial issues, but also legal issues such as copyright, materials published illegally on the Internet and protection of personal data, especially when it comes to giving access to collected materials.

This will require not only close cooperation between national libraries, but more likely partnerships in sharing the costs of developing technical know-how concerning collecting, giving access, sharing collections and preserving online materials. Equally important is the

need for national libraries to act together in lobbying legislators for permission to collect and give access to the archives and in negotiating deals with copyright holders.

As a beginning step towards that goal, The Royal Library, the National Library of Denmark, sent out a questionnaire in March 2007 to the CDNL and the CENL mailing lists of national libraries altogether 95 libraries.

We got 23 responses at the time of the deadline (March 27) and another 14 after the deadline but in time to be incorporated in this paper (a response rate of 39 %). Not all responses dealt with or were able to deal with all questions so the responses to some of the questions are lower. The response rate was enriched by the many comments that the questionnaire allowed and which demonstrate that the respondents are very involved in this issue and believe in the purpose of web archiving.

In the following the questions and comments will be discussed accompanied with a note on Danish practice, the latter less to promote our own experience but to clarify what this means on a practical level.

> Briefly about the Danish experience in general: we have been harvesting the top-level domain .dk since July 2005 when a new legal deposit law went into effect. We aim at preserving the Danish part of the Internet as part of the cultural heritage for future generations to experience; however, alone we will not be able to duplicate in entirety the typical Internet surfing of today unless we can provide access to the other parts of the network in other net archives.

Questions and responses

The first question had the heading

1. Interoperability in general

and the question asked was:

Do you agree with the statement "*The long-range goal should, therefore, be seamless access to internet material past and present to all citizens across national borders*

33 agreed with this statement while 4 said no. Seven commented on the question and revealed that in fact only one respondent was against the statement while the remainder thought that the goal was worthwhile but unrealistic or at least not attainable at present due to legal problems and the varied collection policies. The comment against is worth a discussion:

> "I do not agree that access needs or should be 'seamless'. The great strength of the internet is its heterogeneity. I believe that an attempt to impose a homogenous solution would stifle innovation and would be ultimately fruitless."

I read this as arguing not so much against the stated goal (seamless access) as against a form of cooperation that would prevent innovative strategies and solutions. The first aim of any cooperative effort would be, therefore, to insure that continuous development would take place and that development should leave room for unorthodox thinking and experimenting.
This does represent a challenge to national libraries, especially their directors. They will typically be faced with – on the one hand – demands from public authorities and politicians to

find solutions here and now and produce some results to show for all the money earmarked for web archiving and – on the other hand – from those engaged in the task of web archiving demands that they be able to go several ways at one time, to gamble on methods and above all to have room for experiments and failures.

The second question had the heading

2. Net archiving – actual or planned

with the following questions asked

Are you currently doing net archiving?

19 respondents said yes to the question of being currently engaged in web archiving while 18 said no. 11 of the latter are planning to begin web archiving, while 7 are not. This means that web archiving no longer is an exclusive task done by only a few which was the case not so many years ago.

Of those currently harvesting or planning to do so, 15 harvest (or plan to harvest) the entire national domain, 27 collect (or plan to collect) selected websites within the national domain, 13 also harvest (or plan to harvest) outside the national domain while 8 harvest (or plan to harvest) websites according to language.

> In Denmark, the Legal Deposit Act allows us to harvest materials published within the .dk top level domain as well as materials published from other Internet domains which are directed at a public in Denmark, so we harvest the entire national domain as well as selected websites outside the domain. We have found about 30.000 websites outside .dk that are aimed at a Danish audience, primarily sites with Danish text but also sites belonging to Danish companies or institutions, or to individuals (musicians e.g.) who are domiciled in Denmark

We then asked for information on the web archiving activities, specifically if respondents had created websites with this information and if so we asked that URL's be provided. The heading and questions were as follows:

> **3. Information on your activities:**
> **3.a** Do you have a website with information on your net archiving activities
>
> **3.b** Do you have policies on collecting internet materials?
>
> **3.c** Do you have policies for discovering and including relevant websites?

15 respondents said yes to having a website with information on their activities. A closer examination of these sites reveals that most have an English version of the website, so clearly web archiving is seen as an international endeavour and information is directed not only at the home audience but also at colleagues and other interested parties abroad. 4 have websites in their own language and one had a website in German and French.

Sharing policies is another matter. All active collectors have policies or are working on formulating policies.[1] However, only 6 respondents make them available online. This may be due more to lack of time for translation than to a desire (or demand) not to make such policies generally available.

> Certainly, the former is the case for Denmark. We maintain a bilingual website www.netarkivet.dk (Danish) and http://netarkivet.dk/index-en.php (English) and

[1] One respondent who does web archiving did not answer this question

have decided to translate and publish our policies in English as well as Danish. So far one policy has been published.

The fourth question dealt with the legal basis for collection whether legislation (or other government regulation) or voluntary agreement

4. Guidelines for collection

Do you collect (please check one or more):

According to legislation	☐	(please specify e.g. legal deposit law):
Through voluntary agreement with other institutions	☐	(please specify):
Through voluntary agreement with communities	☐	(please specify):
Through voluntary agreement with private organisations or persons	☐	(please specify):
Other	☐	(please specify):

11 respondents collected according to a legal deposit law, one according to other legal mandate, one complying with copyright law and one as a special exception to a law on personal data. Yet another respondent is collecting the national domain as part of a project. 4 respondents collected according to agreements with publishers, other communities and Internet Archive and 1 had an opt-out approach.

It appears that there is still some way to go for those countries that want to collect through legal deposit act or other legislation, but more than half of those who indicated that they are engaged in web archiving do so with a mandate in a legal deposit act while three more had other legislative mandates. It should be noted that one library active in web archiving, the National Library of the Netherlands, has no legal deposit at all and therefore has a long tradition for negotiating with publishers to deposit the national heritage whether online or in physical form.

The Danish legal deposit law that allows for harvesting the Danish parts of the Internet was passed in December 2004 and went into force in July 2005. Prior to that we were able to collect net publications from 1998-2005 according to the previous act on Legal Deposit (in force 1998-2005). In addition, we had permission from the Ministry of Culture to do selected types of harvests during the years 2001-2004 as part of various projects in preparation for a new legal deposit act.

An important issue to a net archive will be its integrity and we asked therefore if those libraries that were engaged in web archiving manipulated the archive by discarding any materials collected and if that could be traced.

5. Manipulation of content of archive

5.a How much do you keep of the harvested materials in your archive?

Everything harvested	☐	
Only part of the materials harvested	☐	(please specify):

5.b If you discard materials can it be traced:

Yes	☐	(please explain):
No	☐	

21 respondents said they kept everything harvested while 1 respondent answered that the library kept only part of the materials harvested, namely files less that a certain limit. The

limit was not specified in the answer, nor on the library's website. In general it seems that the libraries endeavour to keep the archive intact.

Access is a basic element in the raison d'être of a net archive – or any archive for that matter. If there is no access to the archive, there really isn't any point in collecting, as the purpose of a net archive is to document part of a nation's cultural heritage. Furthermore, access should be for everybody just as the access to all other materials belonging to a nation's cultural heritage. We therefore asked about

6. Access to archived materials

Do you allow general, online access to your net archive?

*If "**no**" please answer the following question:*

Limited access:	(please specify, e.g. research only, statistical purposes)
No access:	

8 of those collecting said yes to the question "Do you allow general, online access to your net archive?" However, 3 of those who said yes noted, that technically, access to the public was not yet possible. The remainder had various types of on and offline access, including access with permission of publishers (6), access on premises (4) and access only to researchers (4). We may conclude that general online access to net archives is not yet common for legal reasons (copyright and data protection) as well as technical reasons.

> In Denmark we allow only access for research and statistical purposes and then only for researchers on a post-doc level. The reason is not, as one might expect, copyright legislation but the Danish Act on Processing of Personal Data. The Danish Data Protection Agency has determined that collecting public materials on the Internet and placing it in a net archive in effect may lead to the processing of personal data, which is covered by the act on Processing of Personal Data. We are about to begin negotiations with the Agency to allow for access for at least some parts of the data collected, for example websites of public institutions. Our original goal was to allow for general access from the reading rooms of the Royal Library and the State and University Library, in order that this part of the cultural heritage was accessible to all citizens just as all books and other types of published works are available to everybody, whether on loan or to be used on the library's premises.

Exchange of materials harvested, of URLs collected or other information on one's net archive might be of interest to other archives and therefore an important element in international web archiving cooperation. Questions 7 dealt with that:

7. Seamless harvested materials

Are you able to provide copies of harvested material for other national net archives?
*If "**yes**" please answer the following question:*
Can you provide (please check appropriate box(es)):

Copy of harvested materials	☐	
Information on URLs collected and timestamp for harvest	☐	
Other information	☐	(please specify):

Only 3 respondents were able to provide copies of the harvested materials, 2 didn't know and 15 answered no to the question. 8 respondents were able to provide information on URLs collected and give a timestamp for materials harvested and one was uncertain as to this type

of international cooperation. In addition, two pointed out that metadata and catalogue records for some materials harvested may be found in their online catalogue.

> In Denmark, we are unable to provide copies and information on what we have harvested to other than researchers who have obtained permission to get access to the archive. We do not have any records in our OPAC or any plan to have part or all of the web archive catalogued as records in the OPAC. We do have records for those net publications that we collected 1998-2005 but cannot give access to these publications at the moment, as they are part of the net archive and thus covered by the Danish Act on Processing of Personal Data

Another aspect of international cooperation is assisting each other in finding materials which can be said to be part of a nation's cultural heritage but which would be difficult to discover without some assistance from colleagues. We asked, therefore, if such a type of cooperation would be possible.

8. Cooperation with other net archives

Would you be able to (please check appropriate box(es))

Give information on URLs within your national domain that are of interest to other national libraries	☐
Collect websites within your national domain that are of interests to other national libraries	☐
Allow other national libraries and their users to access your archive	☐

14 respondents were able to provide information on URL's within their national domain that could be of interest to other national libraries, 13 would also be able to collect websites while 10 allowed access to other national libraries.

> The Danish net archive cannot provide information on URLs when it comes to harvested materials, nor collect websites for other national libraries. We can, of course, give information on URL's and content that is publicly available on domain .dk if we have come across such information.

Closely connected with the issues of cooperation is the choice of partners which was the topic of the ninth and final question, where under the heading

9. Partnerships in harvesting internet materials

we posed the following questions:

9.a Do you harvest in cooperation with other institutions?

*If "**yes**" please answer the following question:*

9.b Do you work with (please check appropriate box(es)):

Other national libraries	☐	
Other libraries	☐	
Other public institutions	☐	Please specify type (archive, museum, university, etc):
Communities	☐	Please specify:
Private companies/institutions	☐	Please specify sector (IT, trade, publishing, etc.):
Individuals	☐	

13 of those harvesting said yes to that question and also provided more details of that partnership

4 worked with other national libraries, 9 with other libraries, 8 with other public institutions, 4 with communities and 4 with private companies or institutions. As the numbers indicate

several of the respondents worked with more than one type of institution, libraries being a preferred partner but other public and private institutions mentioned were Internet Archive, universities, national archives, film and sound archives, museums, publishing companies, a private trust. From the comments it appears that what governs libraries' choice of partners are those partners' commitment to web archiving and/or their technical know-how.

Having partners also means complications in terms of who owns the archive and to the question:

> **9.c** How have you solved issues concerning ownership of the archive and its contents:

4 respondents said they had joint ownership, while 7 stated that their institution owned the material and one respondent said they hadn't sorted it out completely, while 3 had various agreements with partners.

Among the costs of building and maintaining a web archive are the costs of development, which is also an important factor in the libraries' thought on partnership.

To the question

> **9.d** How have you solved technical issues
>
> | | | |
> |---|---|---|
> | Joint development | ☐ | |
> | My institution is responsible | ☐ | |
> | My partner is responsible | ☐ | |
> | Other arrangement: | ☐ | (please specify) |

6 answered that they are engaged in joint development with partners, while 4 institutions are in charge of development, 2 leaves it to the partner and 1 relies on software developed by IIPC (International Internet Preservation Consortium).

> In Denmark, the two libraries engaged in web archiving, The Royal Library (the National Library) and the State and University Library, have established a virtual institution, "netarkivet.dk" (netarchive.dk) which is governed by a Steering Committee of 6 members (3 from each library) representing expertise in web technology, IT, legal deposit and collection building. The committee meets 2-3 times a year to discuss and decide on economic, technical, policy and legal issues. Netarkivet.dk has a daily manager who reports to the Steering Committee and supervises the daily work. Development is partly shared between the two libraries, partly by partners in the IIPC.

We also asked those who did not cooperate with other institutions in harvesting to answer this question:

> If you were to engage in partnerships when harvesting internet materials, which type of institution would you want to work with and what is your preferred solution to issues concerning ownership of the archive and its contents and to technical issues?

Actually, several of those who are doing web archiving, also answered this. The comments underlined the need for cooperation with institutions that had technical as well as collection expertise to offer along with a commitment to preservation issues. In some cases the libraries distinguished between technical partnerships and curatorial partnership. Technical partners mentioned are universities, research institutions and private companies and as curatorial

partners: archives, museums and other trusted repositories. As the National Library of Australia puts it “the more shared work that can <be> done the better”.

> In Denmark the key word since the first thought on web archiving appeared has been cooperation and partnering. We are active in the IIPC and in European initiatives and projects and also working very closely with the other Nordic countries who are all doing web archiving

Concluding remarks:

All respondents support the idea of web archiving although for some it appears unrealistic that we will be able to archive the entire Internet. The challenge for national libraries will be to make it a goal that can be realized.

Partnering and close cooperation is a key ingredient in moving forward. While this is not new to libraries, the new feature is that partners come from a variety of fields both public and private. It appears to be useful to distinguish between curatorial partners and technical partners. In finding curatorial partners it will be necessary to look beyond libraries to archives, museums and other institutions with depository functions. Probably these will be all public. Technical partners could be the same as the curatorial, but the dominant type of institutions mentioned is private companies, universities and other research institutions. Working with private, for-profit companies may have the added advantage that national libraries will be forced to keep in mind and plan for the fact that web archiving must be accompanied by continuous research and development which must also leave room for unorthodox thinking and experimentation. The challenge for national libraries will be to find the right partners for the various functions.

The biggest challenge for national librarians right now is, however, to make legal deposit, copyright and other legislation adapt to an Internet world, so we can preserve this piece of the national heritage and make it available to present and future generations.

At the end of the questionnaire we asked for comments on the questionnaire and the issues raised and one of these provides a perfect conclusion to this paper and an exhortation to IFLA and the national libraries “We hope this questionnaire will help all of us so that IFLA is able to push forward the legislative issues/problems of web archiving.”