Experiences and Conclusions from a Pilot Study:
Web Archiving of the District and County Elections 2001

Final Report

for

The Pilot Project "netarkivet.dk"

Contributors:

Birte Christensen-Dalsgaard, SB
Eva Fønss-Jørgensen, SB
Harald von Hielmcrone, SB
Niels Ole Finnemann, CFI
Niels Brügger, CFI
Birgit Henriksen, KB
Søren Vejrup Carlsen, KB

Danish Version: September 2002

English Version: February 2003

Contents

List of Figures:

List of Tables:

1 Abstract

In the introduction to our application we wrote:

> *Over the last decade or so the Internet has become one of the mainstays of our society's communications infrastructure. For a rapidly growing part of the population the net is now a daily-used tool, serving variously as a source of news, a reference work, a library or knowledge archive, as a public forum for discussion or a postal system. It is used alike for business purposes, for communication between central or local government and the individual citizen, and between private individuals in the public or semi-public arena. Just as private and public institutions and the majority of firms have by now established their presence on the net, so many interest groups, associations and individual citizens have set up their own web sites and web pages, and many private individuals are regular participants in several of the numerous public and semi-public chat forums and news groups. Thus the Internet today is a medium at once for the state authorities, for the private market and for civil society. It is a place to seek information, enlightenment, entertainment, and diversion, a place to trade and a place to meet other people.*

The present report by the group behind the ”Netarkivet.dk” project describes the experience gained from a pilot study, in which existing software was used to harvest and subsequently test out materials relating to the County and District elections of 2001. The pilot study showed that a great deal of material could be harvested in this way, but also that much of the interactive use of the net cannot be caught by ordinary methods.

The pilot project also offers an indication of the financing needed if Denmark is to safeguard an important part of its cultural heritage. Estimates are given both for the archiving of this heritage under present conditions, where the work is carried out on the basis of voluntary agreements, and on the assumption that the law on legal deposit of material may be changed, making it legal for institutions receiving statutory deliveries to acquire online materials.

2 Introduction

The present project embarks from the view that the Internet today is the medium that gives the most comprehensive, multi-dimensional and accurate picture of contemporary cultural life, and contains some of the key sources for understanding the modern network society.

If we understand our cultural heritage to mean "that part of a generation's established customs, in terms of outlook, lifestyle, social conventions, aesthetic and other artistic norms and forms of expression, that is wholly or partially adopted by succeeding generations, sometimes for a sufficiently long duration that it may be considered in a historical perspective"[1], then the Internet is undoubtedly one of the most original contributors to that heritage today.

But even as the Internet, both as a medium of communication and as a repository of knowledge, is growing daily more important, a very large proportion of the material published on the net is disappearing with disturbing rapidity. Several studies have shown that 40% of the material on the net disappears within one year, while a further 40% is altered, leaving only 20% in its original form. Other studies indicate that the average lifetime of a web page is 44 days.[2]

As a result we may soon find ourselves in the position that, when it comes to writing the history of our times, a significant part of the source material will be missing.

In certain respects we are already in that position. Important parts of the source materials documenting the way in which the Internet came into being and developed from the 1960s to the present have already disappeared. Similarly, the history of the Internet's breakthrough in the early 1990s will have to be written without recourse to most of the original source materials. The same applies when it comes to describing the dramatic rise of net trading and 'dot.com' companies in the mid.1990s, and their subsequent decline immediately after the millenium. And it is also true of the Internet's significant role as a source of news and a medium for public debate: here too a substantial proportion of the source materials must be written off as lost.

Other such materials will likewise be lost in the near future unless an effort is made now to start acquiring them, and still others will exist only as a more or less haphazard, unregistered and inaccessible form of publication in the archives of institutions, organisations and companies. Any material that is handed over for archiving will represent only an arbitrary selection from the material found in random and publicly inaccessible places, and may well be in a form that is no longer usable, because the hardware and software needed for reading it is no longer available.

The overall purpose in establishing an Internet archive is thus to ensure the preservation of this contribution to the cultural heritage and thereby the source materials that will provide the foundation for future research not only into the Internet's own history, but also into all the ever more comprehensive cultural, institutional and business activities that take place especially – and in some cases exclusively – on the Internet or in close connection with it.

---

[1] Den Store Danske Encyklopædi vol. 2, p. 33, column 2.

[2] See among others Peter Lyman et al. *How much information*: http://www.sims.berkeley.edu/how-much-info/; Johan Mannerheim, *The WWW- and our digital heritage*, http://ifla.org/IV/ifla66/papers/158-157e.htm

However, preserving the products of such web activity is neither technically, logistically nor legally straightforward. Several intitiatives towards this end have been undertaken internationally, each with its own strengths and weaknesses. Three strategies in particular have been pursued: an event-based strategy such as is practised, for example, by the Library of Congress (dealing with such major events as the Presidential elections or 11th September); a selective method, as practised for example by the Australian National Library (Pandora), and finally a cross-section approach, as practised for instance by National Library of Sweden. None of these initiatives however has attempted to consider their harvesting strategies and results in relation to the anticipated needs of researchers. The aim of the Danish pilot study is to obtain through practical experience a proper technical understanding of the problems involved, and to test the material obtained in relation precisely to such concrete research needs.

There are a number of legal problems associated with web archiving. Different approaches to these have been taken in the various countries where experiments with archiving have been conducted. Some countries have aimed to secure a comprehensive legal framework before undertaking any initiative; others have embarked on the basis of the given country's existing laws and have dealt with particular problems as they have arisen.

An important aspect of the present project was to elucidate the legal problems involved. On the one hand, there is the question of whether web archiving can be handled within thepresent framework for legal deposit of publications; on the other, there are problems concerning intellectual property rights and the treatment of personal data, to which solutions must be found.

An important element in the project, therefore, was also to find out

- Whether it is necessary to establish archiving agreements.
- How such agreements can be set up.
- What types of access can be offered – and
- to what extent such agreements can meet the needs of researchers.

The concrete purpose of the present pilot project was to test out various methods for selecting, acquiring and archiving Internet materials within an area that was both relatively limited and of undoubted interest to a broad swathe of the ordinary public.

The field selected was that of net activities relating to the Danish local elections of November 2001. This was considered an appropriate field for a number of reasons:

1. It was anticipated that the local elections would give rise to a broad range of dynamic and interactive net activities, thus presenting the project with a number of important, new and hitherto unsolved archiving challenges.
2. The elections involved a series of markedly differentiated regional events involving participants throughout the country, and were therefore well suited to highlighting the importance of a national Internet archive.
3. The activities in question were limited in time, and could therefore be followed and archived from start to finish.
4. The material gathered would be thematically consistent, which would make it possible, as part of the project, to test out and judge its research value.

The focus of the project was on harvesting – on what materials should be acquired, how this should be done and how the necessary agreements could be secured. Obviously, in order to

give a proper answer to these questions it was necessary to consider the *outcome*: e.g. to consider the question of eventual access to such materials by the public. The project planned to use the access software developed in connection with the Nordic NWA project [NWA]. Unfortunately the launch of this software was delayed, which had an impact on the project. The long-term storage of materials was not part of the remit of this particular project; nevertheless, it is clear that this point too needs to be addressed.

A series of reports were made while the work was in progress; the following can be found via netarkivet.dk:

Interim report 1: Strategy for gathering, monitoring and archiving [report1]

> ***Abstract***: The archiving of Internet materials demands strategic thinking on what materials should be gathered and how. The report offers both a media-research and an archiving/technical approach to the problem. Three programmes (The Legal deposit System, NEDLIB and RoboSuite) are identified and their use both in downloading, following different strategies, and their support for presentation are discussed. On the basis of their respective strengths, and in order to test out various strategies, a downloading matrix was established with harvesting frequency as one axis and the programme as the other. The choice of materials was made on the basis of four parameters: type of participant, demographic criteria, type of communication and type of file. Finally a monitoring strategy was established.

Interim report 2: Experience so far with regard to harvesting [report2]

> ***Abstract***: This report describes activities up to the local elections of 20 November 2001. The focus is on work with selected relevant URLs (an absolutely essential activity if one wishes to do event-based harvesting), work on gaining permissions to harvest material and, where relevant, to make the material acquired publicly accessible; experience of configuring harvesting applications and the first experience of using these applications.

Interim report 3: An observation report [report3]

> ***Abstract***: This report describes net activities in connection with the local elections in November 2001. It begins by describing the nature, extent and significance of this event on the Internet, and concludes that in 2001 the local election campaign on the net reached a level and acquired a significance that represented a breakthrough comparable to that achieved in the first publicly recognised Internet campaign during the local elections of 1997. Second, it contains a brief description of the strategies that some of the existing media, for example, used on the Internet.

The present report supplements the previous ones, particularly in three areas:

- Analysis of results (sections 4 and 5).
- The section on Danica (section 6).
- The presentation of financial models for web archiving work.

3 Strategies and methods

3.1 Web archiving

Several different strategies and methods for acquiring materials from the Internet have been put forward. Below we briefly define these and discuss the advantages and disadvantages of each.

3.1.1. Harvesting strategies

- Snapshot: The idea with snapshot archiving is to save at well-defined intervals a reasonable quantity of publicly available material. Examples of this approach are the Internet archive in the USA and the strategy of acquiring all radio broadcasts in a particular week here in Denmark.
- Selective: The value for posterity of particular objects, works or total web sites is evaluated and those that are considered valuable are preserved. An example of this approach is the Pandora initiative in Australia.
- Event-oriented: When a given event is judged significant, materials relating to it are preserved so that both the event itself and the reactions to it can be followed by posterity. Examples of this approach are the "September 11" archive, the archive on the presidential election in the USA and the pilot study on the local elections here in Denmark.[3]
- (Statutory) delivery: Categories of material that must be delivered to certain libraries are defined by legislation. The publishers decide whether a given material should be delivered according to the law. The difference between this approach and the selective approach is that, in the case of legal deposit, *types* of material are defined as worthy of harvesting, rather than individual documents. Government publications are an example of one type of material that is currently subject to harvesting by this method, whereas online newspapers are not at present covered by the law on legal deposit

3.1.2 Methods of harvesting

When we talk of active or passive harvesting, it is important to distinguish whether we are speaking of an initiative in relation to the selection/discovery of material, or referring to the method whereby the material is acquired. In what follows we are referring to the method of harvesting, rather than to the way in which the material is discovered or selected.

- Pull: The material is gathered manually or automatically either by using a harvester or by some other method.
    - Manual harvesting: Individual files are gathered and put into an archiving system. One example is pligtaflevering.dk, where individual works are designated for archiving and acquired.
    - Automatic harvesting: Materials are harvested automatically on the basis of criteria such as domains or a series of manually or automatically generated URLs, and automatically included in an archive. Examples of this are the web archiving projects Kulturarw3 in Sweden and the Internet Archive in the USA.

[3] The web addresses of these archives can be found in the List of References

- Push: The material is delivered or donated to the archive.
    - Delivery: The publisher is responsible for delivering the material in a manner agreed in advance, in certain cases including a description of the material. The materials delivered to the State archives offer an example of this.
    - Donation: A collection of material is donated. In many cases this will subsequently need to be structured.

3.1.3. Evaluation of harvesting methods in relation to type of materials

Different types of material lend themselves to different strategies of harvesting. Different countries have chosen different methods. As was made clear at a web archiving conference held at the Royal Library in Denmark in June 2002, the choice of method tends to be based more on financial than on technical considerations.

The following table gives an overview of what different countries and institutions do in this regard:

	Event-based	Selective	Snapshot
Active (automatic and manual harvesting)	Strategy in USA, Library of Congress	Strategy of Pandora, Australia. Strategy of pligtaflevering.dk	Kulturarw3, Sweden EVA, Finland Archive.org
Passive (delivery and donation)		Electronic journals at the Royal Library, Holland	

Where the selective strategy is used, a very detailed description of the selected subjects is given, whereas in the case of event-based harvesting, such as is practised by the Library of Congress in the USA; the library expects to offer only a description of each event. The retrieval of material relating to the event will be carried out in a different way.

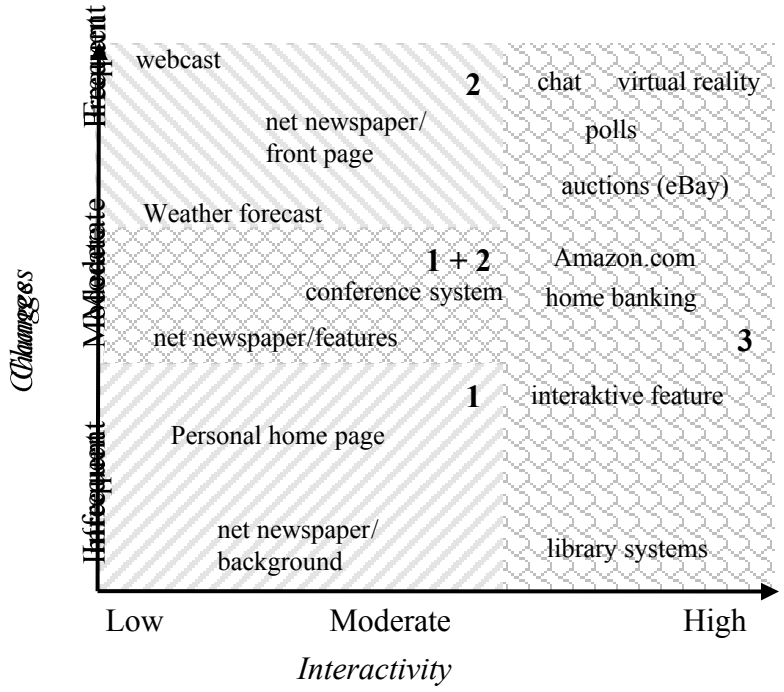
With the American strategy, therefore, the library system would offer a description of the event as a whole, but not of the individual components.

In the case of snapshot gathering it is impossible to give a description of the individual sites other than that offered by the site itself. An eventual adoption of the 'semantic web' (see e.g. http://www.cs.concordia.ca/~faculty/bcdesai/grads/haddad-thesis.pdf) may lead us to revise this conclusion, in so far as the site itself will carry relevant metadata. Today however it is necessary to use alternative strategies such as free text searching, which is currently being tested out in connection with NWA.

Different types of information lend themselves to different strategies and methods of harvesting.

Two important parameters in connection with the strategy for acquiring web materials are the frequency of updating and the degree and complexity of interactivity at a given site. Depending on these factors, different strategies of harvesting should be chosen, as illustrated by the figure below. Area 1 is the area in which we would expect snapshot gathering to be the best strategy. Section 2 is the area where we would expect a selective model to apply best.

Finally there is the area where we would not expect to be able to use either the selective or the snapshot method, because the material is not of a kind that can automatically be acquired. This area is a candidate for the so-called filming method, in which only certain 'scenes' illustrating its function are preserved.



Figur 1 The applicability of different harvesting approaches as a function of the rate of change and the level of interactivity. On the figure 1,2 and 3 stands for: 1: Snapshot, 2: Selective and 3: None of the traditional approaches works.

It should be pointed out that in many cases the various components of a given site will belong to different areas in the figure above. Here it will be a matter of individual evaluation to decide on the appropriate method or combination of methods.

3.2 The Local Elections on the Net

3.2.1 The nature, extent and significance of the event

The decision to use the Danish local elections of 2001 as the target of this pilot study was based on the following considerations: First, that they represented a discrete event that could be followed from start to finish; second, that the event was of both local and national significance; third, that it could be expected to give rise to widely ramifying and comprehensive activity on the net; and finally, that the various forms of these activities would include those that are most complex to deal with from the archiving point of view. On the basis of the observations carried out concurrently with the archiving, we can conclude that all these presuppositions proved correct.

These presuppositions rested in turn on the assumption that it was probable in this instance that there would be some kind of breakthrough in the use of the Internet for political electioneering. This assumption, too, appears to have been correct, if one looks at the amount of resources used, the number of people involved, the wide-ranging nature of the activities and the references to the Internet in the election material put out in other media. It is not part

of the remit of the present project to analyse the extent of traffic on the relevant sites, and as far as we know there is no material gathered anywhere else that would indicate the extent of such use.

Although these presuppositions were fulfilled, however, events in a number of other respects did not work out as expected.

First, the election campaign on the net began unexpectedly late. It is evident that one of the main reasons for this was the overwhelming impact of the September 11 terror attack in the USA, which was the main focus of public attention for a long time afterwards. Second, the local election campaign had scarcely begun before (on 30 October) the Danish government decided to call a general election, to be held at the same time as the local ones. Since the archiving was based on a database that had been built up before the parliamentary elections were called, and since many of the actors involved (both in the various parties and the media) were the same, the material gathered included a great deal of material on the general election.

Moreover, it would be reasonable to assume that activities relating exclusively to the local elections were less extensive than they would have been if the general election had not been called at the same time. However, the fear that the local elections of 2001 would be entirely swamped by the parliamentary elections proved unfounded, since many of the large-scale media simply divided their web sites into two sections dealing respectively with the different elections.

The impact of the general elections on the local elections is described in more detail in Interim Report 2.

Despite these unforeseen upsets the Internet campaign for the local elections of 2001 was sufficiently extensive and distinctive as to suggest a breakthrough comparable to the first publicly recognised net campaign, which occurred in the run-up to the local elections of 1997. The net campaign on that occasion has not been thoroughly described in the relevant literature, and since such material as existed would be missing today, there is no possibility now of describing it more precisely.[4]

There exists, however, a thorough description[5] of what can be assumed to be the most significant activity in this connection, namely the initiative undertaken by the organisation Kommunedata (KMD in the following) to establish a countrywide election platform, offering six elements for each district: 1) Presentation of parties; 2) debate; 3) overview of newly-elected politicians; 4) results and prognoses; 5) information about the election law; 6) statistics etc. A subsequent report concluded that ”the attempt was a fiasco” and that the reason for this was above all that ”far too great an extent the definition and formation of the democratic content in the concept was managed by the organisations involved: Local Government Denmark (Kommunernes Landsforening), the KMD, and Mostrup Publishers, and far too little by those whom democracy is actually about: Namely, the citizens and the politicians.” (Hoff et al, p. 100). The report also laid part of the blame, however, on ordinary citizens, politicians, the education system and local party organisations. In a book entitled *Kommunalvalgene i perspektiv* (The Local Elections in Perspective), which came out shortly

[4] A search in Archive.org’s ’way-back-archive’ of party web siteweb sites relating to the Social Democratic and Venstre parties, and of the media sites dr.dk., tv2.dk and politiken dk., yielded only four pages from 1997 (all belonging to the Venstre party, with one page per quarter year). Search carried out on 14.3.2002.

[5] J. Hoff, K Löfgren & S. Johansson. *Internet og demokrati - Erfaringer fra kommunalvalget 1997* (The Internet and Democracy: experiences from the Local Election Campaign of 1997), Jurist- og Økonomforbundets Forlag, København 1999

before the local elections took place, election researcher Roger Buch, in an extension of the argument put forward by Hoff et al., questions whether ”the net campaign in its present form is not fundamentally a contradiction in terms” since election campaigns generally involve one-way communication, while ”the essence of the Internet is …the opposite – here it is the users and recipients that are in control.[6] One needs actively to seek out a home page or a debate forum”. (Buch p. 129). Buch devotes fewer than four pages to the subject and does not hold out any real prospect in future for election campaigning on the net, concluding that votes cannot be shifted through net activities. Such activities certainly do no harm, but they are presumably directed only towards those who are already active, and who perhaps have already made up their minds; no great resources, therefore, should be expended on such efforts (pp. 130-131). Nevertheless it is open to question whether the ”action requirement”, which Buch refers to in relation to Internet activity, is different in kind from, for example, that involved in taking out a subscription to a newspaper, reading an article, or going to a meeting.

As far as the use of resources is concerned, no heed was paid to Buch’s warning. Apart from the private resources devoted to the campaign by thousands among the altogether 16,000 candidates that stood at the election, and by their parties and local party branches, all the major media in the country invested quite substantially in building up the local election platforms. Even more surprising, perhaps, is the fact that considerable imagination was shown in designing the individual web pages in as lively and attractive a way as possible. The means employed included not only the usual interactive functions (e-mail and mailing lists), but also numerous other forms: Quick polls, attitude tests, robot services, photo satires, games, competitions, get-a-poster-via-the-net, send-an-SMS, send-a-postcard, screen savers, banner advertisements on heavily used portals, video- and audiostreaming, WAP-materials, as well as numerous kinds of debate forums and chat activities.

The number of participants involved in each such debate forum was relatively modest, and the forums varied considerably in significance. The ”nationaldebat.dk” forum, which in November 2001 claimed that it had had some 50,000 visits and 4,000 contributions (2,000 of which were in the Politics subject group), had since August 2000 been the site of intense discussion as to which candidates and parties to support during the election, and how. Here, the site in question was a pre-existing forum, bringing together politically and ideologically like-minded people, who during the election then used it as a medium for political debate and as a tool for organising and activating grass roots support. The large nationwide media similarly set up debate forums around the local elections in 2001, which in these cases functioned more or less like the letter pages of newspapers, but in a more extended form, with more reciprocal debate and a more distinct thematic focus. Meanwhile, state-initiated projects such as ”Nordpol” in Nordjylland County offered a somewhat different model, designed to broaden political involvement; similarly, individual districts (Odder, Køge) and party branches in various places throughout the country experimented with local debate forums.

Debate forums obviously have an advantage over chat sites in that, since the debates are archived, participants can enter them at any point; they also have the opportunity to rethink or correct their contribution, whereas chat debates require the contributor to be present here and now, and tend to show signs of forced tempo, with short sentences and answers that always limp one or two points behind the question they are answering. For a chat forum to function,

[6] R. Buch. *Kommunalvalgene i perspektiv* (The Local Elections in Perspective). Odense Universitetsforlag, Odense 2001

the participants obviously have to be accustomed to the particular conditions and possibilites that this form of communication offers.

As far as the send-a-postcard function is concerned, one candidate drew attention to herself with a virtually nude picture, which together with a humorous text, went the rounds of the international press and provoked – albeit somewhat equivocal – comment, even in serious publications such as the German *Der Spiegel*. The Venstre candidate in question, who also incidentally conducted a very active campaign outside the net, was later elected to both the City and the County council.

This is perhaps the only example that indicates a direct connection between net activity and an election result, but this is not to suggest that the net did not in other respects have a quite considerable impact on the course of the election campaign. Whereas in 1997, according to the data available, there appeared to be only one key site relating to the local elections, which had been set up for strategic purposes on the initiative of a public body, KMD. In 2001 there was a whole series of such sites, initiated not just by public bodies but by the printed media (both national and local), by the electronic media (especially DR and TV2's regional stations), and by the pure net media (city portals).

In 2001 the public institutions (particularly Nordjylland County and Bornholm) in fact made just as great an investment in the net campaign as they did in 1997, but their initiatives represented a much smaller proportion of the whole picture, even though they had evidently learned from the KMD's experience and put great weight this time on attracting potentially interested participants (mainly among young people) in their preparations, and on ensuring that they were mentioned by the other media and cooperated with them.

As indicated above, we had no opportunity in this project to monitor the amount of traffic or traffic patterns, and there is no systematic and reliable information available about the extent to which the various web sites were used, and only sporadic indications as to who the users were. Such sporadic knowledge as there is stems from the debate forums, where you can register the number of contributions to the debate, and a couple of chat forums where you can see whether the participants were politicians (e.g. election candidates) or citizens (but not whether they were party members or representatives of interest groups). Thus there is no reason to doubt the general thesis that in this election the net was used mainly by those who were most interested and involved; however, this category of 'interested parties' includes not only politicians but, just as importantly, journalists, who made considerable use of the net and also referred to it frequently as a source of information. Thus well before the election the television station TV2/Fyn made available on the net a great deal of their source material on the key election issues for every single district in Fyn. It has not been possible to conduct a systematic evaluation of the extent to which net activities were duplicated in the other media. Nevertheless, the monitoring that was conducted simultaneously with the harvesting of material suggests that most of the printed media made several different references daily to election-related web sites.

3.2.2 Net strategies

From our monitoring of web sites relating to the local elections, carried out continuously over a six-week period in selected districts and counties, we can conclude that, despite the numerous actors involved, most districts had one – perhaps two – key sites that were the central forum for the constituency concerned, while individual pages and local party pages were for the most part rather dull, and bore the distinct mark of not having been properly

edited. Notwithstanding the conclusions of election researchers, and without disparaging the value of the candidates' individual web pages, it is indisputable that one cannot build up an attractive web site without investing considerable editorial energy in updating the content daily and in using various enlivening features. A successful web site not only attracts attention but also maintains and builds upon it. A photo, text and e-mail address are simply not enough. The significance of key sites in the local elections can hardly be overestimated. There needs to be one place where users can expect to find a properly assembled introduction to the event. This is also vital for journalists and for establishing connections with other media.

Likewise it is noteworthy that among the electronic media it was only the big national TV stations that took major initiatives on the net, while in the print media, too, the national papers appear to have put out a great deal more material on the net than local newspapers, which only occasionally played a role in initiating a key site. If this observation is correct, it might indicate that certain new kinds of interaction between the print and net media are now coming into being, with the national print media seeking to use the net to distribute locally differentiated materials and in this way compensate for their Achilles' heel: the coverage of local events.

Our observations also gave a clear indication of the importance of tailoring net activities to fit other election campaign activities, and, equally, of ensuring that net activities break through to other media platforms.

3.2.3 The monitoring procedure

The purpose of the monitoring procedure was to reveal, analyse and document all the Internet activity relating to the election, primarily going by district and county, but also looking at the way in which the local elections were reflected at the national level.

The monitoring was carried out with the help of six students who were divided into two groups, two of which worked primarily at the district/ county level, while the third monitored activities at the national level and took care of certain specific tasks. The monitoring proceeded as planned, increasing in intensity from 30 to 45 hours per week as the time of the elections approached.

Since the aim was primarily to acquire representative materials, no attempt was made to monitor activities throughout the whole of Denmark. The criteria used in building up the initial base were therefore supplemented with two sets of criteria for selecting the particular districts and counties that should be observed in more detail. The main goals were to ensure broad coverage of the different levels of activity, and of the different forms (or funtionalities) that this activity takes. These criteria for selection were applied by using a specific set of variables in each case:

A: Variables with regard to the level of activity:

- Conscious effort to use IT in relation to the local elections 2001 (great/little anticipated activity)
- County/town/suburb/metropolis
- Centre/periphery
- Concentration of particular demographic groups (sex, ethnicity, age…) (great/small)
- Main occupation

- Turnout in last election (high/low); partly in general/partly in demographic terms
- Access to Internet (great/small)
- Overlap between counties and districts
- Existing local/regional press (yes/no)
- Existing local/regional radio (yes/no)
- Existing local/regional TV (yes/no)

B: Variables with regard to function

- hat
- Usenet-groups/news groups
- Discussion forums and similar (e.g. on other sites)
- Streaming (live and on demand)
- Circular e-mails, newsletters, press releases
- Password protected areas
- Wap/pda/sms
- Java scripts

On the basis of these two sets of variables an initial observation was conducted of all 275 districts and 14 counties with the aim of getting a general overview. Against this background 133 districts were then selected as being most suitable in terms of the different variables (at this point none of the counties was de-selected). This group of districts and counties was then subjected to a further critical review, which resulted in the choice of a net group of 58 districts and eight counties. The districts and counties in this net group were then observed more closely for the first fourteen days, and on the basis of these observations a final choice of 24 districts and seven counties was made, which would constitute the target for more in-depth monitoring over the remaining observation period.

Thus the observation targets, which included unforeseen activities (e.g. banner advertisements, e-postcards, shoutboxes, SMSs, e-trade, quick polls, robot answers, calculation of election odds and games) were adjusted week by week, and apparently low levels of activity, e.g. on city-nets and debate forums, were subsequently checked. In addition, the work focused on identifying possible centres and 'key sites' in the individual districts, and on the use of interactive features (postcards, quick polls on electoral themes, and so on). In the final part of the monitoring procedure the focus shifted to checking more special and atypical cases and the most central web sites.

Logbooks of the observations were kept throughout the monitoring period, and continuous reviews and trials of the tracking and harvesting methods were carried out (such methods included index and search machines, webcrawling, link-to-link and so on).

The results of the monitoring procedure are presented in detail in the logbooks. These results are described in Interim Report no. 2 [Report2].

4 Technical implementation

The technical procedure is described in detail in Report 2 [Report2]. Below we present a summary of the results. The following harvesting procedures were carried out:

URLs with an harvesting frequency that is: **Robot**	**Monthly**	**Weekly**	**Daily**	**Hourly**
NEDLIB	Snapshot carried out(*): 31/10-4/11 4/11-23/11 23/11-04/01	Snapshot carried out(*): 31/10-4/11 4/11- 15/11 15/11 - 23/11 23/11 - 04/01	Cumulative, in the period 23/10-3/12[7].	
WGET			Snapshot in the period 5/11-3/12(#)	Snapshot in the period 15/11-26/11
RoboSuite			Attempt at harvesting from www.politiken.dk in the period 12/11 – 3/12	

Tabel 1 Overview of different harvesting procedures. (*): All harvestings were terminated due to lack of time (#): Certain harvestings were terminated due to lack of time

NEDLIB has been developed to save a page only if that page has been altered since the last time it was saved. WGET was re-written to follow the same strategy because of shortage of space. This means that the number of pages gathered is significantly greater than the number of pages saved. The table below shows the number of saved URL objects, e.g. a picture, a text, a sound or some other object. Any given web page typically consists of many such URL objects

URL objects obtained	**NEDLIB**	**WGET**
Monthly harvesting	3.0 mill. URL objects	
Weekly harvesting	4.8 mill. URL objects	-----
Daily harvesting	2.1 mill. URL objects	15.6 mill. URL objects
Hourly	-----	9.8 mill.
Total		25.4 mill. URL objects

Tabel 2 Total number of harvested URLs

[7] Harvested in the period 23/10-3/11 via cache, e.g. these data may be defective

In the course of the practical harvesting and subsequent analysis of the material we identified a number of advantages and disadvantages with this method. Similarly, in the process of establishing voluntary agreements we realised that this procedure is extremely costly. These problems are elucidated in the following section.

4.1 Harvesting

4.1.1 Legal problems in relation to web archiving

There are three laws that are relevant in evaluating the legal framework for web archiving: The law on legal deposit, the law on intellectual property rights (or copyright) and the law on personal data protection.

Legal deposit

The law on legal deposit covers only published works. Under the law, a ’work’ is defined as a limited quantity of information that can be treated as a finite and independent unit.

A home page is not a work according to this definition of the term, but can contain or give access to a number of published works.

The law on legal deposit undoubtedly covers a number of the works that were copied and archived as links in the web archiving project. The remaining content of home pages falls outside the remit of the law on legal deposit, and cannot therefore be copied and archived on the basis of the existing law.

The definition of a ’work’ given under the law on legal deposit is an impediment to the development of a legislative framework for web archiving. It is therefore recommended that this definition should be removed in the forthcoming revision of the law, and that the concept of a work should be given the meaning that it has in common Danish usage, and which is also used within the law on intellectual copyright. This would mean that the work of acquiring and preserving materials could more flexibly follow the overall development of society.

Intellectual property right

Archives, libraries and museums are authorized to take a specimen copy of a work for their own purposes under the Ministry of Culture’s Decree no. 876 of 28 November 1997.

None of the purposes described in the decree can be applied to web archiving. Thus the present law on intellectual property rights does not authorize institutions to take, without permission from the copyright holder, the specimen copies harvested in connection with the web archiving project.

This means that agreements have to be made with the copyright and other rights holders concerning harvesting and archiving. Similar agreements have to be made concerning conditions of access to the material.

Handling of personal data

It is unavoidable that in the course of archiving web materials, personal data that can be attributed to particular individuals will also be acquired.

Unlike in Sweden, where a special regulation has had to be instituted to make web archiving possible, the existing Danish law on personal data does not create any real problem in relation to archiving personal data that are publicly available on the Internet. These data can as a rule be regarded as ”ordinary personal data” made available by the individual in question. The Law on Personal Data does however include a number of requirements concerning the handling of personal information, which a web archive has to live up to: E.g. concerning the correction of incorrect data.

This means that web archiving can be carried out under the present law on personal data. In order to ensure that, in the course of web archiving, personal data are handled in accordance with the law, it it however recommended that a specific clarification of this be made in consultation with The Danish Protection Agency (Datatilsynet).

4.1.2 Negotiation of permission to archive

In view of the time constraints involved in this project (the local elections would after all take place on a set date) agreements with copyright holders were made concurrently with the analysis of material. As noted above, there is no basis for legal deposit in current legislation, which means that it is necessary to establish agreements concerning archiving and use.

These issues were comprehensively dealt with in Interim Report 2 [Report2]. In this chapter therefore we present only a summary of the results.

Types of producers contacted	No.	Positive response	Reluctant response	Negative response	No response	Agreements concluded
1. Selected producers (standard agreement)	21	8	0	0	13	7
2. Selected producers (special agreement)	6	5	1	0	0	3
3. Contact on grounds of complaint	5	2	1	1	1	1
4. Donations, applications via Centre for Internet Research	3	3	0	0	0	2
Total	**35**	**18**	**2**	**1**	**14**	**13**

Tabel 3: Summary of practical experience in making agreements, August 2002. Note that certain producers come into more than one category.

Comments on the results:

As described in Interim Report 2, the producers selected with a view to contracting voluntary agreements were primarily political parties, TV channels/newspapers and producers that researchers had singled out as being of particular interest.

The overall result was that agreements were concluded with over 30% of the producers selected. The political parties in particular tended, when prompted, to respond positively, while the TV stations/newspapers were more cautious. However, it should be noted that the Danish Newspaper Association was positively disposed towards making a ”collective agreement” for all newspapers, while TV2/Fyn was similarly open to the idea of concluding

an agreement; however, there were not sufficient resources during the project period to conclude these negotiations. The political parties also tended to give permission for broad access to their materials (e.g.e.g. through all libraries), whereas the newspapers and TV stations were inclined in the first instance to give access only through the Royal Library in Copenhagen and the State and University Library in Aarhus.

Those producers that researchers had singled out as being of particular interest (e.g. Nordpol.dk and Køge and Odder Districts) were more positive, but it was frequently necessary to conduct long negotiations, partly because in certain cases special forms of agreement had to be drawn up.

In order to achieve the results above

- 156 letters or mails were written.
- 30 telephone conversations were conducted.
- Three meetings were held.

The estimated amount of work required came to approximately one man-month. Despite these letters and reminders, 14 out of the 35 producers did not respond. It should be emphasised that we decided at the outset of the project to contact only a selection of the larger producers who could be expected to be interested in concluding an agreement and to have sufficient resources for doing so.

4.1.3 Harvesting via NEDLIB

The NEDLIB robot functions by starting a series of harvesters (programmes designed to fetch web pages from given sites), all of which use a common schedule and queueing system. Every harvester acquires pages from the same site by following links on the site's pages, and will continue until it cannot find any further pages to be acquired. NEDLIB's archive is built up in such a way that one catalogue per day is set up in its file structure. Within this, sub-catalogues are established with 2000 documents per catalogue, in such a way that all harvesters archive in the same catalogue until 2000 files have been reached, whereupon a new catalogue is established, and so on. This means in practice that, unlike WGET or the legal deposit system which we have developed ourselves, NEDLIB does not acquire all files from the same site in one place; rather, the pages from a single site are spread over a series of sub-catalogues and mixed with pages from other sites. File names are changes to an MD5 code, which is generated on the basis of the URL and the content of the file. Data concerning the URL are put together with a series of technical data in a metadata file – one for each document.

This means in practice that:

- It is difficult to locate an individual URL in the archive, since one has to go through the content of the metadata files in order to find it.
- It is difficult, though not impossible, to erase all files from an individual site if the owner wishes them to be removed from the archive.
- It is difficult to work out how far one has gone in the archiving of a particular site.

If one uses NEDLIB for cumulative archiving, the material acquired is compared with that acquired at the last harvest, and if the examples harvested are identical, the latest one is thrown away. Otherwise the latest material harvested is archived in the usual way.

Two specimens are considered identical if their MD5-code is the same. NEDLIB works out an MD5-code for all objects to be archived. This stamp is preserved in a database so that

NEDLIB can check whether an object with the same MD5-code has already been archived simply by searching its own database.

Configuration problems

Proxy problems

In October 2001 the NEDLIB robot was placed within the Royal Library's network in such a way that all pages acquired would have to pass through the Royal Library's proxy. In practice this meant that there was no guarantee that the right pages had been archived unless they had *not* previously shown up on the Royal Library's network. On 4 November 2001 the NEDLIB robot was placed in a new position in the Royal Library's network, so that it no longer had to carry out harvesting via a proxy, and we chose at that point to re-start the harvesting procedure. Thus only the material gathered in the period after 4 November 2001 is included in the study.

Archiving of style sheets

Because of an error in the configuration, style sheets were not acquired, and as a result the information acquired does not always appear in our presentation system as it did in the original site. Since the access module was first introduced late on in the project, we were slow to notice this configuration error. Subsequent tests have shown that it would have been technically possible both to acquire the files and to get them to function correctly in relation to the presentation of archived materials.

Statistics on harvesting via NEDLIB

NEDLIB can be configured to acquire both static and dynamic URLs. Static URLs in this connection mean for example those without parameters, such as, e.g.,

http://www.geocities.com/selsoe/hil/station.htm,
http://www.vu.dk/webmaster/ and
http://www.tv-stop.dk/videoklip/oversigt.shtml,

while dynamic URLs refer to those with parameters, e.g.

http://www.sfvejle.dk/php/view.php?id=38&picbased=1,
http://www.sfu.dk/poldebat/index.php?id=514 and
http://www.unet.dk/dk/studenterliv/studorg.php3?todo=show&ID=7

Since we wished to make as complete an archive as possible, we decided to acquire both static and dynamic pages. This had the following consequences for the harvesting:

a) The harvesting of material took considerably longer than expected, since 60% of the material turned out to be dynamic. By deciding to include dynamic files we therefore more than doubled the number of URL objects that needed to be archived.
b) There proved to be a greater number of serious, blocking errors in this part of the software, which first had to be corrected by the developers of NEDLIB in Finland, before the process of harvesting could proceed. As far as we know, no other attempt has been made to acquire dynamic URLs using NEDLIB. This means in practice that Netarkivet.dk is the first project seriously to have used this sofware to test out this part of the active net.

URL elements acquired	#	%
Static (without parameters)	2,122,939	40
Dynamic (with parameters)	3,159,760	60
Total	5,282,699	100

Tabel 4: NEDLIB, The proportion of static versus dynamic URL elements

If we compare these figures with the Royal Library's experience in the Danish part of the NWA project, where only 36% of the 7.4 million URL elements were parametrised, it is clear that, in the Danish part of the net, it makes a difference whether one conducts an event based harvesting or a broader kind of harvesting such as that conducted by the NWA project. A reasonable explanation for this is that the kind of sites harvested in connection with the local elections are more likely to be changed frequently, meaning that in this case we are ahead of the general trend, which is to maintain web material in a Content Management System. The general development is likely to be in the direction of having more and more parametrised URLs.

4.1.4 Harvesting via WGET

WGET is an open source product that was developed via shell scripts. Starting from one or more URLs it can break up a site and acquire all the web documents that belong to it.

The programme offers several different ways of limiting the material harvested. It is possible to specify how many links one wants to follow with a given start-URL, and one can limit the harvest to a single host. One can also introduce pauses in order not to overload the site one is harvesting. In this project WGET was configured to harvest within only one host.

The WGET programme is single threaded e.g. it acquires one URL object at a time. In order to harvest sufficiently quickly it was necessary to run WGET in more than one instance at a time. We ran 100 instances simultaneously.

Two types of harvesting were undertaken using WGET:

- Daily, with WGET configured to harvest a maximum of 50 links down in relation to the start-URL. It was eventually configured to wait two seconds between each inquiry to a given site in order to counteract possible technical problems arising from an over-aggressive harvest.
- Every hour, when only a single page was harvested. In the case of hourly harvesting it was necessary to check whether the page was a frame set, in which case one additional level would be harvested.

The URL-objects harvested were saved in a file system structured in the same way as on the server. WGET can change the linking within the URL objects acquired in such a way that their internal linking works. Links that point outside the URL objects harvested point at the original documents. For security reasons we chose also to preserve the original URL objects. This afforded an opportunity to clean up the WGET data and check how many files were full up. We will return to this point in our summary of storage needs.

WGET handles both static and dynamic URLs – apparently with the same harvesting speed.

4.1.5 Archiving with RoboSuite version 3.2.1.5

Robosuite version 3.2.1.5 was included in the test. We did not succeed in getting this version of RoboSuite to handle our needs, e.g.to parse relative links in web pages. Thus we did not succeed in acquiring data concerning the local elections from www.politiken.dk, which we had counted on being able to do.

Subsequently we had the good luck to be able to archive materials harvested from www.jp.dk using this version of Robosuite.The only fundamental differences between the two sites is apparently that www.politiken.dk makes more extensive use of relative links on their site than www.jp.dk, but we cannot exclude the possibility that the failure to archive in the former case was caused by faulty installation of the programme.

4.1.6 Donation

The project concluded an agreement with TV2 Bornholm for delivery of all their material. It turned out that TV2 Bornholm used a Microsoft Server, and all their web pages appeared as ASP. These were installed without difficulty and run on the State and University Library's web server. As part of the installation the URLs were adjusted so that they fit the new server environment. Technicians however warned that in order to ensure correct running it would normally be necessary to have exactly the same setup as existed on the delivery side (including Patch level).

The experiment was a success in so far as it was possible to run the pages, including the debates. However, if one looks at the project with long-term storage in mind, this method is not without its problems. If the system is to be run in the future, it would require that one was able to set up an identical copy of the running environment –e.g. to simulate the control system, applications and relevant help programmes – all in the correct version.

4.2 Technical and legal problems identified

In the course of running the various programmes, and in the light of the tests carried out by the researchers, a number of problems were identified. From a technical point of view, many of the problems that the researchers reported (see table in section 5.4.1), can be reduced to a series of fundamental problems, some of which can, and others of which cannot be immediately solved.

There are three main categories of errors:

- Errors in connection with harvesting – parts of pages or entire pages were not harvested.
  There were various reasons for this, such as redirects (the harvester was configured to remain within one domain); URLs were embedded, for example in flash animations; or the harvester was forced to stop because of time problems.
- Errors in connection with internal representation of URLs.
  In WGET the internal representation is transcribed last. If WGET is stopped too soon, the URLs will not be transcribed.
- Errors in the presentation system.

Most errors occurred because of redirects, e.g. the harvester was linked to another address, which might or might not be on the same host. Both NEDLIB and WGET were configured to remain within the same host, so a redirect to another host would mean that harvesting would

cease. An analysis of the materials has shown that around 3% were redirects, of which over 1% referred to another host. Further analysis has shown that certain URLs in the URL base refer to a page where one is immediately linked to an alternative start-URL on another host, so that the page cannot be harvested.

Such errors have different implications and were revealed in different ways when researchers attempted to go through the material.

Below, we go through the most important of these.

4.2.1 Redirects

The main reason why the harvesting was incomplete was the problem with redirects. Problems arise for example when you want to acquire an item from one URL, e.g.. www.tv2.dk. In practice a site such as www.tv2.dk is not just one host but a series of hosts, and this means that some of the links that come up on the site will in fact send users out to other hosts, either via automated mechanisms in the webserver's own setup, or by establishing new hosts in the TV2 domain which the user is then linked to. Since these hosts have not been entered on the list of URLs that should be archived, the harvesting will skip them, and they will then be found missing when the archived material is compared with what the researchers intended to harvest when they included the URL in question as one of their targets.

In the archive's system of presentation this error will emerge as:

- WGET: Problems with connection to link outside the archive, File not found.
- NEDLIB: Error 404: File not found.

Solution: This problem can arise in connection with both event-based and selective harvesting, because – in order to limit the harvesting to what is considered relevant – the harvesters are configured to remain within one host. By continuously following the harvesting manually one can catch these redirects and get the new hosts entered into the URL base, so that they too will be archived. The problem does not arise with cross-section harvesting.

4.2.2 URLs embedded in the code

There are several instances in which URLs are embedded in the code: this can happen because of Java script or arise in a flash animation. If this occurs the harvester will not know which links should be followed. The results will depend on the harvester and the type of code. Finally there are instances where embedded URLs are placed together in the browser.

In the archive's system of presentation this error appears as follows:

- WGET: The flash animation is harvested and shown correctly. If you click on a link, you will either get the message 'file not found' or problems with connecting to the link outside the archive. No explanation has been found as to why one or other of these errors occurs.
- NEDLIB: the same symptom occurs here.

Solution: Generally the problem has no simple solution. The reason it arises has to do with fundamental problems in the method of harvesting. The only way to identify an error is to interpret the page, for example by harvesting via a browser, or by following the harvesting and inserting the relevant URLs. However, there is no guarantee that the presentation system will subsequently be able to connect the two relevant files.

4.2.3 Time problems

Time problems were observed in working both with NEDLIB and with WGET. Since the project had greater freedom in relation to WGET, by starting 100 instances in parallel some of these problems were averted. .

The original plan was that NEDLIB should be used to

- Archive a snapshot of selected URLs on 20 October 2001 and again on 4 December 2001.
- Archive a snapshot of selected URLs once a week.
- Undertake the cumulative archiving of the selected URLs on a daily basis.

This strategy however proved impossible to maintain, simply because the NEDLIB robot was unable to finish its harvest before it had to begin archiving again. For that reason we decided to stop it and to start the harvesting process all over again at agreed times. This meant that in each round of archiving (snapshot) there are URLs that were not archived as intended. (See Chapter 4 for a summary of the snapshots carried out).

Subsequent analyses showed that it was particularly the archiving of the dynamic URLs that was time-consuming, and that no other country had previously carried out archiving of dynamic URLs using NEDLIB[8]

In a few cases it was necessary to stop WGET.

In the archive this error is shown as:

- WGET: Problems with connection to link outside the archive, since the URLs were not transcribed
- NEDLIB: 404 – File not found (since there are URLs that NEDLIB will never be able to get to).

Solution: The problem arose primarily because of very aggressive harvesting, recursive behaviour on the part of the page, or the size of the site/bandwidth. In a permanent configuration these parameters would be taken into account and the harvesting would be better scheduled. As far as NEDLIB is concerned the revised version of the programme should be used.

4.2.4 Robot.txt

Harvesting cannot be conducted on the basis of present legislation, which is why voluntary agreements must be negotiated. In view of the nature of the project, it was decided to test out the negotiation process only for selected sites, as described in section 4.1.2.

One way in which a given site can prevent visits is via the text file robot.txt, which makes it possible to prevent robot visits to the whole site or parts thereof. In our pilot study it was decided that WGET should respect robot.txt, whereas NEDLIB should ignore it. However, we made it possible for a given site to request that particular materials be deleted from the archive. Only in one instance did the owner of a site request us to delete material gathered by

[8] Subsequent experiments with NEDLIB have shown that NEDLIB's problem to a great extent lies in two processes that cannot be done in parallel, namely the link parsing (finding new links in documents harvested) and link filtering (sorting out those links (URLs) that should not be harvested). The latest version of NEDLIB is now equipped with an extra link filter, and this has brought about a significant improvement in performance

NEDLIB. The owner in question was a small publishing house, and no explanation was given for the request.

This meant that the conditions under which we acquired everything on a site must have been optimal, since we avoided the usual problem with pictures: Namely, that they are usually found in catalogues that are excluded by robot.txt. In our case, however, they were nevertheless included in the archiving. One possibility to be considered in the long term is the model that is used in Archive.org and Library of Congress, where robot.txt is not respected at the point of harvesting, but files that have been protected by a robot.txt request are not shown.

4.2.5 Robot denied access

Only in one instance was the NEDLIB robot denied access. Thus we did not succeed in using NEDLIB throughout the period to archive from one of the big daily newspapers. The owners were concerned that because of the increased load their webserver would cease to function. It was therefore possible only in part of the period to download and archive the desired URLs from this site.

An analysis of the material from WGET showed that it was excluded from a smaller number of sites. Random tests of the two types of harvesting showed that we were denied access to 0.4% of the requested sites in the case of hourly harvesting, and from 0.5% of requested sites in the case of daily harvesting.

In the archive's system of presentation this error was shown as:

- WGET: file not found.
- NEDLIB: Error 404: File not found.

Solution: Legislation or voluntary agreements.

4.3 Completeness of the archive harvested using NEDLIB and accessed via NWA

In the course of evaluating the usability of the archive, resarchers identified a number of errors that had led to gaps in the material. The researchers had access to the archive only through the presentation software, so when it was not possible to access a document, the problem might have lain in one of the following:

- The material had not got into the archive.
- The material was in the archive, but because of an error in the presentation software it could not be located and shown, giving the impression that it did not exist in the archive.

In order to get an idea of the extent of the problem we went systematically through two randomly selected smaller sites: a small site with static URLs and a small site with dynamic URLs, see Appendix 9.

The development of software for identifying and presenting the archived material did not come into the remit of this project, since we had counted on using for presentation the software developed in connection with the Nordunet2 project Nordisk Web Arkiv (NWA). Unfortunately this project was so greatly delayed that only a first version of the software was available to us, and this was configured to function only in relation to the NEDLIB archive.

In the case of the WGET archive we therefore had to use the interface that comes with WGET, which is far from being user friendly.

In general we know that there are pages that we did not archive because of the redirecting problem, and because the NEDLIB robot had to be interrupted before it had finished its work: e.g. certain URL objects were not archived because the robot did not manage to reach them in time. But in the case of static URLs, NEDLIB did in fact archive all material, with the exception of style sheets, wherever it had the possibility of doing so. Generally there were more archiving errors on those sites where there were parametrised URLs, perhaps because these sites tend to be more complex in general.

In the case of the material to be delivered by TV2 Bornholm, everything that we expected to be archived was indeed archived. Only the shout boxes and sound files, where the links are embedded in Java script, were not archived.

As can be seen in Appendix 9, however failure to archive was not the only problem.

The present presentation software is not able to handle URLs with more than one parameter. For example the software can handle URLs such as http://www.domæne.dk/hent.pl?id=1 but not those such as http://www.domæne.dk/hent.pl?id=1&arg=2 or http://www.domæne.dk/hent.pl?id=1&ny=1&f=2

The problem may be either that these URLs have been wrongly indexed, or that the software looks them up wrongly in the index.

4.4 Archiving and long-term storage

It is one thing to acquire material, but quite another both to store the material (that is, bytes) and to find a way to present an understandable version of it.

Various strategies have been used for storage, the most common being:

- emulation, where one draws up thorough descriptions of the conditons for running the material, and preserves these.
- Conversion, where one converts the material into a series of standard formats (possibly in a simplified form), which future users are more likely to be able to run.

Long-term storage is a broad field and is not part of the remit of this project. However, it is important to take this element into consideration in making financial estimates. Below we therefore present information that may be of help in calculating the overall costs.

4.4.1 Types of application

One characteristic of the harvesters is an involuntary selection mechanism, which has important consequences. As Interim Report 2 describes, it was possible to obtain only a limited number of the applicationtypes and forms of communication that the researchers considered relevant in connection with the local elections.

The results are presented on the next page:

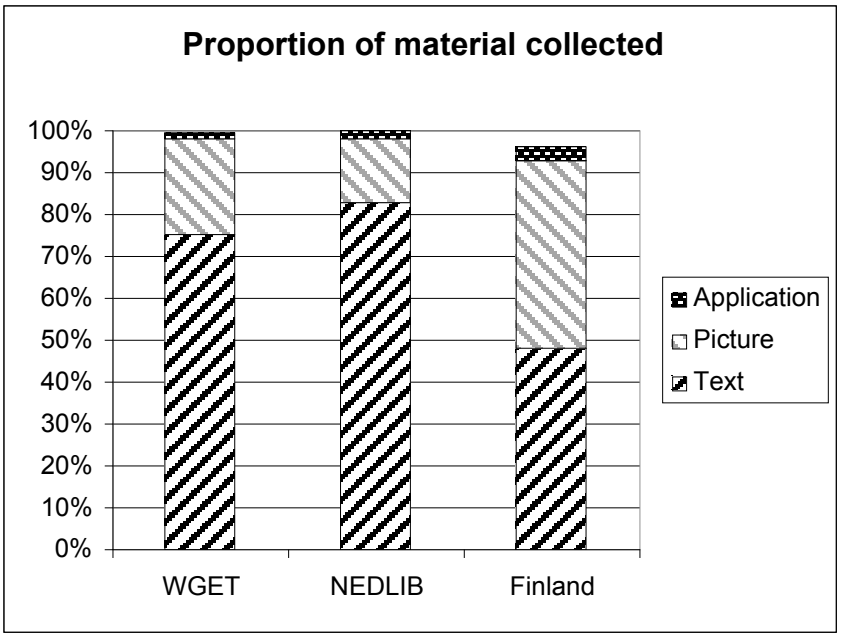
Type of material	RoboSuite	NEDLIB	WGET

Uncomplicated web pages	Uncomplicated	Uncomplicated	Uncomplicated
Slightly complicated web pages (java script)	NO, but can probably be developed	NO	No, but certain things can probably be caught via a programme
Streamed video	NO	NO	NO
Streamed audio	NO	NO	NO
Chat	NO	NO	NO
Net conferences	NO	NO	NO

Tabel 5 The different Harvesters ability to collect certan types of material

Even where it was possible to harvest material, much of it will need considerable preparatory work to make it usable in the longer term. This problem is considered below.

The overwhelming majority of the materials harvested fell into three categories: Text, pictures and applications (mainly pdf files, but also for example flash). Below, we compare the experience in Denmark with the most recent snapshot harvesting in Finland[9]. It can be seen that the Danish figures differ somewhat from the Finnish, which can be accounted for by the fact that, as a consequence of the thematic limitation we imposed, we actively chose pages whose purpose is primarily informative.



Figur 2 Distribution of the main types of material in three different harvesting experiments

The graph covers several formats for the different categories. The simple ones are, for example, pictures and sound, where the following types were harvested:

[9] Juha Hakala on the mailing list web-archive@cru.fr

Pictures	Sound
image/x-ms-bmp image/x-icon image/tiff image/png image/pjpeg image/jpeg image/gif image/bmp	audio/x-wav audio/x-realaudio audio/x-pn-realaudio audio/x-ms-wma audio/x-aiff audio/wav audio/mpeg audio/midi audio/mid

Tabel 6 List of different types of sound and picture mime-forms. Not all correspond to different formats.

If these pictures are to be shown and the sound files heard it is essential that one has the software to run them. This software is often non-standard and the source code not available. It is therefore necessary to save samples of this software together with a description of the running conditions, in order to make it possible for these files to be interpreted in the future. In the long run a better alternative for this type of file will be to convert it to a standard format for each type, and thereby reduce the number of software components required for running such files.

The ”application” group represents a much greater challenge. Let us look at what it includes:

application/zip
application/x-zip-compressed
application/x-tar
application/x-stuffit
application/x-shockwave-flash
application/x-sh
application/x-ns-proxy-autoconfig
application/x-mspublisher
application/x-msmetafile
application/x-msaccess
application/x-macbinary
application/x-javascript
application/x-java-class
application/x-java
application/x-hexedit
application/x-gzip
application/x-dvi
application/x-director
application/vnd.ms-works
application/vnd.ms-powerpoint
application/vnd.ms-excel
application/self-extracting
application/rtf
application/postscript
application/pdf
application/octet-stream
application/msword

application/mspowerpoint
application/mac-binhex40
application/java-vm
application/doc

As can be seen, this category is extremely broad. In certain cases it will be possible to convert the material into a standard format, but this is by no means true in all cases. It is known, for example, that flash presents problems. There are already problems with running flash-animations that are 3 years old, and if one has access only to the finished flash animation, there is no possibility of converting it.

4.4.2 Storage requirements

A summary of the storage requirements for the material harvested is given below, showing the total storage requirement and the average derived from it. The latter can be used to extrapolate the requirement for other events.

The storage requirement is also connected in various ways with the presentation method. The index generated by the NWA software in connection with the NEDLIB harvester increased the total storage requirement from 143 GB to 240 GB; e.g. the index represented 40% of the material archived.

In the case of daily harvests only the new pages were kept. NEDLIB itself checked in the course of harvesting whether the page had been changed, whereas WGET checked the pages subsequently to see if there were any duplicates, in which case the duplicate was thrown away and any hardlink was rewritten. This analysis showed that approximately 50% of the material harvested was redundant. However, deleting already-archived material has consequences as far as presentation is concerned. At the State and University Library we attempted to save the material on tape. One result of discarding several pages was that it made it difficult to call up a coherent unit – there may be links pointing to other material that subsequently needs to be called up as well. If all the material is stored, for example on a hard disk, this is not a problem, but if the response time is 2 minutes the user interface will be very sluggish and inconvenient. In so far as the material is stored on tape, one should avoid this kind of retrospective treatment, even if that means doubling the storage space.

	NEDLIB (compressed) for the period 4/11-23/11		**WGET** (uncompressed)	
	Total	Average[10]	Total	Average
Monthly harvesting	52.08 GB	52.08 GB	----	----
Weekly harvesting	48.11 GB	24.06 GB	-----	
Daily harvesting: Cumulative	40.24 GB	1.9 GB	Approx. 172 GB	5.9 GB

10 The average refers to the average per harvest. Where a cumulative method is used, the average is increased

Hourly harvesting	-----	----	Approx. 108 GB	0.4 GB
Total:	140.43 GB		Approx. 280 GB	

Tabel 7 Storage requirement for the harvested material

The figures for WGET are approximate in that the total strorage space is reduced, depending on the number of files that have been rewritten and where the original was kept. In an up-and-running system these would be removed. The State and University Library has a total of 437,154,430,466 bytes divided into 42,598,135 files. Of these, 13,004,155 were rewritten where the original had been saved, e.g. counted twice, which amounts to a reduction of 31%. If the same reduction is made in the storage space it means that 29,593,980 files can be expected to fill 303,702,016 bytes, which corresponds to 290 GB. Finally the number is reduced by 10 GB to 280 GB, which is the size of the log files.

4.4.3 Storage methods for preservation of material

As the previous section illustrates, the amount of data in question here is extremely large. There are two schools of thinking today concerning storage in relation to web archiving: one favouring online storage and the other storage on tape (possibly by robot). Archive.org at present keeps all its material immediately available on hard disk. The Kulturarw3 project, by contrast, has used a tape-based robot system for archiving. The consequences of using a robot system were investigated in relation to the WGET material.

It became evident that the usual methods of storage, in which information is generated on the placing of every single file, do not work, since the individual items are too small.

For example it took 4 days and nights to restore the WGET material because of the very large number of files involved.

The immediate alternative – to fetch the files if and when they are needed - is also not viable, since it takes up to 2 minutes to find and fetch one file.

Because of these problems it is considered advantageous to save all files instead of ”tidying up” as you go and saving only those files that have been altered. If one keeps complete sets of all the pages, it is always possible to establish sensible ways of assembling and packing them.

Ideally the system of packing should take future users into account. It is probable that users in the future will be interested both in getting a time-snapshot and in being able to compare a given set of sites over a period of time. One possibility therefore would be to use a matrix method, where the unit in one dimension is time (for example months) and in the other a natural group of sites (for example newspapers, districts and so on).

5 The research angle

5.1 Synopsis of research interests

This synopsis was made with a view to help to identify materials, to establish a monitoring strategy and to develop research tests. The purpose here was not to give an exhaustive description of the possible research value of the material, but to give a specific and meaningful introduction to the monitoring and tests. Two general questions were selected, which were then split up into a series of sub-questions.

The two general questions were:

1) Did the Internet have any impact on the local elections?
2) Did the local elections have any impact on the Internet?

The first of these questions was then broken up into a series of sub-questions, among them:

- What was the extent of use (this is not part of the remit of the present pilot project).
- What was the impact of strategic initiatives on the part of the authorities, parties, media in terms of increased turnout (e.g. the turnout of young people in Nordjylland County), influence on election results, increased participation in debates, and so on?
- Was material from the Internet picked up by other media?
- Can cross-media relations be established – partly linking the same actors' activities, partly in relation to the public?
- Is the Internet as a new medium of particular relevance to local communities?
- What is the relevance of Internet material in relation to current Internet/democracy theories?
- Does the Internet open new channels of communication between citizen and citizen, citizen and politician, politician and politician?
- What is the relevance of the Internet in relation to particular fields, e.g areas that typically fall under the responsibility of local authorities: institutions (for young people, pensioners), health, schools, traffic, work contracts; and demographic issues: special regional features, sex, age, class, culture, language, ethnicity, etc.

The second question was similarly broken down into a series of sub-questions, including:

- Were particular techniques (e.g. audio-videostreaming, wap, pda) implemented for the first time or used much more extensively than hitherto?
- Were new forms of interactivity and hypertextuality (chat forums, discussion forums, link patterns etc.) implemented for the first time or used much more extensively than hitherto?
- Were there notable innovations in terms of genre, design and content (interactive genres, new forms of group-targeted communication, fan- and hate pages)?

5.2 Location of material

As described in Interim Report 2, the material we wanted to archive was identified and categorized on the basis of four overall criteria:

1. Types of actor involved.
2. Demographic and other related factors.
3. Types of communication.
4. File types.

Each of these four main criteria was subsequently broken down into a series of sub-categories, on the basis of which we built up a URL base that was used as the foundation for harvesting the material for the archive. A full description of the content of the URL base is given in Interim Report 2.

The harvesting of URLs was carried out with the help of index pages and search machines and by unravelling link chains. No attempt was made systematically to acquire material via other media.

One important conclusion we drew was that certain actors (e.g. TV2) use a varying number of servers (IP- and web adresses), and that it is therefore problematic in such cases technically to identify their web sites. Since the web sites must be regarded as the foundation for archiving (see Appendix 1), a particular effort must be made to address this problem. Newly-formed URL addresses also pose a particular problem, because in this connection the available search machines work with considerable delay. This question is addressed below in our summary of the particular problems that arise in the archiving of net events (Appendix 2).

5.3 Monitoring strategy

The purpose of monitoring net activities concurrently with the archiving process was to obtain materials that would make it possible to evaluate:

- Whether the archive material acquired was sufficiently comprehensive in relation to the actual amount of net activity in progress.
- Whether the techniques of harvesting yield the desired material, and what technology if any is missing for acquiring and archiving material.

The monitoring, which was carried out in the period from 22 October to 3 December 2001, also served to provide input on how the harvesting frequency should be adjusted, and to add in subsequently discovered URLs (including newly-established pages)

5.4 Archive testing

One purpose of the project was to test the quality of archived materials and the accessibility/user-friendliness of the archive. The main emphasis was on testing the quality of the material, while the testing of accessibility and user-friendliness was conducted simultaneously (for an overview of the material that was presumed to be in the archive prior to the test, see Appendix 4. The appendix also shows what was actually tested. For a detailed presentation of the arrangement of the test, see Appendix 3).

5.4.1 Testing of the material harvested

Purpose and arrangement of the test

From the research point of view the overall purpose of the test was to discover to what extent the material archived was usable for research purposes.

The plan was for the quality of the archive material to be tested in four respects:

1. Internally: The archive material would be tested 'in itself', as it appears in the archive against the background of the various types of archiving software and harvesting frequencies.
2. Externally: The archived material would be tested out in relation to the initial collection, e.g. the types of material that were selected in advance for archiving.
3. Externally: The archived material would be tested out in relation to the monitoring process
4. Externally: The archived material would be tested in relation to other existing archives, primarily archive.org (http://www.archive.org/) and the Danish legal deposit system (http://www.pligtaflevering.dk/).

In all four instances a number of questions were posed with regard to the quality of the material (the concrete questions relating to all the tests are presented in Appendix 5). For the internal test (of the archived material 'in itself') the questions fell into the following main categories: The presence of the material in the archive; frequency/content; the quality of the material. In relation to the second dimension, where the archived material was tested in relation to the initial collection, questions as to both the presence and the quality of the material were posed. Where the material was tested in relation to the monitoring, the additional question was posed as to whether the harvesting of the material correponded time-wise with the monitoring process. Finally, the testing of the material in relation to other existing archives was carried out in two stages: First, the other archive in question was subjected to the same internal test as the material archived for this project; second, the two archives were compared on the basis of the same three main categories.

Four cases were chosen for testing:

- Parties (defined as one type of actor); here, the Social Democratic party and the Venstre party were selected, both on the national and district/country level.
- Media (defined as one type of actor); here, the newspapers JyllandsPosten and Politiken, and the broadcasters TV2 Fyn, TV2 Bornholm and DR were chosen.
- Districts and counties (defined as geophysical entities); here, Århus County, Odder District, Nordjylland County, Hals District, Copenhagen County, Copenhagen District, and Gentofte District were selected
- A 'remainder' category, covering all the interesting items that were not documented in the three other cases, and applying both to types of actor and to forms of communication.

Since each case was to be followed over a period of time, a series of test dates were selected.

It was mainly the two political parties and the five media institutions that were to be studied 'in depth': E.g. they would be tested daily, using all the archiving software, in relation to all the main questions listed above; while the county/district-related sites were to be tested more

'in breadth', e.g. not necessarily daily and with all the software, but with reference to all the main questions.

In addition, examples of inconvenient forms of arrangement and functioning in the archive, and good ideas for improving them, would be noted on an ongoing basis.

The testing procedure

The testing was carried out with the assistance of two students. As mentioned in Chapter 4, different types of archiving software were used. Since NEDLIB and WGET were the principal forms, they were the main focus of testing with regard to the 'archive in itself'. Because of the problems in presenting the NEDLIB material, much of the focus during the test period was on the material gathered via WGET. Moreover, the fact that there were greater difficulties in navigating with NEDLIB than with WGET meant that the NEDLIB material was not tested to the same extent as that harvested with WGET. Random tests should however be sufficient to show the strengths and weaknesses of the NEDLIB material. Thus in the testing of both the NEDLIB and the WGET material several different types of site of different sizes were tested, and a certain level redundancy was reached in investigating the strengths and shortcomings of the different archives (the testing of NEDLIB was carried out in relation to the home pages of the Social Democratic Party and the Danish liberal party, Venstre, all the media sites, and Århus County in the category 'geographical areas'). Finally, the material archived with the help of Linux platform and Robosuite — together with the material donated — could not be systematically tested (material donated by TV2 Bornholm underwent a subsequent test, see below).

The external testing of the archived material in relation to the Danish legal deposit system was not carried out in the end, since it was estimated that there would only be a very limited overlap.

Results of the test

The archived material 'in itself' (and in relation to the initial collection):
Remarks in relation to the presence, frequency/content and quality of the material harvested are summarised under the following main points (see Appendix 6 for further points, details and examples):

NEDLIB:

- A large number of sub-pages turn up in the index when one wants to view a site (e.g.: Jyllands Posten (daily) approx. 6,000, Berlingske Tidende (daily) approx. 14,000), which could constitute a problem, since it would mean having to try to navigate through numerous random pages.
- The number of levels that can be shown varies considerably (this also applies within one individual site); on larger sites this probably means, for example, that one cannot read the articles in a newspaper unless one is lucky enough to find the article in question among the thousands of individual URLs that are in the archive's index for the site.
- Different forms of graphics are by no means always shown (besides the announcements and banner advertisements that are fetched from another server, this applies most often to pictures), and some instances can be found in which, although the text is shown, it is not shown in the original form or colour.
- There appears to be problems in relation to sites that use frames.
- Sometimes an extra, virtually blank page is incorporated in addition to the actual site.

- One cannot always rely on the time line's indication of how many times a page has been harvested.

WGET:

- Although the harvesting was often carried out continuously from the start, there are many instances of "groundless" interruptions in the harvesting.
- Differences in depth: There are great discrepancies in the number of levels that can be shown.
- Graphic differences: There are variations in the graphic elements that are shown.
- Differences in feedback when activating features: There are discrepancies in whether the features function, whether there are problems with connection to links outside the archive, or whether the message 'file not found' is shown.
- There are cases in which links on the same site that were harvested on different dates react differently on being activated that is connecting to links outside the archive or not.
- Problems may arise when sites use frames: A link outside the archive may be activated without the browser showing it, or the message "file not found" may appear.
- The fact that a site name is included in the index is no guarantee that it is actually present in the archive.
- It can happen that individual URLs are not coupled to the home page, even though they are in the archive, which in concrete terms means that you cannot use the menu.

As mentioned above, a number of special (and in certain cases new) functions were used in connection with the local elections. The table below shows to what extent these appeared in the archive (Y = yes, N = no). Appendix 6 gives a detailed account of the observations made, where problems arising in relation to presentation or configuration are identified.

Function	NEDLIB	WGET
Streaming of sound, moving pictures	N ("document not found")	N connection to link outside the archive or "file not found on the server")
Sound/moving pictures which start up automatically as you enter the site	Y N (flash is missing on www.ungtvalg.dk)	Y
Material that can be downloaded	N ("document not found") Y (e.g. on www.louisegade.dk)	N (but flash animation works on Jyllands Posten) Y (e.g. on www.socialdemokratiet.dk and www.louisegade.dk)
Discussion forums	N (but Y on the local election page on DR's site, where the debate is conducted under the different individual regions) N (Politiken) N (but Y in so far as one can find debate contributions that	N (because of connection to link outside the archive on DR's site) Y (Jylland Posten) N (Politiken) Y (TV2/Bornholm: via index) Y (Social Democratic party)

Function	NEDLIB	WGET
	are included as individual URLs) N (Socialdemokratiet)	
Chat	N (”document not found”)	N (because of connection to link outside the archive)
SMS[11]	The pages from which these are ordered are in certain cases present	The pages from which these are ordered are in certain cases present
Games	N	N connection to link outside the archive)
Opinion polls/profile tests	N	N
Quickpolls	N (e.g. http://www.venstrenet.dk) N (e.g. do not appear on www.ungtvalg.dk)	N e.g. http://www.venstrenet.dk Y e.g. http://www.ungtvalg.dk)
Robot answer (Rosa)	N connection to link outside the archive when you press enter and shows ”document not found” if you click the mouse) *(exclusively on A’s site)*	N connection to link outside the archive) *(exclusively on A’s site)*
E-trade pages	N (because of connection to link outside the archive)	N (because of connection to link outside the archive)
Calculation of mandate test	Can’t be tested, since Nicolai Wammen’s page (which is the only place where this function is found) is harvested only on 2 levels	Can’t be tested, since Nicolai Wammen’s page (which is the only place where this function is found) is set for hourly harvesting, so only one level is taken
Shoutboxes	N (in the frame where the box should be, ”page could not be found” appears (probably due to connection to link outside the archive)	N (shoutboxes saved, but in a wrong and not very legible layout; one can’t shout in the archive)
Newsletters[12]	• ordered newsletters are present in full quality in	• ordered newsletters are present in full quality in

[11] In connection with SMSs the only test made was whether the page from which SMS messages can be sent or ordered was in the archive. The SMS messages themselves (meaning the messages that can be ordered) were not archived, but since — as was the case with the local elections — they are often part of a news circular (along the same lines as a newsletter), it will in future be important to archive them (together with those web sites they belong to, again as in the case of newsletters; see below).

[12] As part of the project we subscribed to a number of newsletters and so on. In this connection we tested out whether the newsletter we had subscribed to, and the page from which they could be ordered, were present in the archive. It should be noted here that the newsletters ordered and received were not connected in the archive with the web siteweb sites they belonged to. Moreover some providers always place their newsletters on their site (as they are sent out); here too we tested out whether these were archived

Function	NEDLIB	WGET
	the special newsletter archive • The page from which they are ordered is present • a very few sites contain the content of the newsletters themselves	the special newsletter archive • The page from which they are ordered is present • a very few sites contain the content of the newsletters themselves
Password protected areas	On Jyllands Posten: ”Document not found” (probably because that level was not harvested)	On Jyllands Posten: password requested (no connection to link outside the archive)
Java scripts	N (e.g. www.stoplufthavnen.dk) N (dr.dk/kommunalvalg: ”document not found”) N (TV2/Bornholm: ”Document not found) N (Rosa on Social Democratic Party) *(The systematic search for this was dropped, since it can be complicated to find out if something is in Java script)*	N (e.g. www.stoplufthavnen.dk) N (dr.dk/kommunalvalg: ”file not found on server”) N (TV2/Bornholm: ”The browser cannot find the page”) N (Rosa on Social Democratic Party) *(The systematic search for this was dropped, since it can be complicated to find out if something is in Java script)*

Tabel 8 List of different function types occurring in the two archives

In most cases the quality of the material is the same on NEDLIB and WGET. However, WGET appears to be better at harvesting debate forums and quick polls.

The archived material in relation to the monitoring
Among the new materials that were observed and reported during the monitoring period, a fairly large amount was not archived. WGET archived least (see random tests in Appendix 6).

The monitoring reports do not contain many comments relating to specific points of time, but it was nevertheless confirmed that there was a correspondence timewise between the monitoring reports and the archive (except of course where the reports mention new sites, since these are only included in the harvesting at a later point).

As far as the quality of the archived material is concerned, the monitoring reports did not establish anything ’new’ in relation to what emerged from the testing of the ’archive in itself’. Thus the problems that emerged there in relation to certain functions, certain graphic elements and navigation within the archived material also apply here.

The comments arising from ’archive observation’ were made in relation to WGET, since it was this material that the test in the first phases was concerned with (inclusion of the NEDLIB material would probably not afford offer? any new information, but simply confirm the remarks made in connection with the ’archive in itself’.)

The archived material in relation to other existing archives (archive.org)
The harvests carried out by archive.org take place months apart and at very irregular intervals, which make direct comparison very difficult, and means that archive.org is not very usable in a case like the local elections.

Often however the quality is better on archive.org than on WGET and NEDLIB. This is evident in the fact that in archive.org's versions (a) there are more complete graphics (e.g. on jp.dk), (b) more levels are harvested (e.g. from juelsminde.venstrenet.dk). However, we cannot draw any firm conclusions from this comparison, as the quality of NEDLIB/WGET in some instances is better than that of archive.org (see for example kommunalvalg.tv on WGET).

Accessibility/User friendliness

In Appendix 8, the "Note with comments and desiderata concerning accessibility/searchability" sets out a series of remarks and desiderata with regard to the accessibility/user friendliness of the archive in the longer term. The remarks are based on observation of the way that NEDLIB/NWA worked during the test period.

The following points were noted in relation to NEDLIB:

- The logic in the presentation of different URLs was not transparent.
- The only way to get back to the index was via the 'back' function, which is not very practical when one may have moved around different sub-pages on different dates.
- Under the different categories in the index, the links to 'First/ Previous/ Next/ Last' page are inconveniently placed at the top of the page: It is almost always at the bottom of the page that one needs them (by the same token, these navigation possibilities are placed at the bottom of the index under the individual sites).
- The time line at the top is sometimes difficult to use in navigating. You often have to be precise to the last millimetre in placing 'the brown frame' over 'a black stroke' in order to fetch a page.
- When you use the function 'Go to page –', it jumps back to 'whole archive' instead of remaining in, e.g. the 'daily' section.

The following points were noted in relation to WGET:

- It is impossible to navigate via the index.
- Connection to a link outside the archive often occurs without one immediately being aware of it (see e.g. valg.bornholm).
- It would be more user-friendly if one could avoid going through the 'index' in order to get access to the archived material (as is the case with archive.org).
- A search function belonging to the programme itself would also make navigaton easier (the search function we used was in Windows).
- The variable speed in the archive's listing of web sites sometimes slowed down our work and would also be a nuisance to potential users coming from outside.

Finally, the following points were noted in relation to archive.org:

- Navigation: in terms of user-friendliness, archive.org has the advantage that — unlike with NEDLIB/WGET — you can click onto a page without having to go through the index.

- Speed: NEDLIB and WGET are much faster than archive.org in reacting to a click on a link, which is an obvious advantage if one is making frequent use of the archive.

Conclusion

As discussed above, the overall purpose of the test from the research point of view was to find out to what extent the material archived was usable for research purposes. On the basis of the test we can conclude that frequently it will not be possible to use the material as a reliable research object. The reasons for this are primarily the following:

- In too many instances the archived material does not appear in the form of a web site, but rather as a 'collection' of individual URLs (e.g. in a form quite different from that found on the active net); this presents a problem to the user, if the links/navigation between the various parts of the site do not work, and if the start/home page can only be found with difficulty: See Appendix 1 on web sites as basic elements in an archive (it should be noted in parenthesis that it would moreover be an advantage if newsletters and so on that were ordered and received over the net were archived in conjunction with the web sites they belong to).
- It is a problem that the material archived differs considerably from web sites on the active net, which reduces their usability and reliability; in this regard it is especially problematic that the material is not harvested continuously and in full depth, and that certain navigation menus and graphics are missing. Finally, the complete or partial absence of important forms of expression and functions (streamed material, discussion forums, chat pages, games, opinion polls, quick polls, robot answers, and password protected areas) considerably reduces the usability of the archived material.
- A serious problem in relation to the reliability of the archive is that if it is not 'closed' in relation to the active net, it often results in a connection to a link outside the archive.. This also affects its usability, since it demands too much attention from the user continuously to have to keep an eye on whether a given document is valid (and in some cases it is impossible to decide this for certain).
- Finally, the highly complex nature of the archive and its very variable quality creates a more general problem, in so far as uncertainty over the status of the material *in itself* constitutes a limitation; for at the 'entrance' to the archive material there is nothing that indicates to the user how close or otherwise the archived material is to the original, which produces a general uncertainty: Can I be sure the whole web site is here? What is missing and why? Am I in the archive or on the net, or both? And so on. Thus the archive generates its own kind of insecurity because of the absence of clear markers and because of its fluctuating quality.

5.4.2 Testing of donated material (TV2 Bornholm)

As mentioned under point 5.1, the material that was donated to the archive could not be tested systematically. It was therefore decided to choose one donated web site for testing, namely the material from TV2 Bornholm.

Subsequently we decided to extend this test, partly in order to try out once more the archiving tools we had already tested (NEDLIB and WGET), and partly to try out some as yet untested archiving software.

The purpose and arrangement of the test

The overall aim was the same as with the previous test: To consider the usefulness of the material obtained for research purposes, but as noted above this time the testing was carried out with three types of material:

- Material donated to the archive, which after treatment again was placed on the net (in a closed area).
- This donated, but now familiar and stable material, harvested with already tested archiving software (NEDLIB and WGET).
- The donated, known and stable material harvested with other types of archiving software (Adobe Acrobat for PC and Mac, HT-track, Internet Explorer for Mac).

In this instance we tested only (1) whether the desired material was present in the archive (2) the quality of the archive in itself, rather than posing the other questions mentioned in section 5.1.

Implementation

In this case the testing was also carried out with the assistance of two students.

Results of the test

All the tools tested, except Adobe Acrobat for Mac, were judged to be useful archiving tools (see Appendix 6 for details). Adobe Acrobat for PC and Internet Explorer for Mac were of limited use, however, since they archive only two levels down. WGET, NEDLIB and HT-track can generally archive the whole site (3-4 levels down). None of these, however, is capable of archiving streaming. A comparison of the three latter tools shows that HT-track has advantages over the other two, for in contrast to the material harvested through WGET and NEDLIB, that harvested through HT-track has none of the 'holes' in terms of level, speed and graphics that the other two suffer from. The difference between the three is not very marked, but is sufficiently great for HT-track to be regarded as preferable. In terms of content, the main advantage of the donated material in relation to that harvested through WGET, NEDLIB and especially HT-track thus consists in the fact that it includes streamed material.

Accessibility/User friendliness

User friendliness was generally high. WGET's sluggishness on the first occasion one uses the archived version is thus the only 'serious' problem. Moreover, in all the tested versions the usual problems with connecting to a link outside the archive apply, thus reducing their user-friendliness.

Conclusion

On the basis of this test we may conclude that harvesting on a smaller scale and under 'controlled' and stable conditions tends to mean that the material archived is more usable for research purposes. Nevertheless, streamed material – a significant form of web activity - still presents problems with connecting to links outside the archive which reduces the usability of the material.

5.4.3 Time required for acquiring URLs as part of identifying an event

In the pilot study carried out in 2001-2002 the participating research group from the Centre for Internet Research conducted the identification and harvesting of URLs. The researchers involved formulated the principles on which this procedure was carried out, while the practical work of building up a URL base that could provide the foundation for harvesting was undertaken by students and staff from the participating libraries, who also took care of the technical work of setting up the data base.

The task proved more time-consuming than anticipated, but a significant part of the time involved was spent on finding solutions to a series of 'first time' problems such as finding search machines capable of providing usable indexing and technical problems with reporting tool kit.

The time devoted to this part of the project by the two participating researchers amounted to approximately 3 weeks' work (75 hours), while the students devoted a total of 120 hours to the work. The librarians' work in building up a database for the specific event should be added to this. Once the necessary methods and routines are developed, a task of this kind (excluding the work of librarians) could probably be carried out within roughly 100-120 working hours, of which the hired assistants would contribute to the majority (at a rough estimate 75%), while the remainder would demand event-related expertise.

If one regards this pilot project as typical (on the one hand it was predictable a long time in advance, and therefore not too unwieldy; on the other it was wide-ranging and from district to district the content varied considerably, so that it was in this sense relatively difficult to form an overview of the event as a whole), the preparatory work of acquiring URLs relating to an event of this kind would thus require approximately 120 hours of work.

In subsequent calculations it was assumed that the work of researchers/librarians in relation to such an event represented 0.1 year of work.

5.5 Evaluation of the number of sites in connection with selective harvesting

Below we give a rough estimate as to how many sites should be included in a national strategy with regard to selective harvesting. Three principles on which to select web sites for regular harvesting are proposed:

a) The first and most important criterion is that the web sites in question should be ones that are significantly updated or renewed between quarterly cross-section harvests, and that they do not practise cumulative updating.

The two other criteria should be considered side by side:

b) One criterion suggested is that web sites of significance to the Danish nation should be chosen.

c) The other is that a representative selection of typical/characteristic web pages should be made, since such a sample would include both widelyused forms of activity (e.g. portals of the net doctor type), local media users (city portals etc.) and examples of experimental and/or unique net activities (such as for example, Chili.net, torshammer, open debate environments and net communities).

d) We have not had access to traffic statistics of a sufficiently detailed nature to support the selection procedure, and in any case such data in certain cases could only be used as a supporting criterion. It should be noted, in general, that traffic figures are not appropriate as a definitive selection criterion, since the largest numbers of visits are most often obtained by navigation pages (search machines, portals) rather than by content pages, and do not therefore reflect any value in terms of content or usefulness. On the other hand, traffic statistics would be of considerable research interest.

Criterion (a) would get rid of a number of public institutions and larger Danish enterprises that seldom if ever update their web sites, and/or that carry out cumulative updating..

In addition, a great many of the more dynamic, interactive parts of the net, including sites devoted to culture and the arts, which are not included in the rough estimate below, would be missing.

If equal weight is given to the two other criteria, then the proposed quota of 75-80 web pages could be divided roughly as follows:

5.5.1 Web page that function as media for the national public

The following types of sites in particular are included here:

Government forums, particularly the web pages of Parliament and the ministeries:

It is considered that the Danish Parliament's web page should be included on principle, while in the case of the ministeries one could consider leaving certain ones out – on the assumption that the content of these web sites is most probably also available in other media.

By the same token the Danish web pages relating to the EU and other international organisatons to which Denmark belongs would also be left out.

NUMBER FROM THIS CATEGORY: 1.

Political web pages (parties, movements, interest groups):

We calculate that the web pages of at least the officially recognised political parties should be included.

The web sites of other small parties, movements and interest groups would therefore be excluded, as would those representing the main actors in Danish culture.

NUMBER FROM THIS CATEGORY: 10-12.

Media (both the printed and electronic media on the net and the pure net media.).

Here we suggest:

The net editions of the national print newspapers
Berlingske, Politiken, Jyllands Posten, BT, Ekstrabladet, Børsen, Information, Kristeligt Dagblad, MetroXpress, Erhvervsbladet, Søndagsavisen, Weekendavisen.

The largest regional newspapers:
Århus Stiftstidende. Jyske Vestkysten, Fyns Stiftstidende, Nordjyske Sitftstidende.

TOTAL: 14-16.

The net publications of the national, Danish language TV- and radio stations:
DR, TV2, TV3, DK4, TV Danmark 1 and 2, CNN-Danmark and DR-radio + new channel.

TOTAL: 9.

Pure net media:
Infopaq, euroinvestor, rb-børsen (http://www.rb-borsen.dk/)[13], jubii, altinget, worldradar, horisontnet (http://horisontnet.dk)[14]

Here, the web pages of a large number of specialised or small national and local news media, including net TV and radio, are excluded.

TOTAL: 7.

TOTAL FROM THE NATIONAL DANISH MEDIA: 41-45.

5.5.2 Representative and characteristic web sites

Web sites that would ensure a representative sample of typical, widely practised forms of activity and examples of experimental and/or unique web activities:

Here, a provisional list is offered, and it is suggested that this list should be continuously revised in relation to the development of activities on the net:

- torshammer – a one-man net newspaper, particularly serving Danes living abroad (unique in this medium).
  NUMBER: 1.
- chili.net, nationaldebat.dk and a couple of other community sites
  NUMBER: 4.

13 RB Børsen say of themselves that they are Denmark's leading business, economic and financial news bureau, broadcasting domestic and international business news from 7.00 to 18.00 hours every day the stock exchange is open, together with market reports, key figures, estimates, newspaper excerpts, National Bank information, events on the business and financial calendar and the calendars of Parliament and the EU, and so on.
RB-Børsen's news is distributed through its own Internet system but also via DanskeBank TeleService, Moneyline/Telerate, Bloomberg, SIX Trader, WinTrade (Nordnet), The Online Trader, Ecovision and Ritzaus Bureau.

14 Horisont.net state that they: deliver subject-related and professional knowledge within clearly defined fields. There is free and easy access to daily news, an article archive, a debate forum, expert panel, supplier and product guide, new literature and research, a job data base, press releases and a free e-newsletter.

- 3-4 city portals (to be selected after more detailed observation)
  NUMBER: 4.
- Web sites that together cover a local area (including a local newspaper that is on the net, a city portal, other local portals (belonging to associations, citizen's groups, local parties and movements, etc.).
  NUMBER 15-20.

Reprsentative and characteristic web sites: TOTAL: approx. 30-40.

5.5.3 URL base for the 2001 local elections

For comparison: The URL base for the 2001 local elections included the following media:

Printed newspapers on the net:	141
Electronic media on the net:	
radio:	66
TV:	64
Pure net media:	
Net newspapers, net radio, net TV:	36
Portals:	415
TOTAL	722

6 Definition of *danica*

The current *Announcement of Legal deposit of Published Works (Bekendtgørelse om pligtaflevering af udgivne værker)* (BEK no. 1041, dated 17/12/1997) does not contain a definition of *danica*. The main rule, however, is that works that are published here in Denmark should be regarded as *danica* and are subject to legal deposit.

The internationalization of the print and publishing industry, where it frequently happens that works are produced in one country with a view to publication in another, and the publication of databases that can be physically located on servers abroad, has meant that §1, item 4 of the Announcement has sought to limit the legal deposit requirement in relation to:

1. Publications produced abroad with a view to dissemination in Denmark.
2. Publications produced in Denmark with a view to dissemination abroad.

Publications produced abroad with a view to dissemination in Denmark

If publications are produced abroad especially for publication in Denmark they are subject to legal deposit. It can be difficult to determine whether this is the case. The criterion given in the Announcement is primarily that of language. For example, works in Danish or translated into Danish, or works including Danish speech or texts, are mentioned as examples of *danica* and as thus being subject to delivery.

Publications produced in Denmark with a view to dissemination abroad

Works that are produced in Denmark with a view to publication for an international public are not subject to legal deposit unless the content of the work is connected with Denmark. The content is considered to be 'connected to Denmark' if:

- The originator of the work is Danish,
- The work relates to Danish conditions
- The work is presented in Danish or
- The performers involved are Danish.

If one of these conditions is fulfilled, the work is treated as a *danicum* and is subject to legal deposit.

6.1.1 The administration of *danica* today

The following categories of printed material are today kept in physical form at the Royal Library in Copenhagen as the property of the national library:

1. Works published in Denmark (acquired by legal deposit).
2. Works by or about Danes that are printed abroad (purchased).
3. Works about Denmark that are printed abroad (purchased).

The State and University Library is Århus has the equivalent responsibility for acquiring, via legal deposit, audiovisual works issued in Denmark on CD and video/DVD, while foreign *danica* in the audiovisual field are sought out and purchased. In addition, the State and University Library acquires Danish radio and TV broadcasts for the State Media Collection

through voluntary agreements with stations, since the law on legal deposit does not cover radio and TV. Where it is not possible to conclude voluntary agreements – e.g. if the station is anchored abroad, like TV3 – the harvesting is carried out by actively ”downloading the broadcast from the air”. The copyright law permits the State Media Archive to recieve and archive radio and TV broadcasts for its collection.

6.1.2 *Danica* in relation to the Internet

The practical administration of legal deposit and harvesting of *danica* thus embraces:

- Works that are published in Denmark.
- Works by and about Danes published abroad.
- Works about Denmark that are published abroad.

If we translate this into Internet terms, the corresponding categories would be:

1. URL objects in the Danish subdomain (= .dk).
2. URL objects created by Danes outside the Danish subdomain, that are:
    a. In Danish.
    b. In a language other than Danish.
    c. Presented by Danish authors or artists.
3. URL objects outside the Danish subdomain that relate to Denmark.

In so far as the law on legal deposit is changed, so that a legal framework is created for carrying out comprehensive archiving of Danish material on the net (which corresponds to point 1), it will be technically possible to embark on such activity immediately, in that complete lists of Danish domains could in that case be obtained via DK Hostmaster A/S, which is the administrator of DK-TLD (Danish Top Level Domain).

The general point to be made about URL objects in categories (2) and (3) is that they are located on servers outside Danish jurisdiction. The key point in copyright law, however, is that the place in which the harvesting is carried out is decisive in determining which country’s law applies. Thus, if it proves technically possible to download these URLs, then under Danish law the legal deposit institutions (the Royal and State and University Library) would be permitted to acquire them.

It is however conceivable those URL objects that will in future be protected by technical protection devices such as encryption or copy protection. The Danish legal deposit libraries (omformuleres) could not demand that such protection devices be removed.

It would therefore be sensible if, in relation to points (2) and (3), agreements were made on an international basis, so that each country would have the possibility to archive the relevant parts of their cultural heritage, in so far as this is deemed necessary.

The most straightforward way to locate the Danish material under (2a) (Danish URL objects created by Danes outside the Danish subdomain) would be to cooperate with one of the biggest search machines. The big search machines see most pages on the net and do a language analysis of the pages in advance with a view to offering services that enable the user to limit his/her search to pages in a particular language. Under such a cooperative agreement the search machine would deliver those URLs which a language analysis has identified as being in Danish, or which include linguistic variants of the word ’Denmark’. These URLs

could be sorted by a manual search (and those that were not in fact in Danish could be removed) and the collection could then be corrected (by adjusting the URL to make it the start URL for the part of the foreign web one wished to preserve, rather than preserving just the actual page).

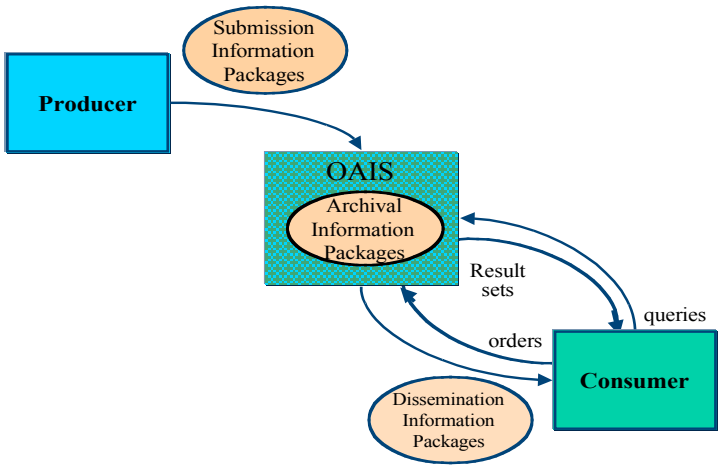
The Finnish National Library attempted to set up an agreement with the firm that administers .com addresses in Scandinavia, so that it could have all Finnish .com domains delivered. This initiative failed, since the information involved was considered too confidential. However, the Finnish National Library did succeed in coming to an agreement with a Finnish firm that has maintained a portal over Finnish sites since 1997 (Info Center Finland). As far as we are aware, there is no corresponding firm in Denmark.

With regard to (2b) (URL objects created by Danes outside the Danish subdomain, that are in a language other than Danish), a comprehensive search would be expensive and a mechanical one too arbitrary. However, since this category includes among other things many of the big Danish firms on the .com domain, we suggest that within certain limits material should be acquired within this category. With regard to (2c) (URL objects created by Danes outside the Danish subdomain, that are in languages other than Danish but involve Danish artists/authors) a search could be carried out on the basis of a list of names, possibly in cooperation with organisations of artists, writers, researchers, designers, sportsmen and so on.

7 Finance

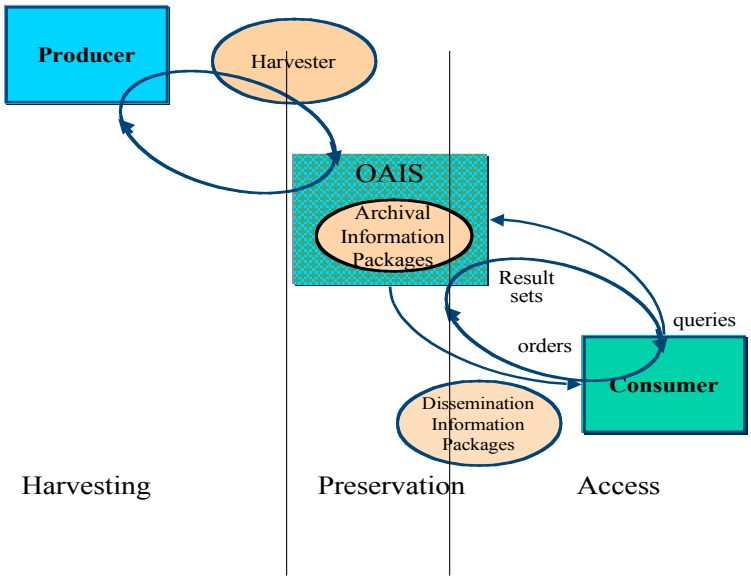
One purpose of this project was to produce an estimate of the costs involved in web archiving. As a first time in making such an estimate, it is necessary to consider the size of the storage procedure one has in mind. For this purpose it is useful to look at the activities involved in terms of an OAIS model (Open Archive Information System), which is currently winning recognition as the architectural model for long-term storage.

At the general level the model distinguishes between three functions that are sketched below:



Figur 3 OAIS overview

In the case of web archiving, the traditional delivery process from producer to archive is changed to an active gathering process, so that the figure is modified as follows:



Figur 4 Modyfied OAIS model

The following three work tasks, which typically are involved in harvesting, are implemented at the same time:

- Selection.
- Handling of rights.
- The physical process of harvesting.
- Quality control of the material: quality of copies and completeness.

The work tasks typically involved in storage are:

- Drawing up a storage strategy.
- Description with a view to retrieval.
- Description with a view to storage.
- Protection of data.
- Protection of interpretation of data.

The work tasks typically involved in access are:

- Administration of access rights.
- Development of software for retrieval.
- Development of an environment for presentation.

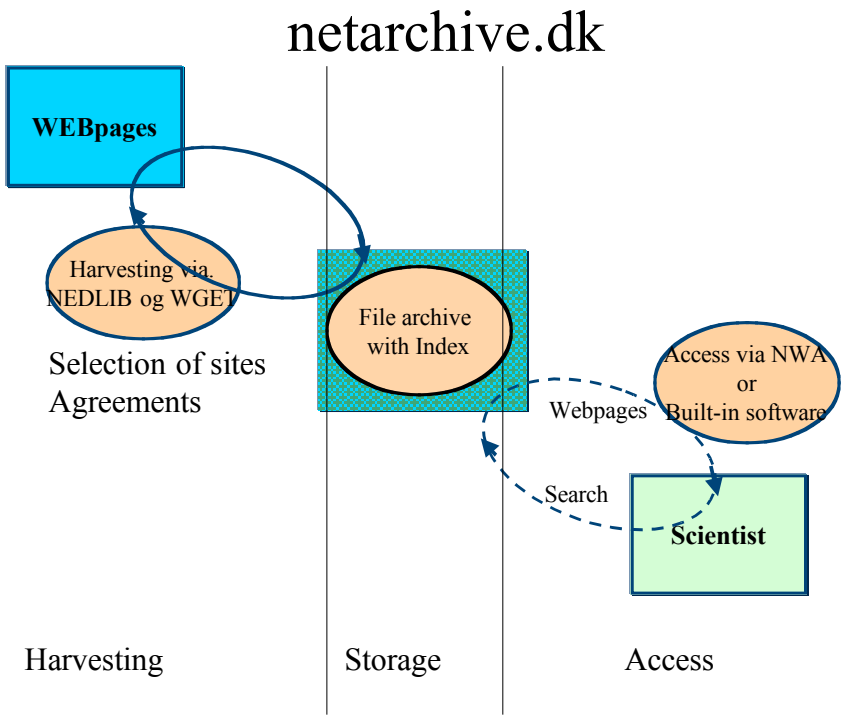
The project has focussed on harvesting and has gathered experience with regard to the following questions:

- What is required in selecting relevant sites in connection with an event
- Technical obstacles.
- Legal obstacles.
- The quality of the material acquired.

The project did not deal with the question of long-term storage. The material was saved as it was harvested, e.g. sound files were saved in a large number of formats. With regard to long-term preservation this strategy should be reviewed and related to the general strategy for the long-term preservation of digital materials. One possibility that should be investigated is to identify formats for long-term preservation and immediately convert the material to these.

The question of retrieval is only peripherally dealt with, since this is part of another project (NWA). No software development has been undertaken – NWA software was used in connection with NEDLIB, while in the case of WGET we used the software that came with the programme. Nor did we deal with the problems involved in the possible generating of metadata that would make possible the storage of relevant plugins and other features necessary for seeing the material, together with metadata that enable one efficiently to retrieve the material.

.



Figur 5 OAIS model as applicabel to the Netarchive.dk

In the follow-up analysis of the material it was realised that a more detailed quality analysis of the harvested material was needed. This would have caught some of the problems that resulted in gaps in the archive, among other things the missing stylesheets and the many redirects.

The financial estimate that follows should therefore be seen in the light of these uncertainties.

7.1.1 Harvesting

As described in 5.4.3, the researchers used approx. 200 hours in carrying out the selection. They have estimated that a selection of material in connection with a subsequent election would take approx. 120 hours, while the librarians' work in preparing the URL base should be counted on top of this. The total extent of the input of the researchers was reckoned to be approximately 0.1 year of work.

As we saw in the section on voluntary agreements, this part of the work involves even greater problems. It took approx. one man-month to draw up agreements with 13 out of the 35 owners whose agreement had been sought. In all we had approx. 3,500 URLs in our database: If the remaining ones had required just as great an investment of time, it would have taken 42.5 man-months to secure the same proportion of agreements: e.g. 540 out of the 1453 requested. However, we may assume that this figure would be greatly reduced on a subsequent attempt, since a precedent would by then have been set for such agreements. Thus, even though the above figures indicate a requirement in the order of 10 man-years, we have reduced this in our estimate to 2 man-years, e.g. 24 months.

In the trial it was clear that even though standard software was used it was necessary constantly to make modifications and adjustments, to modify the settings and so on. We estimate, therefore, that in order to have a coherent procedure (e.g. one based on a single

harvesting machine) it would be necessary to allocate six man-months to IT-management in connection with the harvesting process.

While the harvesting is in progress the material should be checked to ensure that errors such as redirected do not spoil what is harvested. In the present trial this work is calculated to have required one man-month.

Total budget for a harvesting such as the one described in this report:

Function	Work-years
Selection	0.1
Agreements	2
Ongoing quality check	0.1
IT-management	0.5

This calculation is based on the assumption that existing software is used. An amount should therefore be set aside either for active participation in an international software development project or for the harvesting of the software.

7.1.2 IT development.

The amount required for IT development in a given project naturally depends on how many times a given item of software can be used before it becomes obsolete. Brewster Kahle, director of the Internet Archive has said that with the current rate of development on the Web it is necessary to renew the software every 18 months. It is difficult to put a price on this: Kahle estimates that the total cost for this development is 500,000 USD, which corresponds to approx. 4m DKR. The question is what proportion of this Denmark should bear. We have arbitrarily estimated the proportion at 5%. The figure is set high if we consider it in relation to Denmark's population, but we cannot expect that all countries will assume their share of the expenses; moreover, there will be additional expenses connected with the ongoing adjustment of the software in relation to specific needs.

Function	Cost
IT-development	200,000 per update

It is anticipated that updates should be done every 18 months.

7.1.3 Storage:

The cost of bytes storage depends on the quantity of bytes and on the storage strategy. If we look at three relevant storage institutions in Denmark we can see that there are three different storage strategies in use that to some extent reflect the present needs of the institutions in question.

- Public record office: Writes materials on 2 CD-ROMs and conducts detailed quality control.
- Royal Library: CD together with back-up tape, until it acquires its own system for long-term storage.

- State and University Library: Stores material on LTO tapes in a SUN mass storage system (robot).
- Archive.org has chosen to keep all material online and accessible via hard disk.

Granted the large quantity of material involved, the CD-ROM strategy is excluded. The choice is therefore between keeping the material live on hard disk or storing it on tapes, either in a robot (and therefore virtually accessible online), or on tapes that are stored in an archive.

Here we will focus only on the costs involved, and assume that a functional user interface can be designed.

The price of storage depends on the solution chosen. Three solutions are sketched out below:

Storage device	Initial cost	Price per TB
SUN with hard disk	280,000 DKK (inc. software for managin storage up to 15 TB)	200,000 DKK
SUN Robot	8,000,000 DKK	5,000 DKK
Intel-based system		50,000 DKK

In our further calculations we assume that the lifetime of a SUN robot is 5 years, for SUN server 3 years and for the Intel-based system 18 months. From the above it is clear that the Robot solution should be chosen only in so far as one has archiving needs, which it alone can cover (or if one can get a good offer). It is assumed in the calculations that the web archiving project is responsible for 10% of the expense of the robot.

All solutions would enable data migration to take place automatically.

If the materials are transferred to tape it is necessary to develop strategies on how such material can be preserved. As described in section 4.4.3, a simple transfer of all files will result in a system in which it would take many days to recover all the stored material.

Function	Work-years per year
Storage strategy and development	0.25

7.1.4 Storage of relevant software

It is meaningless to store a collection of bytes over the long term unless an effort is made at the same time to ensure that these can be decoded into an understandable signal. As already discussed, there are several strategies for ensuring this – and at the time of writing there is no answer as to which strategy will be most appropriate and when.

As with the software for harvesting, the solution to the problem lies in international cooperation.

The cost of this work will of course depend on which strategy is chosen. If the material is converted to a very limited number of formats there will be a high immediate cost, but a lower

cost in the long run. If one chooses not to do anything in the short run – let us say in the next 2-5 years – then one has presumably chosen the emulation strategy.

There are great uncertainties concerning this strategy and the investment of work required, and it is therefore difficult to estimate the cost of this part of the procedure. It is anticipated however that it will constitute a significant part of the total budget – particularly since it involves an ongoing process. The requirement here, including the development of a strategy for storage (6.1.3), has been very roughly estimated at 0.25 man-years.. This estimate does not cover potential active involvement in an international project devoted to these issues.

7.1.5 Access

There are several aspects to the question of access:

- The establishment of metadata (possibly an index) that can be used to find relevant material.
- The establishment of a user interface.
- Physical conditions (primarily server capacity) for servicing end users.

Concerning metadata there is an ongoing discussion among those concerned with storage as to how many metadata should be attached to digital objects. Here we make a minimal assumption, namely that metadata are generated automatically, e.g. mime-type information is preserved and a full text index is generated. It is possible to attach metadata even to an event, as the Library of Congress does, but the individual pages have no independent metadata. If the metadata are generated in this way the human input is minimal and the cost is simply the cost of the extra storage involved. In connection with NEDLIB we found that by generating an index we increased the amount of material harvested by 60%

Experience in establishing a user interface has been obtained primarily through the NWA project. Here it is estimated that it will cost one man-year per year. In our calculations we have not set a price on this contribution, since the international consortium under Brewster Kahle is also going to develop a user interface.

In so far as the material is stored in a robot, the necessary disks must be established online in order to give access. Irrespective of the strategy with regard to the storage of the byte collection, it is necessary to have a server devoted to this purpose. Below it is assumed that this server is a SUN with 1 TB harddisk/per user and that this has a lifetime of 3 years.

The above results in the following estimates:

Function	Cost per year
Metadata, automatic	60% increase in storage

7.1.6 Financial estimate

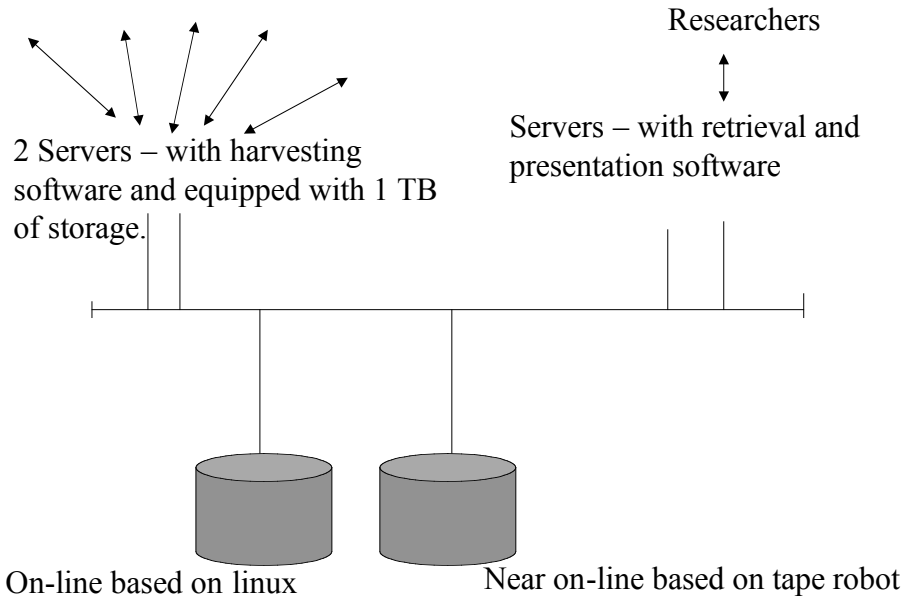
In the above section we estimated the costs on the basis of the experience gathered in the course of the pilot project. The costs fall into four groups:

- Web archiving infrastructure, the cost of applications and of an infrastructure that can be re-used for both harvesting and access. 0.25 year of work for the formulation of strategy

is included. These costs are the same regardless of whether one harvest is carried out or 100 harvests per year.

- Event costs: The cost of selecting and following sites, together with the cost of concluding agreements. This will depend on the size of the event.
- Cross-section costs: The cost of carrying out a cross-section harvest every quarter.
- Online harvesting. As a compromise among the partners involved it is assumed that we select 80 sites to follow.

In the pilot project one server was used for several purposes (harvesting, storage and access). We anticipate that in a production situation the following configuration would be necessary:



Figur 6 Architecture

With regard to the questions both of presentation and security it is proposed that two types of long-term storage should be used. It is suggested that one type would be a tape archive with LTO tapes. This has the advantage that one can get the system automatically to check the condition of the tape and that the cost of tape per TB is cheaper (but not the purchase cost). Second, it is proposed that there should be a server park consisting of PCs with Linux and a RAID based storage system.

All these items are reflected in the calculations given below.

It is assumed that the two harvest marchines would require upgrading of 1 Mbps in each of the two harvesting sites. This figure has been produced through retrospectively calculating from the expected quantity of material (altogether approx. 9 TB stored – e.g. approx 18 TB harvested). 18 TB per year corresponds to an average network load of approx. 0.6 Mbps.

In connection with the financial estimate the following prices and depreciations have been used:

The harvesting takes place with the use of SUN E250 with 1 GB storage, which costs 160,000 DKK per machine and 183,000 DKK per store. We assume that the lifetime of the SUN machine and the disks is 3 years. Two such machines would be purchased.

The project here would ”pay” for 20% of the purchase price of the tape archive and for the tape that is used. The purchase price is 4 million DKK and the archive has a lifetime of 5 years.

An online store based on PCs with disks costs 50,000 DKK per TB. As

The presentation configuration corresponds to the harvesting configuration. This investment is built in, but can be removed.

The cost of the network is estimated at 75,000 DKK per year per Mbps (average traffic increase). This figure is based on the State and University Library’s current payments and on the assumption that the Ministry of Science would not support the line with a 50% contribution (as is otherwise the case on the research net).

Finally, one work-year is assumed to have the value of 36,000 DKK per month.

Infrastructure

In calculating the cost of establishing and running the infrastructure we have assumed the following:

		No.	Unit	Cost per year DKK	Investment DKK
Harvesters	Server	2	SUN	106,667	320,000
	Store	2	TB disks	122,324	366,972
	Software	1	software	133,333	200,000
Presentation	Server	2	SUN	106,667	320,000
	Store	2	TB disks	122,324	366,972
Network	Bandwidth	2	Mbps	150,000	
Storage					
Store, location A	Server	1	SUN mass	160,000	800,000
Store, location B					
	Price of Linux store includes price of PC				
Storage strategy formulation		0,25	work-year	108,000	
Basic infrastructure				1.009,315	2,373,945

Depreciation is built into the cost per year. The amount of investment is therefore interesting only from the point of view of cash flow.

Event based

In the calculation below we have assumed that the average event is of the same size as in the pilot project.

1 event	Cost (DDK)

selection	0.1work-year	43,200
IT-management	0.5work-year	216,000
Quality control	0.1work-year	43,200
Voluntary agreements	*2work-year*	*864,000*
Storage requirement - store A	300 GB	300
Storage requirement - store B	300 GB	15,000
Total		1,181,700

The work input will depend on how many online media are selectively harvested. If we assume the number to be 80, no great change is anticipated in comparison with the pilot project. In the pilot project we had approximately 3,500 URLs in the URL base. Harvesting from 80 online media would alter this figure, but not significantly. The significance of the number of online media in terms of the workload will naturally depend on the type of event in question. As a rule the need for event-based harvestings is reduced if the number of online media included in continuous archiving is increased.

Cross-section

1 snapshot		Cost (DKK)
IT management	0.5work-year	216,000
quality control	0.05work-year	21,600
Voluntary agreements	*5work-year*	*2.160,000*
Storage requirement, store A	500 GB	500
Storage requirement, store B	500 GB	25,000
Total		2,423,100

In the calculations below it is assumed that four cross-section harvests would be carried out. This figure is based on discussion with partners in Netarkivet.dk, who argue in favour of four harvests on the grounds that this would reduce the extent of the considerably more expensive selective harvests

Selective – 80 online media

Selective: online newspapers, media – 80 agreements		Cost DKK
IT management	1work-year	432,000
Quality control	0,5work-year	216,000
Voluntary agreements	*0,25work-year*	*108,000*
Identification, agreement	0,25work-year	108,000
Storage requirement, – store A	8 TB	8,000
Storage requirement, – store B	8 TB	400,000
Total		1,272,000

The figure of 80 URLs has been much discussed, especially in the light of the proposed frequency of the snapshot harvests.

Total

The figure below shows the relative cost of:

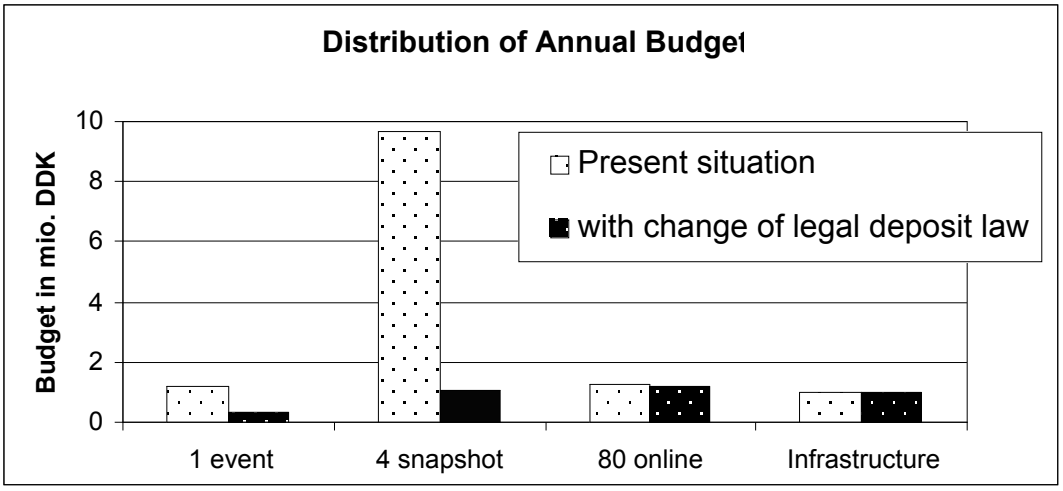
- 1 event based harvest.
- Four snapshot harvests.
- Selective harvesting of 80 online newspapers/media.
- Together with the basic expenses, which should be seen as a start-up cost that will be there if one aims to carry out harvesting at all, and the size of which depends to a large extent on how much one aims to harvest.

Calculations have been made both on the assumption that the law on legal deposit is changed, and on the assumption that it is not.

The costs given take into account that some form of international cooperation will be established. The impact of this is illustrated by the fact that the total cost in Sweden for two cross-section harvests, without any presentation system, is 3 million DKK whereas here the total cost of four snapshot harvests, together with the basic infrastructure, is approx. 2 million DKK.

The costs do **not** express the investment needed in order to get web archiving established, but are designed rather to give an average estimate. For example, if you carry out harvesting over a period of five years, the average costs over the five years is shown in the figure below, but the expenses – particularly with regard to large-scale hardware – will tend to come in large lumps at one time.

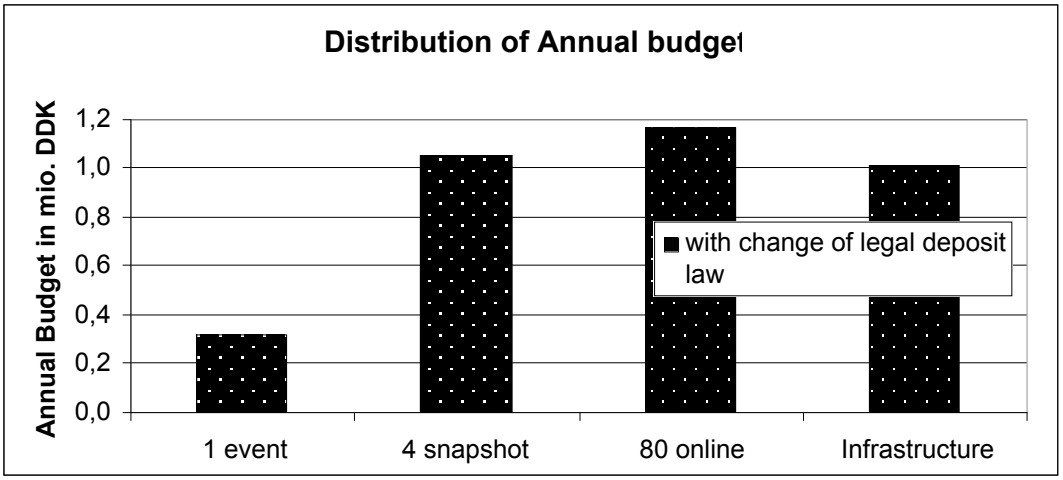
The total for the level of activities proposed is approx. 3.5 million DKK if there is a change in the law, and just over 13 million DKK if the work has to continue to be conducted on the basis of voluntary agreements.



Figur 7 The distribution on the different activities under the assumption of an unchanged legal law of deposit and under the assumption, that it is changed and there is no need to negotiate rights to aquire the material.

EMBED

As the figure is totally dominated by the expense to negotiate right, it is shown below under the assumption that there will be a change in the law



Figur 8 The distribution of the budget under the assumption, that the legal deposit law is changed. It should be pointed out, that the expense to storage is part of the three collections and not part of the infrastructure – in accordance with the nembers provided in this report.

EMBED

The above shows the cost of carrying out the level of harvesting proposed, and would buy sufficient storage capacity to enable the material to be preserved.

8 Conclusion

The pilot project has tested out a number of the tasks involved in web archiving, and has come up with some useful insights based on practical experience. The following are some of the main insights gained:

From a technical point of view we can immediately start harvesting from the Internet and thereby provide a useful foundation for analysing its current use. One of the conclusions that the researchers came to was that:

- despite problems in the pilot project, particularly in regard to presentation, it was judged that the material that can be archived will be sufficiently complete and reliable for a number of research purposes. This conclusion is based on the consideration that a large part of the shortcomings noted can be attributed to the pioneering nature of the project.

The pilot project also showed, however, both that continuous work is required to make web archiving possible: It is not sufficient simply to start up a software programme; and that the method has limitations in terms of the material that can be acquired.

- A good deal of material can be harvested with the existing software. However, there will remain certain things that for technical reasons cannot be harvested. The border onwhat can be harvested is constantly shifting, and we expect that it will continue to do so. Nevertheless, streamed material and problems with connection to links outside the archive currently constitute a serious problem.
- Continuous technical and professional supervision is required in order to carry out harvesting. The ongoing follow-up will concern both the quality of the material archived, and the identification of new formats and functionalities.
- Even if ready-made technical equipment is used, it is likely to require constant adjustment. The need for such adjustment in the case of the pilot project was due to some extent to the fact that the project's requirements differed from those envisaged, for example, by the developers of NEDLIB.
- The researchers came across many situations where the situation did not yield the desired results. An analysis illustrated that this was due to a combination of the presentation software, the configuration of the harvester and straightforward errors.
- The researchers found that the usability and reliability of the material as a research object seemed to increase when less material (for example, just one site) is archived, and when the material in question is not dymanic; whereas its usability and reliability are apparently reduced when larged quantities of material are archived and when the material is dynamic, as most of the Internet material is. One reason for this is that the material from larger and interactive sites was presented in a chaotic form, partly because of problems with the harvesting of dynamic material, and partly because of presentation problems.
- Many different types of competencies are required in order to carry out event-based web archiving. Professional expertise is needed to identify where relevant sites can be found, and detailed follow-up is required, partly to ensure that these relevant sites are included in the collection of start-URLs, and partly to ensure that they are actually harvested, and that this is done in the manner anticipated. Moreover, technical knowledge of harvesting software is required to make sure that technical problems do not disrupt the harvesting process, just as technical expertise is needed for participation in software development;

and conservation/curating preservation ? knowledge is required to make sure that the right means are used for long-term storage. Finally, technical knowledge is required to handle this storage (to define and implement long-term preservation strategies), as is the involvement of researchers who can represent the needs of future users, and so on.

- Considerable work is needed to negotiate voluntary agreements. If web archiving is to take place on any noteworthy scale, the necessary legislative basis for it needs to be established. Securing agreements with individual rights holders proved practically impossible – even though we met with plenty of goodwill. There are simply too many individual rights holders to be asked, and as a rule people do not respond to an initial request.

8.1 Recommendations

Over and above its general political and financial recommendation that the necessary economic and organisational foundation should be established for embarking on web archiving, the project has identified present legislation as the single biggest obstacle to this process.

Web archiving requires a proper legislative framework.

- The relevant institutions need to be given the legal authority to acquire and store material, just as the conditions to make such material accessible must be established. This could happen for example within the framework of the law on legal deposit.
- The definition of a 'work' within the current law on legal deposit is an impediment to developing the law. It is recommended that any forthcoming revision of the law on legal deposit should discard this definition, giving the concept of a 'work' the meaning that it has in ordinary Danish usage, and which is also used within the law on intellectual property rights.
- The present law on Personal Data authorizes the conduct of web archiving. Nevertheless, in order to ensure that personal data acquired in the course of web archiving are handled in full accordance with the law, it is recommended that a specific clarification should be made in consultation withThe Danish Data Protection Agency, Datatilsynet.

8.1.1 Strategy, economy and organisation

As mentioned in the conclusion, one of the experiences gained from the pilot project is that many different kinds of competencies are required in order to carry out event-based web archiving. An initiative should therefore be taken to acquire knowledge of web archiving and digital storage in general – for example by establishing a forum similar to the Digital Preservation Coalition (DPC)[15] in England involving the participation of a broad spectrum of people with different types of expertise and representing all different kinds of interest.

Much effort has been expended internationally on solving the types of problems involved in web archiving. It is important that Denmark should participate in some of these initiatives in order to ensure the necessary knowhow in this country, but it is also important to be aware that such participation requires resources, since participants are not welcome in the long run unless they are prepared actively to contribute.

[15] The role and activities of DPC can be seen from their homepage: www.dpconline.org

We recommend the creation of a group of experts in Denmark who could begin actively to preserve that part of the cultural heritage that is published on the net, and at the same time actively take part in both national and international initiatives concerned with digital storage.

The establishment of this body of expertise could have several phases and several spheres, depending on the financial situation. On the basis of the considerations that have been put forward in connection with this pilot project, we would suggest a several pronged strategy, consisting both of cross-section, event-based and selective harvesting, with the same infrastructure used in each case.

This report has indicated why such a severalpronged strategy is necessary. Even if the harvesting strategy is supported by the same software, there are different ways in which this software can be configured, and differences in the way the harvesting is handled technically. It is therefore suggested that expertise should be built up around different methods within different setups, instead of seeking to make all participating institutions expert in all aspects.

If we assume that the law on legal deposit will be changed, such that it will cease to be necessary to negotiate voluntary agreements, an up-front investment in the order of 2 million DKK per year will be necessary to ensure the preservation of Danish cultural heritage. This amount is considerably less than is used for example in Sweden, reflecting the fact that the calculations assume Denmark's participation in international development work in this sphere.

9 References

[Report 1] Interim report 1: http://www.netarchive.dk/rap/webark-report-1-nov2001.pdf (in Danish)

[Report 2] Interim report 2: http://www.netarchive.dk/rap/webark-report-2-dec2001.pdf (in Danish, is being translated)

[Report 3] Interim report 3: http://www.netarchive.dk/rap/webark-report-3-mar2002.pdf (in Danish)

[NWA] http://nwa.nb.no/

[Kulturarw3] http://www.kb.se/kw3/

[Pandora] http://pandora.nla.gov.au/

[Pligtaflevering] www.pligtaflevering.dk

10 List of appendices

The appendices listed below are collected in a separate document: Appendix to the Final Report for Netarchive.dk. All the reports below are in Danish.

Bilag 1:	The web site as the foundation for web archiving.
Bilag 2:	On the archiving of dynamic Internet materials (net events).
Bilag 3:	Description of test.
Bilag 4:	Graphic depiction of the archive.
Bilag 5:	Test form: Archive 'in itself'; archive-initial base; archive-monitoring; archive-other archives
Bilag 6:	Detailed commentary on test results.
Bilag 7:	Notes with comments and wishes concerning accessibility/searchability.
Bilag 8:	Notes on the draft French law: Loi sur la Société de l'Information — with particular reference to the question of legal deposit.
Bilag 9:	Completeness with regard to documents archived with NEDLIB.