LIBER QUARTERLY 18 (3/4), December 2008 – ISSN: 1435-5205. P366–376
http://liber.library.uu.nl/
Igitur, Utrecht Publishing & Archiving Services

Web Archiving: Issues and Problems in Collection Building and Access

Grethe Jacobsen

Head, Department of Legal Deposit and Department of Maps,
Prints and Photographs, The Royal Library,
P.O. Box 2149, DK-1016 København K, Denmark,
gja@kb.dk

Abstract

Denmark began web archiving in 2005 and the experiences are presented with a specific focus on collection-building and issues concerning access. In creating principles for what internet materials to collect for a national collection, one can in many ways build on existing practice and guidelines. The actual collection requires strategies for harvesting relevant segments of the internet in order to assure as complete a coverage as possible. Rethinking is also necessary when it comes to the issue of description, but cataloguing expertise can be utilised to find new ways for users to retrieve information. Technical problems in harvesting and archiving are identifiable and can be solved through international cooperation. Access to the archived materials, on the other hand, has become the major challenge to national libraries. Legal obstacles prevent national libraries from offering generel access to their archived internet materials. In Europe the principal obstacles are the EU Directive on Data Protection (Directive 95/46/EC) and local data protection legislation based on this directive. LIBER is urged to take political action on this issue in order that the general public may have the same access to the collection of internet materials as it has to other national collections.

Key Words: Web archiving; webarchiving; internet archiving; national collections; access; Denmark.

Introduction

Denmark began web archiving in 2005 when a new Legal Deposit Act went into force. This article will analyse our experiences in the context of the other Nordic countries and focus specifically on three issues:

- collection building
- national vs. international collection
- access to web archives.

Collection Building

The Royal Library, founded 1648, has since the late 18th century functioned as the National Library of Denmark and as such has a long experience in building a national collection. We also have had mandatory legal deposit since 1697.[1] Since 1902, two institutions have shared the administration of the Legal Deposit Act, the Royal Library at Copenhagen and the State and University Library at Århus. While the handling of works in physical form is divided between the two institutions, we share the task of collecting the internet, and for that purpose we have established a virtual institution called Netarkivet.dk (Netarchive.dk) to take care of web harvesting and web archiving. The Netarkivet.dk is staffed by members from both institutions with a day-to-day manager currently based at the State and University Library.

When in 2004 new legislation was being prepared which would allow us to harvest the Danish parts of the World Wide Web, we built on previous definitions of what a Danish national collection should contain. Since 1781 the definition of the Danish national collection has been that it would contain printed works which are produced within the country as well as works which are produced outside the country, but which deal with Denmark or Danish people or are translations of works by Danish authors. This concept was used in the new act to apply to materials published online as well. The current act, Act on Legal Deposit of Published Material no. 1439 of 22. December 2004, states in article 8 that 'Danish material published in electronic communication networks is subject to legal deposit' and defines 'Danish material' as materials 'published from Internet domains etc. which are specifically assigned to Denmark', or materials 'published from other Internet domains etc. and … directed at a public in Denmark.'[2]

The first type of material — and the largest bulk — is the easiest to collect, as it concerns materials made public on Top Level Domain (TLD) DK. We have three strategies for collecting this material in order to assure that we catch everything that is covered by the act:

1. a general snapshot of the entire <.dk> domain (ideally) done four times a year
2. selective but more frequent harvests of more dynamic sites
3. irregular harvests of selected websites in connection with events

1. Snapshot (or Cross-sectional) Harvesting

The first step is feeding the harvester with a list of all domains under TLD DK which we get from the Top Level Domain administrator.[3] Issues concerning collection building come up when we run into the problem of websites with materials which may be considered not published and hence not subject to legal deposit. In that case, however, we can build on our experiences with physical material. Online materials are considered public if anybody can get access to it, whether for free, by paying a fee or by providing information to the website (giving your name, e-mail etc). Websites of associations are similarly considered public, if membership is open to all or segments of the general public. Internal websites of companies, institutions, associations and research projects are another matter. They are not published and not subject to legal deposit. Explaining that difference to website owners and others who ask us, is not the problem. The problem lies instead in the fact that while it is technically possible to circumvent a log-in site, the harvester cannot distinguish between websites with published materials and sites with private materials. The only recourse we have at the moment is manual handling, which with the rapid increase of the Danish Internet (the number of domains under TLD DK has doubled in three years), and given the resources we have, is rapidly becoming unrealistic when we talk about cross-sectional harvesting. Instead, we will try to identify technical features that would distinguish access to public sites from access to private sites in order to capture more of the public websites with log-in.

2. Selective Harvesting

When doing this type of harvest, log-in is less of a problem, as we do a much smaller number of sites. The idea behind the selective harvest is to gather web pages that are frequently updated and which would be missed by the snapshot harvest. These have been defined as:

- news sites (national and regional media)
- 'typical' dynamic and heavily used sites representing civic society, the commercial sector and public authorities
- experimental and/or unique sites, documenting new ways of using the web (e.g., net art).

Currently we collect 82 sites. The challenge here is to select the right sites and to determine how much of the site should be harvested. News sites are the easiest category to select, amounting at the moment to 30–35 sites, but the media market is constantly changing, so developments must be continuously monitored. The other two categories are more difficult to select. The web has fostered new ways of human communication (virtual communities of all kinds), and this is important to document. An Editorial Advisory Board of five people, representing researchers and media professionals, assists the Netarchive.dk in the selection process, and this has been most helpful.

Once a site has been selected, it will be examined carefully to find out which parts of the site should be harvested and how often it should be harvested. The frequency may vary from several times a day to once a month. News sites will often have archival functions for articles that will be collected through snapshot harvesting; therefore, only the 'front pages' (and sometimes a few levels below that) of the website need to be harvested frequently. Selective harvesting requires close monitoring by the staff to find the sites and to determine frequency as well as depth of harvest — content as well as archival function must be taken into consideration.

3. Event Harvesting

The idea behind event harvesting is to collect web pages from new sites dedicated to one event, which are expected to disappear once the event is over and therefore will not be caught in the snapshot harvesting. We have defined an event as something that:

- creates a debate among the population and is expected to be of importance to Danish history or have an impact on the development of Danish society
- causes the appearance of new websites devoted to the event
- is dealt with extensively on existing websites.

Some events we know of in advance, such as elections or important political meetings that are taking place in Denmark, the 2009 Global Climate Summit in Copenhagen. Others, such as the newspaper cartoon crisis (2005–06), just happen unannounced, and we have to act quickly to catch the web pages generated. One of the lessons we have learned is that unplanned events may take many forms and that we should not try to categorise them but stay alert and collect what look like temporary web pages. This type of work is difficult to standardise and it requires staff that is tuned to this type of collection, but if you have such staff (and we are fortunate to have them) it works quite well.

The other Nordic countries (Sweden, Norway, Iceland and Finland) are now carrying out cross-sectional harvesting, and Sweden is the most experienced having harvested TLD SE since 1997. Sweden's legal deposit legislation does not yet allow web harvesting and therefore special governmental permission was needed to start collection.[4] The other Nordic countries have legal deposit acts that permit web harvesting. Practices elsewhere in Europe vary from cross-sectional harvests to harvesting selected web sites. (Here 'selection' refers to a collection building strategy and not to a harvesting strategy).[5]

We are often asked questions about the description of the harvested materials. Currently we only do indexing on demand, so the entire archive is not indexed. We will be implementing a 'way-back' type of access next year. In the longer run (depending on the resources allocated) we would like to develop more sophisticated tools for searching. Our overriding policy for description is that is has to be automatic, not manual — in other words we are not producing, and will not produce, library catalogue records (e.g., MARC records) for our archive. We would like to harvest embedded metadata for more specific searches and also utilise the tools which we are creating in connection with harvesting, especially selective and event harvesting. For instance, when harvesting election websites, we categorise the various types of websites and make lists according to these categories. This has a practical purpose, namely distributing the work among the staff, but this categorisation could easily be used in searching the materials afterwards.

At the moment a possible national strategy for cataloguing web resources is being debated in Denmark. The National Bibliographic Agency (a private organisation) is currently cataloguing about 2–3,000 web publications a year, mostly resources which are analogous to printed materials, and they would like to continue doing so. The institutions behind Netarchive.dk, on the other

hand, would like to see automatic cataloguing tools developed which would enrich metadata provided by the producers of internet materials to help our users obtain more accurate search results. The National Bibliographic Council expects to take a stand on this issue in the fall of 2008.

National vs. International Collection

The nature of the internet is such that even a complete national web archive will not provide users (present and future) with the experience that users of the live internet get, so interoperability among web archives is a necessity. This requires political and legal actions in each country. Ideally, if all countries (and international organisations) would collect 'their' part of the internet, using compatible tools and providing seamless access, we would have a virtual internet archive. A survey last year among national libraries all over the world revealed general support for the goal of a global virtual net archive[6]; however, that goal is not quite within reach at the moment. Currently (May 2008) some fifteen European national libraries are harvesting their national domain, either partially or fully, and another nine are planning to and/or testing. To this can be added institutions outside Europe such as The National Library of Australia, Library and Archives Canada, Bibliothèque et Archives nationales du Québec and the Library of Congress. Still, only a minority of possible collectors of the World Wide Web are involved in actual web archiving.

Meanwhile, some countries, like Denmark, will collect relevant web pages outside the national TLD. We collect some 40,000 websites from other TLDs like .com, .org, .nu etc. These are found partly by manual searching (Danish place names, Danish trade terms etc. outside .dk), partly by automatic searches (servers located in Denmark by IP number and an automated geographical tool location check (Geo-IP)) in order to comply with the provision in the Legal Deposit Act that states that we should collect materials 'published from other Internet domains etc. … directed at a public in Denmark.'

A final issue in collection building of internet materials can just be touched upon here, namely, the different needs of users. As a national library, our focus is on documenting the national heritage and providing research materials for scholars who need this type of material. However, researchers also use material which is not part of the national heritage and they need the entire internet to be preserved as documentation of their findings — hence the need for a global virtual archive and for seamless access to this.

Access to Web Archives

So far, I hope to have shown that principles for collecting internet materials in many ways can be based on previous practice in building a national collection. Rethinking is necessary when it comes to the issue of description but, again, cataloguing expertise can be utilised once we abandon the dogma that only cataloguers can describe information resources, and instead apply our cataloguing expertise to improve information produced by non-librarians. Technical issues of harvesting and archiving should not be ignored, but in here also, problems are identifiable and can be solved, not the least through international cooperation, such as the IIPC which is making good progress towards solving technical issues.

Quite another matter is *access* to the archived material. This has become the major challenge to web archives, especially those who are involved in snapshot harvesting. Legal obstacles prevent the general access that national libraries want to offer and indeed are accustomed to offer their users. In Europe the obstacles are

- the EU Directive on Data Protection (Directive 95/46/EC)
- the national data protection legislation based on this directive
- administrative law
- copyright law.

Copyright law is in fact a minor obstacle, as librarians and rights holders by now have much experience in agreeing on conditions for access and copying, agreements that are of mutual benefit. A far more serious obstacle is the Personal Data Act which in Denmark limits access to the collected resources to scholars, and for purposes of research and statistics only. The argument is that the harvested material may contain sensitive personal data and should not, therefore, be generally accessible according to the EU Data Protection Directive (full name: Directive 95/46/EC on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of such Data), as interpreted by the Danish Data Protection Agency. Norway[7] and Sweden[8] face similar obstacles in giving access, while Iceland[9] hopes to be able to give general access as does Finland — with due regard to copyright law.[10]

This restriction goes against the aim and purpose of a national library, which is to give access to the cultural heritage of the nation to its citizens, with restrictions determined by preservation considerations and copyright. It also goes against the important aim of any democracy of giving its citizens the possibility to be informed. An obvious example concerns resources related to elections. Any Danish citizen should be able to see the web pages generated during the election campaigns that we have archived and hold the elected politicians accountable if she or he so desires. We have archived two national elections in 2005 and 2007 and one local election in 2005. However, due to the legal restrictions, only researchers are allowed to see this material.

The basic issues behind this legal problem are both formal and political. The formal issue is whether Directive 95/46/EC should be applied to web archiving. The Danish Data Protection Agency has decided that is should. The political issue is the balance between protecting the individual and protecting the public's right to know.[11] At the moment the scale is heavily weighted on the side of the former (protecting the individual), and it requires political action to change that balance. Here I see a task for LIBER: to put pressure on the EU as well as on national governments and politicians.

It is not an easy issue for several reasons. For one, a lot of people, especially young people, are willing to publish private data on the web that they may later regret. For another there are also people (young and old) who are willing to publish private data (including photographs) of other people on the web against their wishes. These data are easy to search and to combine with other data to reveal more about a person than that person is willing to disclose (for example, you may not want your employer to uncover your leisure time activities or family parties). Some would argue that we, therefore, must keep the web archive closed, in order to protect people.

Countering this argument is the need for the public to be kept informed and the right of the public to have access to information, once it has been made public. Added to this is the need for the cultural online heritage to be preserved, just like the printed heritage is being preserved and has been preserved for several centuries, especially in view of the fact that more and more public communication is migrating from printed media to online media.

Netarchive.dk has tried to find technical solutions for the problems of access. In 2004, when we became aware of the legal obstacles after the Data Protection Agency had responded to the draft of the new law, we thought of two possible solutions. One was to classify certain websites as safe, i.e., not containing sensitive personal data. This might apply to the websites of official local and national institutions and governmental bodies, private companies with products involving patent rights as well as copyrights and sites of the news media. The assumption was that such organisations would have an interest in protecting sensitive data and therefore would know how to protect them. The problem with this approach is that is requires manual selection of websites, so Netarchive.dk would have to apply lots of resources to survey the TDL DK for new websites. We do not have a special domain (like .gov) for public websites to ease the task. With the rapid growth of the Danish internet it has quickly become apparent that this is a dead-end approach — quite apart from the fact that several incidents have revealed that even public websites cannot be trusted to hide sensitive data. There have been cases whereby official bodies have published on their website minutes of meetings containing sensitive personal data that should have been deleted before publishing the minutes on the web.

The other solution was to develop a computer programme that could identify sensitive data. However, the studies we have conducted so far revealed that even the most neutral and non-sensitive personal data can become sensitive when the context is considered or when searches make it possible to combine data.

Conclusion

Based on the Danish experience with web archiving I would argue that while we, as European national and research libraries, have solutions for formulating collection policies with regard to internet resources as well as techniques for implementing these policies, and while we also know how to preserve the materials collected, we still face a major, as yet unsolved, problem when it comes to giving access to these resources.

This problem is particularly alarming for national collections which should be available to all citizens. It is not a technical problem and it cannot be solved

through technical means. It is a political issue, and thus it can only be solved through political action. I therefore urge LIBER to assume the task of helping to solve that problem through political action.

Websites Referred to in the Text

Netarkivet.dk, http://netarkivet.dk/index-en.php

Notes

[1] *Den trykte kulturarv: Pligtaflevering gennem 300 år*. Redigeret af Henrik Horstbøll og John T. Lauridsen. With an English Summary. Danish Humanist Texts and Studies, Vol. 16. Copenhagen, Det Kongelige Bibliotek og Statsbiblioteket; Museum Tusculanum, 1998.

[2] Cited from unauthorised translation of the act, found at http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html

[3] Currently (July 2008) http://www.dk-hostmaster.dk/

[4] http://www.kb.se/soka/internet/sv-webbsidor/om/

[5] More information and links to web archiving institutions at http://netpreserve.org/about/members.php

[6] Grethe Jacobsen, 'Webarchiving Internationally: Interoperability in the Future?' Paper completed April 2007 and presented to the World Library and Information Congress: 73rd IFLA General Conference and Council, Durban, South Africa, 19–23 August 2007. Found at http://www.ifla.org/IV/ifla73/papers/073-Jacobsen-en.pdf (English); http://www.ifla.org/IV/ifla73/papers/073-Jacobsen-trans-fr.pdf (French); http://www.ifla.org/IV/ifla73/papers/073-Jacobsen-trans-es.pdf (Spanish). Revised version published September 2007: 'Webarchiving Internationally: Interoperability in the Future? Results of a Survey of Webarchiving Activities of National Libraries, March 2007'. http://netarkivet.dk/publikationer/InteroperabilityInTheFuture_IFLA2007.pdf

[7] Kjersti Rustad, 'Our digital heritage as source material to end-users: Collection of and access to net publications in The National Library of Norway',*Journal of Digital Asset Management* (2006) 2, p. 172–177.

[8] http://www.kb.se/english/find/internet/websites/ (information in English); http://www.kb.se/soka/internet/sv-webbsidor/ (in Swedish).

[9] http://vefsofnun.bok.hi.is/index.php (about plans for web harvesting and access; in Icelandic).

[10] http://www.finlex.fi/sv/laki/alkup/2007/20071433 (the law, Swedish version).

[11] I am grateful to Charlotte Bagger Tranberg, Aalborg University, and Hanne Marie Motzfeldt, Aarhus University, both legal scholars, who have helped me (and several others) understand the legal issues involved.