Webarchiving Internationally: Interoperability in the Future? Results of a survey of webarchiving activities of national libraries, march 2007

By
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

This a revised version of a paper published on the IFLANET prior to the World Library and Information Congress: 73rd IFLA General Conference and Council *held in Durban, South Africa, August 2007 (English version: http://www.ifla.org/IV/ifla73/papers/073-Jacobsen-en.pdf; French version: http://www.ifla.org/IV/ifla73/papers/073-Jacobsen-trans-fr.pdf; Spanish version: http://www.ifla.org/IV/ifla73/papers/073-Jacobsen-trans-es.pdf)*

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, webarchives must be interoperable, enabling a user to make seamless searches across archives. A questionnaire on this issue was sent to 95 national libraries in March 2007. The answers show a high level of agreement with the goal of interoperability among webarchives. Partnering is a key ingredient in moving forward and a useful distinction is suggested in the labels curatorial partners (archives, museums), technical partners (private companies, universities, other research institutions) and users (research institutions and communities as well as the general public interested in the cultural heritage). The biggest challenge right now is to make legal deposit, copyright and other legislation adapt to an Internet world, so we can not only collect and preserve internet material but also 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 be, therefore, that all national net archives are interoperable, in order that future users will be able to recreate contemporary use of the Internet from any national (or other) library or webarchive.

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 16 responses before july 1st, when we closed the website containing the questionnaire. This gave us a total of 39 unique responses (a response rate of 41 %). Some institutions answered twice but are only counted as one response. Not all responses dealt with or were able to deal with all questions so the response rates to some of the questions are lower. The response rate was enriched by the many comments that the questionnaire allowed, demonstrating that the respondents are very involved in this issue and believe in the purpose of webarchiving.

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. A separate article dealing with Danish experience 2005-2007 found here http://netarkivet.dk/publikationer/CollectingTheDanishInternet_2007.pdf (English) and here http://netarkivet.dk/publikationer/CollectingTheDanishInternet_2007_ES.pdf (Spanish)

> 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

Under the heading

> **1. Interoperability in general**

we asked:

> **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***

34 agreed with this statement while 5 said no. The comments reveal, though, that 4 of the latter agreed with the goal but found it unrealistic, leaving only one respondent who was against the goal

> "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."

The second question dealing with

2. Net archiving – actual or planned

asked

Are you currently doing net archiving?

21 respondents are currently engaged in webarchiving, 11 more are planning to begin webarchiving, while 7 have no plans. This means that webarchiving no longer is an exclusive task done by only a few which was the case not so many years ago.

It should be noted that 18 of those active in webarchiving are members of the IIPC (www.netpreserve.org) whose mission is to "acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations" (http://netpreserve.org/about/mission.php)

We also asked of those active or planning to engage in webarchiving

Do you harvest (choose one or more):

The entire net within your national domain	☐
Selected websites within your national domain	☐
Outside your national domain	☐
According to language	☐

18 (= 58%) are currently harvesting (or planning to harvest) the entire national domain, 29 (= 93%) collect (or plan to collect) selected websites within the national domain, 13 (= 42 %) also harvest (or plan to harvest) websites outside the national domain, while 9 (= 29%) 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 webarchiving 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?

16 respondents said yes to having a website with information on their activities (see APPENDIX for at list of those). A closer examination of these sites reveals that 4 have an English version of the website (a few more websites are under reconstruction with indications that English language web pages will be forthcoming), the remainder have websites in their own language.

All active collectors have policies or are working on formulating policies.[1] However, only 6 respondents make them available online and fewer in English. 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 have decided to translate and publish our policies in English as well as Danish. So far (September 2007) one policy has been published in an English translation.

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): |

14 respondents collect according to a legal deposit law, 5 according to other legal mandate, while 18 respondents (also) collect according to agreements with publishers, research communities and institutions.

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

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

An important issue to a webarchive will be its integrity and we asked therefore if those libraries that were engaged in webarchiving 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	☐	

23 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:	

6 of those collecting said yes to the question "Do you allow general, online access to your net archive?" 2 will provide such access when it is technically 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 very restrictive for legal (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 who have obtained a master's degree. 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. We are about to embark on a project that will analyse the possibilities for fulfilling the demands of the Agency to protect sensitive personal data among the harvested material while giving general access to material that does not contain such data. Our goal is to allow for general access from the reading rooms of the Royal Library and the State and University Library, thus making this part of the cultural heritage accessible to all citizens just as 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 webarchiving cooperation. Question 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 4 respondents were able to provide copies of the harvested materials, while 10 respondents were able to provide information on URLs collected and give a timestamp for materials harvested. This is the minimum information needed for sharing knowledge of archived materials. One library has some metadata on harvested materials in their catalogue and that can be shared.

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

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 (= 36% of all) were able to provide information on URL's within their national domain that could be of interest to other national libraries, 14 (= 36%) would also be able to collect websites while 10 (= 26%) allowed access to other national libraries.

> The Danish netarchive cannot provide information on harvested URLs nor collect websites for other national libraries. We can, of course, give information on URLs and content that is publicly available on domain .dk if we 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	☐	

15 of those harvesting said yes to that question and also provided more details of that partnership 6 worked with other national libraries, 9 with other libraries, 9 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 webarchiving 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:

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

Among the costs of building and maintaining a webarchive 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)

7 answered that they are engaged in joint development with partners, while 5 institutions are in charge of development, 2 leave it to the partner and 1 relies on software developed by IIPC (International Internet Preservation Consortium).

In Denmark, the two libraries engaged in webarchiving, 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 webarchiving, 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 words since the first thought of webarchiving appeared have 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 webarchiving

Concluding remarks:

The survey shows that the overwhelming majority of respondents support the idea of webarchiving although for some it appears unrealistic. 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 as well as research institutions and communities who will also be the future users. Technical partners could be the same as the curatorial, but the dominant type of institutions mentioned is private companies, universities and other research institutions.

Tasks ahead for national libraries:

1. Revise (where needed) legal deposit legislation to include internet material
2. Get legislation that will allow access to webarchives (copyright law is the lesser obstacle; laws on personal data the greater)
3. Facilitate provision of information on sites of interest to other national libraries
4. Facilitate provision of assistance to other national libraries in collecting these sites

Recommendations

National libraries should

- Examine how much information on webarchives that can be shared
- Identify obstacles to sharing
- Work for legislation or agreement that allow for seamless access through interoperability
- Work through IIPC to attain interoperability

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 webarchiving."

APPENDIX: list of URL of websites dealing with webarchiving

institution	URL for website
The National Library of Sweden	http://www.kb.se/soka/internet/sv-webbsidor/
The National and University library of Iceland	http://vefsofnun.bok.hi.is/
The National And University Library Zagreb	http://www.nsk.hr/DigitalLib.aspx?id=80
Martynas Mazvydas National Library Of Lithuania	http://ei.libis.lt:8080
National Library of Norway	http://www.nb.no/fag/nasjonalbibliotekets_samling/nettdokumenter__1
Llyfrgell Genedlaethol Cymru - The National Library of Wales	http://www.webarchive.org.uk
National Library of Estonia	http://digar.nlib.ee/
National Library of the Netherlands (KB)	http://www.kb.nl/hrd/dd/dd_projecten/projecten_webarchivering.html; http://www.kb.nl/hrd/dd/dd_projecten/projecten_webarchivering.html (English)
The Royal Library, Denmark	http://netarkivet.dk/index-en.php
The Library of Congress	http://www.loc.gov/webcapture/
The British Library	www.webarchive.org.uk
National Diet Library, Japan	http://warp.ndl.go.jp/
Swiss National Library	www.e-helvetica.ch (website is being updated!)
National Library of Australia	http://pandora.nla.gov.au/
National Library of the Czech Republic - Národní knihovna ?R	http://en.webarchiv.cz/
Bibliothèque nationale de France	http://www.bnf.fr/pages/zNavigat/frame/version_anglaise.htm?ancre=english.htm (English)
	www.bnf.fr/pages/infopro/depotleg/dl-internet_intro.htm