WEB ARCHIVING

USER SURVEY

Marcel Ras, Sara van Bussel

National Library of the Netherlands
(Koninklijke Bibliotheek), July 2007

Contents

1. Introduction

In 2006, Time Magazine, the American weekly, chose 'the internet user' as the most important person of the year.[1] Time justified its choice by pointing to the new forms of community spirit and cooperation that 'the internet user' was displaying in a manner and on a scale such as the world had never seen before. According to Time, the unprecedented success of Wikipedia, YouTube and Second Life are examples of initiatives that 'not only change the world but also change the way the world changes'. The user is no longer just a consumer but is also a very important producer of content. The internet has made an enormous impact on daily life, on the way we offer and gather information and on the way we communicate and do business.

The influence of the web has also been felt in the Netherlands. Take the 'stemwijzer', the website that for many Dutch people was a decisive factor in forming their voting behaviour. More and more often we purchase items via the internet or base our purchases on comparison sites. Unneeded possessions change hands by way of market sites and bands become famous on YouTube. Anyone who has anything to tell the world keeps a weblog, and when it's time to book a trip or pay a bill we also rely on the internet. According to a study carried out by the Stichting Internet Reclame (Internet Advertising Foundation), internet use in the Netherlands rose by 8% during the first six months of this year compared with 2006. Almost 11 million people age thirteen and older spend an average of eight hours a week on the internet.[2]

Within the government, the cultural heritage community and academia the internet has also developed into the most important medium for distributing, exchanging and gathering information. More and more of this information is published exclusively on the web. The size of the Dutch portion of the World Wide Web has also grown by leaps and bounds. In 2006 this .nl domain had only existed 20 years, but it is impossible to imagine Dutch society today without it. In 2006 the two-millionth .nl domain name was registered,[3] making the Netherlands the fourth largest 'country code Top Level Domain' in the world.[4]

The growing dependence on the web has its flip side, however. The ease with which information can be deleted or changed makes the web highly vulnerable. On the other hand, we are increasingly coming to see the web as part of our cultural heritage.[5] Information on the web is often quite transitory and has a very brief life expectancy. If no action is taken, this digital heritage – which is important to future research on the development of the web and of society today – will be lost. Or as Malcolm Gillies has put it, '*The daily loss is already huge and we are at risk of losing large parts of our culture*'.[6]

1.1 KB and web archiving

In 2006 the National Library of the Netherlands (KB) started archiving a selection of Dutch websites. As the country's national library, the KB is responsible for the permanent storage of both printed and electronic publications. Because more and more publications are appearing in electronic form, storing them permanently and keeping them accessible has become a very important task.

---

[1] http://www.time.com/time/magazine/article/0,9171,1569514,00.html?aid=434&from=o&to=http%3A//www.time.com/time/magazine/article/0%2C9171%2C1569514%2C00.html%3Faid%3D434%26from%3Do%26to%3Dhttp%253A//www.time.com/time/magazine/article/0%252C9171%252C1569514%252C00.html%253Faid%253D434%2526from%253Do%2526to%253Dhttp%25253A//www.time.com/time/magazine/article/0%25252C9171%25252C1569514%25252C00.html

[2] http://stir.web-log.nl/stir/2007/07/internetconsump.html

[3] As of 25 July 2007: 2.48 million .nl domain names registered. www.sidn.nl

[4] Following Germany (.de), the UK (.uk) and Europe (.eu). Verisign, The Domain Name Industry Brief. Volume 4 – issue 3, June 2007. http://www.verisign.com/

[5] See article 1 of the UNESCO charter on the Preservation of the Digital Heritage. http://portal.unesco.org/ci/en/ev.php-URL_ID=13367&URL_DO=DO_TOPIC&URL_SECTION=201.html

[6] Malcolm Gillies, Born Digital Born Free? The Cultural Impact of the Web. Keynote Address, *Archiving Web Resources Conference, National Library of Australia* (9 November 2004). http://www.nla.gov.au/webarchiving/about_speakers.html#mgillies

Whereas most international initiatives began concentrating on harvesting websites at an early stage and are still following this approach as a general rule, the KB emphasises the long-term storage and permanent access of archived websites. This means that not only are websites being harvested but a strategy for long-term access is also being developed.

The complexity of this task is the reason why the KB did not start web archiving until 2006. From the very beginning, the KB has acknowledged the importance of digitally expanding its national depository function and has taken practical steps in that direction. In 1995 it began investing in research on the development and furnishing of an electronic deposit library. Since 2003, this e-Depot system has provided the KB with an infrastructure that makes it possible not only to store articles from periodicals electronically but also to guarantee the archiving of websites. In addition, the KB enjoys a leading position both nationally and internationally when it comes to researching the digital preservation of electronic publications. As a result, the KB has accepted the challenge within the Netherlands to archive websites as well.
Although websites have been archived since 1996 and web archiving is now being carried out on a large scale in many countries, very little research has been done so far on the users of these archives.

1.2 Selection

There are two basic strategies for web archiving. The first strategy focuses on harvesting websites automatically, usually in large quantities (bulk archiving of a national domain, for instance). The second strategy has to do with making selections based on a specific selection policy. Both strategies have their advantages and disadvantages. Automatic harvesting is relatively cheap compared with the selective approach, which is necessarily more labour-intensive. On the other hand, in harvesting a limited number of sites more attention can be paid to technical details and complete sites can be archived down to the deepest level. The selective approach allows for more tailor-made archiving in which the frequency of archiving can also be determined per site.

The KB has chosen this second approach. Aside from the fact that it allows for more customised work, there are a number of other reasons for making this choice. First of all, because of the absence of legal deposit legislation in the Netherlands it is extremely difficult to archive websites without any form of permission. A selective approach makes it possible to operate with greater care.[7] Second, bulk archiving is based on making a so-called *snapshot*. This means there are strict limits as to the number of documents and amount of data that can be archived per website: no more than x number of files per site and no more than y MB per site, to keep from accumulating an impossible amount of data and files. This is only possible if a national domain is very limited, which certainly is not the case for the Dutch domain. Since for the KB the basic motive for web archiving is permanent storage in the e-Depot, it does not seem like a very good idea to store only a limited portion of the websites. After all, we don't store only the title pages of books.

For the time being, the KB will base its selection of websites to be archived on the KB's own collection policy. Within this framework a well-reasoned selection will be made consisting of a cross section of the Dutch web domain. It should be noted that the Dutch web domain is a broad concept that is by no means limited to the .nl domain; it contains all the websites registered in the Netherlands. The primary selection will be taken from websites with academic and cultural content, although innovative websites that are examples of current trends in the Dutch portion of the web will also be considered. The next step will be to seek collaboration with other knowledge institutes for the purpose of

[7] Online publications are included in the deposit legislation of such countries as Denmark (http://www.bs.dk/content.aspx?itemguid=%7B332484E6-A5B1-4CEE-B953-059843182050%7D.) and Germany (http://www.ddb.de/aktuell/presse/pressemitt_dnbg_neu.htm). In France preparations are now being made to expand the existing legal deposit legislation so that websites will also be included.

broadening the selection, thereby making use of the substantive expertise of these organisations. Another idea is to give websites owners the opportunity to offer themselves up for archiving.

1.3 First phase and second phase

During the first phase of the KB project (January 2006 – May 2007) the goal was to acquire as much knowledge and experience of web archiving as possible. Consequently only a limited number of 100 websites have been archived so far. This provided enough information to make a fair estimate of the resources and infrastructure that will be needed. During this first phase, the 100 selected sites were crawled, involving more than 360 GB of data. These harvested sites consist of more than 16 million files with 200 different file formats. The second phase of the project is concentrated on embedding an infrastructure for web archiving and upscaling the selection to around 3,000 unique sites by the end of 2008. The intention is that this number will grow by the year and the selected sites will have to be archived a number of times per year. Given the amount of data and unique files that will have to be stored, as well as the abundance of file formats, it will take a lot of brain power to develop a strategy for permanent access.

1.4 Digital Preservation

Not until the websites are gathered, indexed and made properly accessible for the user does the problem really begin. How can we make sure that these websites will still be accessible to the user in 50 years or so? We won't be using the browers and platforms that we're now accustomed to using, and it may be that the concept of the web will have changed altogether by then. Even so, we must make sure that scientists 50 years from now can do their research, gather their data and put it to use. It is therefore realistic to assume that a great deal of that research data will come from web archives. The fact that our present websites are stored in the e-Depot is very reassuring, but it's not enough. We will have to do more. Active research will have to be conducted on how we can keep these sites accessible. Preserving the correct metadata so that people later on will be able to figure out what it is and how it should be presented is essential. Because the presentation of a website depends to a great extent on the browser being used as well as the plug-ins needed for the presentation of specific aspects of a website (such as Flash, video and audio), the most common browsers and plug-ins will have to be preserved in a software repository. The KB is doing intensive research on these aspects of web archiving (along with other organisations worldwide, whenever possible), including in the context of the International Internet Preservation Consortium (IIPC).[8]

1.5 Selection and user survey

Because the KB is still just beginning to develop a web archive, it is possible to profit from the knowledge and experience acquired by other organisations throughout the world in a number of different areas. We can certainly do this as far as the technical design of a web archive is concerned, but unfortunately little is yet known about the use and the users of web archives. A tour through the various web archives did provide us with valuable information, but it did not produce a coherent portrait of the user of a web archive. It soon became clear that a distinction had to be made between archives – and the use of those archives – that are based on bulk archiving and those that are included in a well-reasoned selection. In the first case an entire domain is archived in principle, and potentially anyone can be a user (as long as the archive is accessible). A user survey under such circumstances would most likely resemble a usability survey: how do we design the access and search function? A selection approach is based on selection criteria in which the potential use and the potential users are related to the selection being used. Here a user survey, besides being a usability survey, is also a survey of a target group and of user preferences. So the contents and the use of a web archive are closely connected.

[8] http://www.netpreserve.org

The KB has chosen to make a permanent archive of selected Dutch websites. This selective approach requires knowledge of the potential users. This user survey is meant to provide more insight into the preferences of potential users with respect to access and search functionality, but also with respect to the websites to be selected. Because the contents of the KB web archive is still limited at the moment to about 100 unique sites, the survey is qualitative and the results are mainly meant to provide more insight into the selection criteria and user preferences. A quantitative survey is being planned for 2009, when more content is available and the archive (online) is accessible.

The results of the present survey will also be decisive for the selection criteria and for the interface and search options to be put in place.

2. Design of the user survey

As noted earlier, a selective archiving approach requires that we have a profile of potential users. By this we mean the end users of a web archive. It is expected that the way they make use of the collection will be different from the way the traditional library is used. They will not constitute different target groups, however. So who is this user? And what does such a user want with regard to the selection, for example? Without answers to these questions it is impossible to set up a web archive for such a user. Besides the potential end users we must also take into account the owners of websites, the owners of content and other knowledge organisations who may play a role in the selection of the websites to be archived. These stakeholders were expressly involved in this user survey.

So that the preferences of potential users and stakeholders can be considered at the earliest possible stage, this user survey was conducted during the first phase of the project. It is a qualitative survey and will be used as 'food for thought' during the designing and implementation of the project's second phase. As soon as the web archive becomes operational and contains more content, a larger quantitative survey will be conducted.

During the survey, use was made of the web archive as it was designed in the first phase of the project. It contains approximately 100 unique websites in the areas of culture, government and science. Within the archive search operations can be carried out by URL (Wayback Machine) and by full text (NutchWax).

A comparable qualitative survey was conducted by the Bibliothèque nationale de France (BnF). The results of the BnF survey and of this survey will be combined with a survey on Access Requirements commissioned by the British Library.[9] The results of these three surveys will in turn be furnished to the Access Working Group of the International Internet Preservation Consortium (IIPC).[10] The aim is to formulate a definitive list of user preferences, on the basis of which existing search functionality and interface can be further developed for a web archive. A methodology for conducting a user survey designed to benefit a web archive will also be developed.

The survey's central question focuses on the contents and search functionality of the KB's web archive and the related preferences of end users and stakeholders. This central question is divided into six sub-questions having to do with users, use and selection. In order to answer these questions two methods were employed:

1. a target group analysis
2. observations of a test panel

2.1 Defining the problem

The central question of this user survey is:

> **What should the contents and search possibilities of the KB web archive look like, taking into account the potential users and stakeholders?**

The user survey is divided into a target group analysis and the survey itself. The central question can be divided into a number of sub-questions related to the target group analysis and the user survey.

1. Target group analysis

1.1 *Who are the potential users and stakeholders of the KB web archive?*
If the KB wants to take potential users and stakeholders into account, it will

[9] Simon Wild, *Web Archive Access Requirements*. British Library, June 2007.
[10] http://www.netpreserve.org/about/index.php

have to consider who these groups might be. Because the KB web archive is not yet operational, this question will have to be answered mainly through contacts with other web archives and organisations having to do with the digital heritage in the Netherlands. This will also involve having a look at the KB clientele as described in earlier surveys.

1.2 *What might the web archive be used for?*
To make the web archive more tangible, and to obtain and offer insight into the future use of the web archive, user scenarios will be formulated. These will be based on the *Use Cases for Access to Internet Archives*,[11] put together by the IIPC.

2. User survey

2.1 *How will survey participants search in the KB's web archive?*
The archive is being made accessible to the public and can be searched in a number of ways. We will look at how users carry out search operations by full text and by URL, paying attention to general search methods and to the search engines used by the KB for web archiving.

2.2 *What preferences or insights do survey participants have with regard to the selection of websites?*
Because the KB will have to select websites for archiving, it is important that the preferences of potential users and stakeholders be considered.

2.3 *How should access to the web archive be regulated, according to survey participants?*
There are several different possibilities for making the web archive available. Naturally the choice depends on the legal options, but the views of potential users will also be taken into account.

2.4 *How can user preferences be prioritised?*
This question arises from collaboration with the Access Working Group of the IIPC.

In this report the sub-questions will be discussed one by one, including a discussion of the method used. Results and recommendations are contained in the conclusion.

[11] Use Cases for Access to Internet Archives, <http://netpreserve.org/publications/reports.php?id=003>

3. Target group analysis

The analysis of the various target groups was carried out on the basis of:

1. discussion sessions with those involved;
2. specific questions about use and users of existing web archives;
3. analysis of the customer satisfaction survey, carried out in 2006 by TNS-NIPO

→ ***Who are the potential users and stakeholders of the KB web archive? (sub-question 1.1)***

3.1 Who are the users of other web archives?

In order to learn more about the use and users of web archives in general, a brief questionnaire was drawn up and e-mailed to sixteen existing web archives. The questions had to do with concrete information and experiences of use. A list of the organisations approached can be found in appendix 1. With regard to the approaches chosen by these sixteen web archives we can say the following:

- six respondents archive the entire national domain; bulk archiving
- two of these six employ a selective approach in addition to bulk archiving
- five respondents make a thematic selection (such as ministry websites or websites of political parties)
- six organisations archive a selection taken from their own national domain

Our limited survey reveals that knowledge of use and users depends very much on the how the web archive in question is accessed. A number of archives are still not accessible, so they have no end users. In some cases there are interested parties who request to have their website archived.

A number of the categories of end users mentioned are:

- Historians
- Sociologists
- Linguists
- Journalists
- Owners/designers of websites who want to have a look at an old version
- Public institutions who do not archive their own website and refer the public to the web archive
- The general public. This is the public that normally visits or uses the library as well

The survey also shows that when bulk archiving is employed there is less of a need for a user survey. In that case, knowledge of use and user will be more aimed at usability. This also has a great deal to do with how an archive is accessed.

- The Danish web archive (netarchive.dk) is only accessible after permission has been granted by the Danish Data Protection Agency, and only for researchers, which is a strict delineation of a target group
- The web archive of the Swedish national library (Kulturarw3) can only be accessed in the library. Their largest group of users consists of students of computer science. A few users have legal questions. Governments use the web archive to compare older versions of web sites with current web sites.

When a web archive is based on a selection, whether thematic or not, more attention is generally paid to specifying the target groups and determining how the archive is used, or will be used. Most of the respondents working with a selective approach indicated that they conduct research on use and users, or that they plan to conduct such research in the near future.

- The UK Government Web Archive is visited by government institutions, information professionals from the heritage sector, researchers and the general public.

- The Israeli national and university library has recently started conducting a survey of the preferences of potential users and stakeholders, mainly with regard to selection. The target groups that they continue to regard as potential users are instructors, lawyers, sociologists, journalists and linguists.
- The Portuguese web archive (Tomba) indicated that they want to conduct a user survey at some point in the future.
- Users of the web archive of the Library of Congress (Minerva) are mainly students, instructors, website owners and US Congress staff members.
- The British Library commissioned a survey of user preferences with regard to access to their archive.

Almost all the web archives we contacted are members of the IIPC.[12] Discussions have taken place within the Access Working Group of the IIPC on potential users of a web archive in general. User scenarios – *Use Cases for Access to Internet Archives* – have been drawn up for this purpose. These scenarios contain concrete examples for the use of a web archive. The users mentioned are: journalist, 'ordinary' citizen, government, lawyer, patent applicant, student, researcher, genealogical researcher and internal staff members.

3.2 Who are the customers of the KB?

In 2006, a survey was carried out by TNS-NIPO on how satisfied KB customers are with the KB.[13] These data are included in the web archiving user survey because the customer satisfaction survey clearly shows who the traditional KB customers are. Five hundred fifty-one people participated in a telephone survey of KB year pass holders. Five hundred two visitors to the KB website answered a few questions via the website.

Pass holders

The current KB pass holders can be divided into three main categories:

- the group that uses KB services for study purposes: 56%
- the group that uses them professionally: 23%
- the group that uses them for private study: 20%

The educational levels of the KB pass holders were not recorded in this survey.
The two areas in which pass holders are most interested are history (36%) and law (11%).

Visitors to the KB website

Of the visitors to the website who participated in the survey, 64% were year pass holders. The division of the group who reacted via the website, according to main categories, is somewhat different here:

- 38% of the visitors use the KB website for study
- 36% professionally
- 25% for private study

Most website visitors have degrees from higher professional schools or universities (70%). Only 4% did not go beyond secondary vocational training (MBO) or junior secondary general training (MAVO), and 22% did not go beyond senior general secondary training (HAVO) or university preparatory training (VWO). Although history is an important field of interest here as well (29%), social sciences (11%), bibliography, library science and archival science (10%), linguistics (9%) and humanities (9%) are more important than law (5%). Visitors to the website represent all ages, but most of them are between 21 and 65 (only 11% of those surveyed are above or below). Between 21 and 65 the distribution is equal.

[12] *IIPC about members*, < http://www.netpreserve.org/about/members.php>, consulted on 12 June 2007
[13] Nanet Beumer, *Klantentevredenheidsonderzoek Koninklijke Bibliotheek, Onderzoek onder pashouders en webbezoekers* (Amsterdam 2006)

3.3 Who are the possible stakeholders of the web archive?

There are several web archive initiatives in the Netherlands. A few thematic web archives are already operational (including Archipol and DACHS).[14] Projects have also been started by a number of large municipal archives that involve archiving the websites of the particular municipality (including the Rotterdam Municipal Archive). In addition, the KB has entered into a number of joint ventures with organisations that can play a role in the KB web archive as far as selection is concerned. These organisations can be regarded as stakeholders because their interest in the KB web archive is at a different level than that of the end user.

→ ***What might the web archive be used for? (sub-question 1.2)***

3.4 Division of the target groups and user scenarios

Based on the results of sub-question 1.1, a list of target groups for the KB web archive was drawn up. The following list of target groups is therefore based on the answers given in the surveys conducted among web archives, the IIPC use cases and the input made by a number of involved stakeholders. This list was then linked to the traditional KB target groups from the customer satisfaction survey.

Target group	KB target group	Professions
Researcher	Professional	Historian Sociologist Linguist
Journalist	Professional	
Jurist	Professional	Lawyer Public Prosecution Service Patent agent
Writer	Professional	
Information technologist	Professional	Web designer Web archive manager
Student	Study	
Consumer	Private study	
Genealogist	Private study	
Culturally interested person	Private study	

In the next step we formulated one or more user scenarios for each person from these target groups based on the example of the IIPC use cases themselves. Each scenario describes what the user wants to do with the web archive; what is necessary to accomplish this in terms of selection, frequency and access; whether these functionalities are offered anywhere else; and in some cases a link to existing material is provided (a newsletter or article in which the sketched situation appears). The KB has drawn up its own user scenario because the web archiving situation in the Netherlands is different than in most other countries.

As noted earlier, the Netherlands has no legal deposit that requires publishers to deposit published information; such deposits are made on a voluntary basis. From a legal point of view, statutorily required deposit has one important advantage. Under such a law, no permission is required for the harvesting of websites as long as the law is formulated broadly enough to apply to offline as well as online electronic publications. Because of the absence of a legal deposit in the Netherlands, the legal situation is different than in many other countries. This means that for the various stages of web

[14] Archipol, started in 2000, is an archive of websites of Dutch political parties set up by the Documentation Centre of Dutch Political Parties in cooperation with the Groningen University Library. DACHS is the Digital Archive for China Studies, set up by the universities of Leiden and Heidelberg. Websites are stored here as well as digital documents and films having to do with politics in China.

archiving (harvesting, archiving and making the archives accessible) we have to rely chiefly on the possibilities (and impossibilities) provided in the Copyright Act. In a legally watertight approach, the archiving of selected websites involves requesting permission officially by means of a contract. This entails an enormous administrative burden, however. For this reason the KB has decided to take a more pragmatic approach. This approach is based on an 'opt out' system; a site holder gives implicit permission for the indexing and caching of his website if he has not explicitly indicated otherwise (via robots.txt, for example).

The selective approach and the legal approach are incorporated in the user scenarios as the KB has formulated them. These scenarios should not be seen as a new or different version of the IIPC Use Cases but as a re-interpretation in terms of the KB's specific situation. The scenarios will be used during and after the survey to provide us with a better impression of how the web archive might be used in the future.

The scenarios are all formulated according to the same pattern:

- Who: What target group is the user scenario meant for?
- What: In what situation is the archive to be used?
- Requirements: What should the web archive offer so that the situation can be implemented?
- Is this function already being offered elsewhere?
- Practical example

An example of a scenario:

Who
Culturally interested person

What
Julia, an office worker, is very interested in the music of the late '90s. She knows that several small bands started out by promoting themselves via their own websites, so she would like to search the web archive for these special, lost sites.

Requirements
Biannual archiving
Search via WERA/NutchWax/Wayback Machine
Selection of cultural websites

Is this function being offered elsewhere
Partially. If the URL is known it may be possible to find these websites in the IA.

These scenarios are only examples (not the only possibilities for the use of the web archive) and will definitely not be used to limit options. The KB user scenarios can be found in appendix 2.

4. User survey

4.1 Survey design

The analysis of the target groups was mainly intended as a preliminary inquiry for obtaining more insight into who the users are, how a web archive is used or could be used and what user and stakeholder preferences are. Based on data from other web archives, investigation of the KB's present customers, input from stakeholders and the KB's specific approach, it was possible to draw up an overview of the potential target groups and the reasons why these particular groups would use the web

archive. This overview could then be incorporated into the central part of the user survey: the observations of actual users.

A number of persons from the defined target groups were invited to conduct search operations in a test version of the KB web archive on the basis of concrete assigned operations. During these observations a test version of the web archive was used because no final choices have yet been made regarding the search options and interface. It was deliberately decided to let the test person use a full-text search option as well as a URL-based search. For the latter the Wayback Machine was used, while the full-text search was made possible by creating a link between NutchWax and the Wayback Machine.

As noted, at the moment the KB web archive is limited to about 100 unique websites in the areas of culture, government and science. A number of these sites have been archived several times.

To learn something about the knowledge of the observation participants, a brief conversation was held prior to the test in which the participants were asked about how they make use of the internet. See appendix 3, user survey observation questions. This was followed by a brief explanation of the KB web archive and the test. The participants were then invited to search the archive on the basis of concrete assigned operations. Their actions were observed and recorded, and the user's comments were also taken down. The participants used the web archive on a laptop linked to an external monitor so the observer could also watch. The recording was made using the program Camtasia Studio 4 by Techsmith.[15] Following the test another conversation was held in which the users were asked about their experiences of the KB web archive and about their preferences with regard to the selection of websites, access to the archive and search functionality. See appendix 3.

4.1.1 Profiles of the participants

A total of fifteen persons took part in the user test. This group was mainly chosen from the KB's so-called patron panel. This includes KB patrons who have indicated that they are willing to cooperate with a customer survey for the KB.

1	Man, 26 years old, professional, researcher
2	Woman, 33 years old, professional, researcher
3	Woman, 43 years old, professional, lawyer
4	Man, 37 years old, professional, archive researcher
5	Woman, 37 years old, professional, curator
6	Man, 37 years old, professional, collection manager
7	Man, 28 years old, professional, researcher
8	Man, 43 years old, professional, press documentalist
9	Woman, 53 years old, professional, researcher
10	Man, 22 years old, study, student of communication and information science
11	Woman, 25 years old, professional, project leader
12	Man, 19 years old, study, student of business economics
13	Woman, 55 years old, professional, medievalist/Netherlands specialist
14	Woman, 48 years old, professional, primary school support staff
15	Woman, 41 years old, professional, communication

Unfortunately no participants were found from the private study group.

[15] This was a free demo, which works for 30 days and can be found on http://www.techsmith.com/camtasia.asp.

4.1.2 Prior knowledge of the participants

All the participants in the user test use the internet for work or study. Almost all the participants also use the internet for private purposes. Everyone uses the internet daily, whether at fixed times or throughout the day.

The internet is mainly used for looking up information and e-mailing (all participants). It is also used for chatting (four participants), personal weblog/website (four participants), entertainment such as reading the news or watching films (seven participants) and administration, such as banking (ten participants).

Only one participants did not know of the existence of web archives before receiving the invitation for this survey. One participant regarded the history function in Wikipedia as a web archive; this was the only function she knew and regarded as a web archive. A few participants regarded the cache function of Google and archives of the websites themselves as web archives.

Of the fourteen participants who had heard of web archives, nine participants had also actually visited a web archive. The most frequently visited and best-known web archive is the Internet Archive (nine participants), but visits had also been paid to the UKWAC web archive (two participants), Archipol (one participant), Library of Congress September 11 archive (one participant) and Pandora (one participant).

Five of the nine participants had visited a web archive 'just for the fun of it'. They searched for websites that they themselves had often visited or had made. The other four participants did research with the help of a web archive or used a web archive as a teaching aid. These were searching for information that had disappeared from the current web.

When asked about experiences of previous use of web archives, the participants remarked that they prefer to search by full text. There was also a strong preference for archiving full websites as opposed to a snapshot, and clarity with regard to selection and selection policy (what's in, what's not).

4.1.3 Tools used

To test how users want to search the web archive, we let the observation participants work with two search methods: a search engine that searches by URL and a search engine that searches by full text. The option of searching for archived websites via the KB central catalogue was not tested. An experimental link was made between the catalogue and the web archive, by which archived websites could be found via the catalogue in addition to the other KB collection. This was not tested on account of its current experimental character and in order not to offer the test participants too many search options.

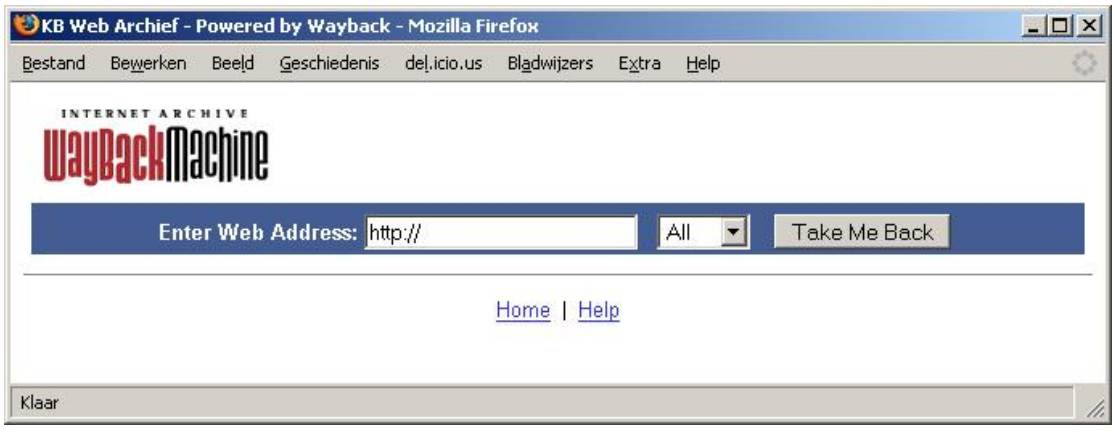


Figure 1. Wayback Machine search engine

The search engine that searches by URL is the Wayback Machine (figure 1). This search engine has no other search function.

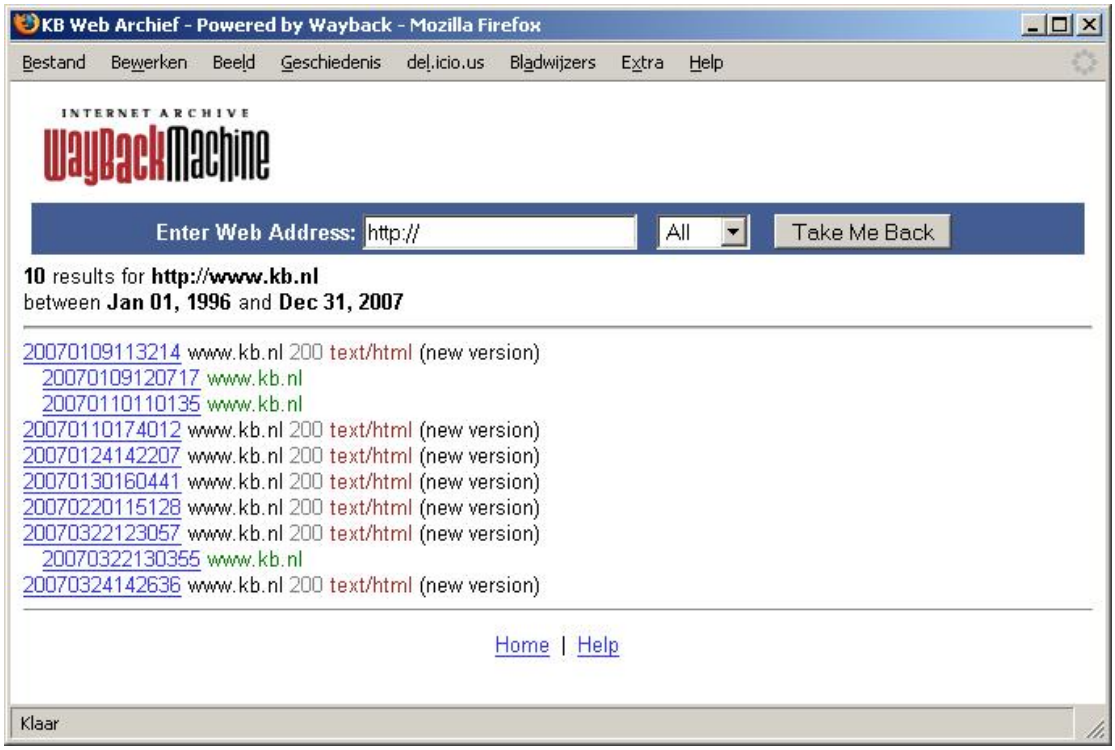


Figure 2. Wayback Machine search results

The data of the various archived versions of the web page being sought are presented as search results (see figure 2). Dates are rendered as a time stamp in year, month, day, hour, minute and second. All archived documents can be found, but the user must know the document's exact URL.

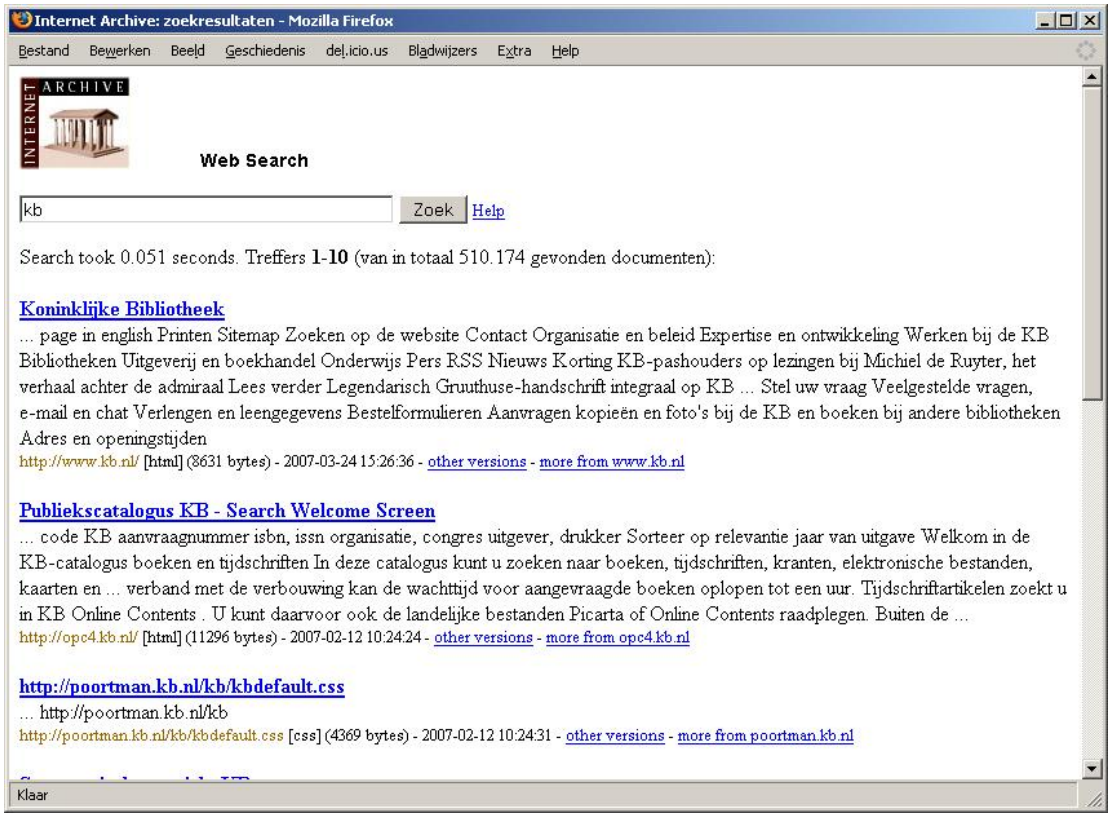


Figure 3. Search results of NutchWax in combination with the Wayback Machine

The search engine that searches by full text is a combination of NutchWax (for indexing and constructing the search functionality) and the Wayback Machine (for the presentation). With NutchWax the user can find archived documents in a Google-like way based on free text input. The search results (see figure 3) consist of:

- The title of the page
- A brief text
- The URL
- The file format
- The size (in terms of data)
- The date and time of archiving
- 'Other versions', a list of all the dates on which the page in question is archived
- 'More from this site' by which the search is carried out within the website

The advantage of this is that the user can search in the archive on the basis of free text input; he does not need to know any URLs. The disadvantage, however, is that the results often produce a large number of hits, often hits that are less relevant for the original search question. In addition, the current search engines are not yet able to handle the time dimension, which is important for a web archive.

Figure 4. Time bar in reproduction of archived website.

To reproduce the archived website both search engines make use of the Wayback Machine, so that the reproduced site is the same for both search engines. At the top of the website is a time bar (using a beta version of the Wayback Machine) in which the user can see how many versions there are of the page in question and what date of this version is (see figure 4). There is also a timeline with arrows for going to the first, the previous, the next and the last version of the page. The versions are drawn on the timeline, the scale of which can be adapted according to hours, days, weeks, months and years. The user can click on the blocks that stand for the various versions. The user can also look into the metadata of the page in question.

The operations that the test participants were asked to carry out were formulated in such a way that participants would be exposed to as many aspects of the two search methods as possible (see appendix 3: user survey observation questions).

→ ***How will survey participants search in the KB's web archive? (sub-question 2.1)***

4.2 Searching by full text

A few participants indicated that as far as the use of the KB web archive is concerned their preference is for a full-text search. One participant said that NutchWax, which for him was a new search engine, did not strike him as particularly reliable because it did not provide good results for his first search. Two participants during the observation said they missed an advanced search function.

As a last task the participants were asked to search a site of their own choosing in the Internet Archive (question 7). The size of archive.org was an eye-opener for many of the participants. The possibility of comparing many different versions of the same website, often many years in the past, clearly showed the usefulness of a web archive. It also became clear why the KB has decided to archive a limited selection of Dutch websites to their deepest level. The absence of parts of a site, which is unavoidable

in a bulk approach, is seen as a deficiency. The participants also regarded searches restricted to URL alone as a limitation.

4.2.1 General comments on the help text

The possibilities of the search engine are described in the help text. Ten participants said they never use the help text, or only in the case of emergencies. A few participants thought the text should more clearly indicate which search syntax can be used. For example, standard Boolean operators such as AND, OR and NOT cannot be used, but quotation marks, + and - can. This was not clearly explained.

The conclusion that can be drawn from this is that a more extensive help text is needed in which examples of the syntax are provided, and search tips on the search page itself would be handy.

The fact that visitors to a website rarely read the help text or the FAQs, if ever, was also noted by other web archives.

4.2.2 General impression of search behaviour

The participants combined several search terms. Interestingly, despite the fact that there are several pages with search results, almost no one continued on to the second page even though in some cases the answer was on that page. A few participants didn't even scroll down to find out what they were searching.

4.2.3 Using the search options

It is striking that most of the participants did make use of standard search syntax such as quotation marks, + and -.

To see whether participants were skilled with the full-text search engine, they were asked to search for specific file types. For example, they were asked to find a photograph of Princess Maxima in the web archive. One way to do this is to carry out the search operation with the addition of a file type qualifier (type:jpg). However, a photograph of Princess Maxima could also be found by means of the correct combination of search terms. A number of participants immediately referred to the way this can be done in Google.

In the end, four participants carried out the search operation by adding type:jpg. Two of them got this option from the help text. Ten participants found a photograph by means of a combination of search terms, such as 'princess maxima photograph'.

Three participants said that for this operation they preferred to search the website of the Royal Family via the Wayback Machine.

4.2.4 'Other versions' and 'More from this site'

The search results offer two additional options: 'Other versions' and 'More from this site' (see figure 3). 'Other versions' provides an overview of dates on which this page was archived – in other words, all the versions of this particular page. 'More from this site' provides search results of conducting the search operation within the archived website from which the page comes.

During the observation we looked to see if these extra options were being used by the participants. As it turned out, only six of the test persons saw then. The others hadn't even noticed.

In addition, the users expected different things from the two options. With 'other versions', a number of participants expected different versions of that particular page (not a realistic expectation). A

number of participants, however, expected different file formats or different versions from ‘more from this site’.

With ‘more from this site’ the expectations were more varied. Three participants expected this to be a link to the live website. One participant expected ‘interesting news about the website’. Other participants did expect search results from the website in question, but not results of this particular search operation. Two participants expected to see every place where this page appears, another participant expected to see everything from this page.

One participant tried ‘more from this site’ and said that the fact that the first ten search results produced many spacer.gif and .css files did not inspire confidence. These should be listed lower in the search results than pages with text.

Clearly the terms ‘Other version’ and ‘More from this site’ were not clear enough for participants in the user test.

4.2.5 Searching for an unknown website

The participants were asked to search for two different versions of the website of the museum “De Dubbelde Palmboom. Virtually none of the participants knew the URL of this website and therefore had to rely on the full-text search engine to find it. However, the search for De Dubbelde Palmboom often resulted in a sub-page of the web page, from which the participants could not go to other versions or to the homepage. So they either had to abbreviate the URL (up to and including .nl) or to try to find the main page by means of ‘Other versions’. Two participants even searched the URL on Google and used this in the Wayback Machine!

Two of the participants who searched in NutchWax did not realise that they had found a sub-page. After taking various detours, all the test persons finally did find both versions of the museum website (its homepage), but the operation was often quite laborious. One of the main reasons for this was that the search results in the full-text search engine are not hierarchically presented. Sub-pages are shown at the same level as a homepage. Even style sheet files and filler images are very often located at that same level. So the presentation of search results is in need of drastic improvement.

4.2.6 Searching within a website

The participants were asked to search in the web archive for information about web archiving on the KB website in such a way that the results show only kb.nl hits. The two ways to do this are by searching for web archiving in NutchWax and then to click on ‘More from this site’ on the hit from KB website, or by adding site:www.kb.nl to the search operation.

Five participants began by searching www.kb.nl or www.kb.nl/webarchiving in the Wayback Machine. They then wanted to search the website or the results but quickly discovered that was not possible. Two participants immediately searched ‘site:’ in NutchWax. Two participants looked in the help text of the Wayback Machine to see if the search operation could be carried out that way. They found out it couldn’t and went to NutchWax. Almost no one realised that the ‘More from this site’ option could be helpful in this case; this option was hardly used at all. At first almost no one made use of the help function. Only when no results were forthcoming did the participants consult help. It is recommended that a number of useful tips be included on the search page so the various options quickly become apparent. Even better: distinguish between a simple search page and an advanced search page.

4.3 Searching by URL

The best way to search by URL is to use the Wayback Machine search screen. This can also be done within NutchWax, but the test showed such a procedure did not produce optimal results, and users who started their URL search here quickly switched to the Wayback Machine. Nine of the participants

were familiar with the Internet Archive and also had prior experience with the Wayback Machine. When participants were asked to search for a website from a particular date, they almost all used the Wayback Machine.

4.3.1 General comments on the Wayback Machine

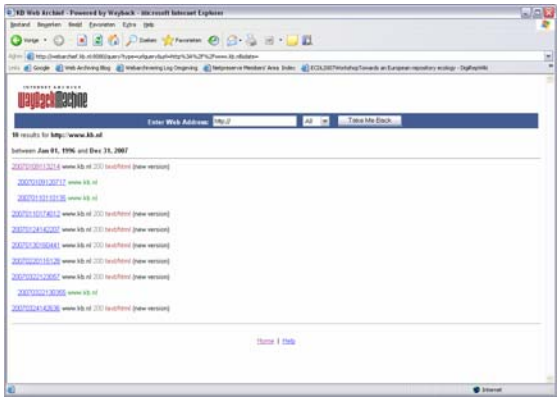
Two participants wanted to use the Wayback Machine but then typed their search task in the address bar of the browser. They said they had not seen the search window. One test person said the search window was too small, certainly for long URLs. Two participants selected the year of the website they wanted to find, the rest of the participants left this option on 'ALL'. One participant searched a URL without prefacing it with www, and that worked as well.

Two participants thought the standard Wayback Machine logo was confusing because they thought they were in the Internet Archive.

One participant said he wanted be able to search for 'pieces of URLs', such as with wild cards, if he didn't know the entire URL.

4.3.2 Reproducing the data

The result of a search operation is a line that consists of a time stamp of the dates when the searched website was stored and the original URL. In this test version of the KB web archive, a conscious decision was made to pay almost no attention to design aspects. For this reason the way the different versions of the same website are presented was sometimes unclear to the participants. The way this is shown in the Wayback Machine of the Internet Archive itself is much clearer.



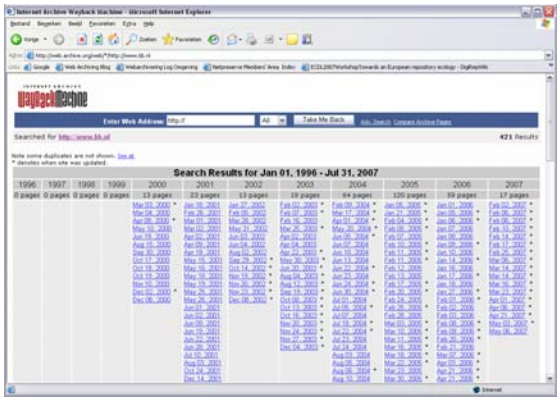


Figure 5: Presentation of results

After a bit of puzzling, most participants saw that these were dates and were able to decipher the time stamp. One participant checked her suspicions of the date by using the help text. While using the Internet Archive, the participants looked at the IA's presentation of the results. These are shown in readable dates listed per year in a table. Eleven participants said they thought this presentation was clearer.

4.4 The time bar

The time bar can be seen in both search engines and is the same in both, so it will be discussed separately.

One person did not see the time bar at any point during the test. A possible reason for this is that the time bar is the same grey colour as the browser's address bar above it.

Interestingly, fourteen of the fifteen participants tried to slide the arrow in the time bar to access an earlier or later version. When they found that this did not work, they clicked on the arrows at either end of the time bar. The blue or red blocks were hardly noticed or used. Two participants did mention the tool tip that appeared when they moved the cursor over a block with the mouse, and they found these handy. This tip is a brief text containing the date of the version which is hidden behind a block. The date of the reproduced version shown on the time bar was seen by virtually all the participants (after being asked what date the reproduced version was), as well as the number of versions.
To go from one version to another, ten participants used the time bar and the rest went one page back and clicked on another version.
After using the Internet Archive (without time bar) during the observation, ten participants found the time bar handy.

→ ***What preferences or insights do survey participants have with regard to the selection of websites? (sub-question 2.2)***

4.5 Selection

After the user test was over the participants were asked what kind of websites they thought should be included in the KB web archive. No list of categories had been drawn up from which they could choose, so the answers were quite diverse. The most important categories mentioned by the participants were:

- News sites
- Weblogs
- Cultural websites
- Government websites
- Scholarly sites

Other categories mentioned were: art websites, university websites, current events, politics, commercial, forums, archives, libraries, web technology, everything (!), science, cultural history, a cross-section of the Dutch web, private websites, 'unpleasant things' (like racism), municipal websites, disasters (such as crisis.nl), letters, museums (exhibitions and collection data bases), book reviews, performances and photo archives.

A striking number of participants (33%) expected the selection policy of the KB web archive to follow the general selection policy of the KB: websites in the area of the Dutch language, culture and society. The stakeholders also thought this was a good option, certainly as a starting point.

→ ***How should access to the web archive be regulated, according to survey participants? (sub-question 2.3)***

4.6 Means of access

The user test participants were all invited to test the web archive at the KB on a PC especially intended for this purpose. Because there are several different options for accessing the archive, the participants were asked what their preferences are with regard to access, taking legal restrictions into account.

As noted earlier, there is no statutory deposit framework for the KB web archive and the approach chosen is a pragmatic one with regard to legal problems. Harvesting and permanent archiving are not expected to cause very many problems, but this is not true of access to the archive. With access we have to deal with the copyright law and the protection of privacy under to the Personal Data Protection Act.

Access to a web archive can be regulated in several different ways, from free access on the internet to restricted availability within the walls of the KB itself. It goes without saying that all participants thought online access to a web archive was a more or less obvious preference, but they also realised

there are legal limitations. During the user test follow-up discussion, they were given four possible ways to approach the web archive. They were asked whether they would use the web archive if:

1. the web archive was only accessible inside the KB
   - Nine participants said they would not use the web archive in that case.
   - Six participants said they would use the web archive in that case. These are the participants who would use the web archive for research and who already use the KB facilities.
2. the web archive was accessible via for KB year pass holders via the internet
   - All the participants would use the web archive in this case, but seven said this would constitute a barrier for them.
3. the web archive was accessible via the internet after the registration of name and e-mail
   - None of the participants found this a problem. The director of the communications bureau said she did not think this was a problem because the KB is a reliable institution.
4. the web archive was accessible via the internet without registration or year pass.
   - None of the participants found this a problem (naturally). They said in this case they would occasionally look around 'for the fun of it' and that they would use it as casually as they use Google.

A number of participants said they would find it very strange if the web archive was not accessible online. The participants said they expect to be able to view websites via the internet, even if the site is a web archive. A few participants said that if the web archive was only accessible in the KB they might even suspect the KB of wanting to use it to earn money (!).

→ ***How can user preferences be prioritised? (sub-question 2.4)***

4.7 User preferences

In the user test follow-up discussion a list of user preferences was presented to the participants. This list was based on the list compiled by the British Library in their study of 'web archive access requirements'.[16] On the basis of participant experiences during the user test and this list of user preferences, participants were asked to indicate which conditions the KB web archive ought to satisfy. The ten most frequently mentioned conditions, in order of importance, were:

1. full-text searching
2. searching by URL
3. clear difference in the presentation between the archived website and the live website
4. possibility to search within one specific archived website
5. the presentation of the archived website must be the same at the original version of this site
6. the search results must indicate how a website can be accessed (freely or only in the KB reading room)
7. the possibility to search via the KB central catalogue
8. the possibility to search by key word
9. distinction made between simple searching and advanced searching
10. presence of metadata having to do with the archived websites; these metadata must be available to the user

As we said, these were the ten most frequently mentioned conditions. The full list contains more than thirty points. The purpose of this extra question to the participants was to place our findings next to those of the British Library. For this reason we consciously chose to base our list on that of the British Library. The resulting findings will then be incorporated in the study to be carried out by the Access Working Group of the IIPC, of which the KB is a member.

[16] 'The Access Requirements project will generate a list of requirements for accessing information stored within The British Library's web archive.'

5. Conclusion

The KB has decided to create an archive containing a selection of Dutch websites for long-term storage and permanent access. This selective approach requires a knowledge of users and potential users. So the central question of this user survey was: what should the contents and search options of the KB web archive look like, taking potential users and stakeholders into account?

The KB's present web archive is still in its test version and only a limited number of websites have been archived. In addition, very little is known worldwide about the use and users of web archives. Nevertheless, this user survey did provide enough information to enable us to take the next step. This qualitative survey is only a beginning. It has given us 'food for thought' and has provided us with input for a second phase of the web archiving project.

On the basis of data from other web archives, the *Use Cases for Access to Internet Archives* compiled by the International Internet Preservation Consortium, the KB's customer satisfaction survey and the input of various stakeholders, we were able to formulate user scenarios specific to the KB and to use them to create a user test.

The user test itself provided us with very valuable information about actual use, user preferences with regard to search functionality and interface, and clues for making selections.

The conclusions formulated below are answers to the sub-questions.

1. **Potential users and stakeholders (sub-question 1.1)**
   Despite the fact that there is still little information about the use of web archives, we were able to distil a picture of potential users from the IIPC use cases, our supplements to these cases, the input from other web archives and from stakeholders in the Netherlands, and the KB's customer satisfaction survey. The list with target groups consists of researchers, journalists, lawyers, writers, web designers, students, consumers, genealogists, people with a general interest in culture and the manager of the web archive. It will take a wider quantitative survey to determine if this is correct.

2. **Reasons for use (sub-question 1.2)**
   It is expected that the main reason for using the web archive will be research. As the web archive grows and ages it will become more of a source for scholarly research. Journalists and lawyers can use the web archive for background information. The web archive will also attract interested 'lay people', of course, as the user scenarios already show.

3. **Searching in the web archive (sub-question 2.1)**
   Searches in the web archive can be conducted in two ways, URL and full text. Searching by full text was clearly preferred by those who participated in the observation. Searching by URL was also felt to be very useful, however, and should certainly not be missing from the web archive. No matter what the search engine finally looks like, it should have a good help text. It is even more important that search tips for users be available on the main page. Clearly Google is the norm for the participants. They expected a search engine to work on full text, just like Google. One important aspect of a web archive is the time dimension. This will be expressed by the presence of a time bar, but time should also be incorporated in the search functionality and the presentation of the results of a search operation. A more hierarchical presentation of the results will also be important.

4. **Selection (sub-question 2.2)**
   The starting point for the selection of Dutch websites is the KB collection policy. The KB collection plan is concentrated on Dutch history, culture and society in an international

context. From this basis the archive can then be expanded with the help of users, stakeholders and specialists. Collaboration with other institutions involved in web archiving is desirable and can take place at different levels.

5. Access (sub-question 2.3)

People who participated in the observation saw no difficulty in having to register via the internet to obtain access to the web archive. If access to the archive were restricted to within the KB building alone only a few participants would make use of it, and then only as a last resort.

6. User preferences (sub-question 2.4)

Participants and stakeholders thought it was important that a clear distinction be made between the live website and the archived version. They also thought it should be possible to view metadata and information about the page, the archiving and the website. The participants wanted to be able to search the web archive in different ways, analogous to searching the live web. The decision to archive selectively was appreciated, certainly if this means the websites will be archived in detail.

Appendix 1: List of web archives consulted

Country/Project	Institution	Selection or bulk	URL
Archipol	Groningen University Library	Selection: thematic	http://www.archipol.nl/
Digital Archive for China Studies	Leiden University	Selection: thematic	http://www.sino.uni-heidelberg.de/dachs/
UK Government web archive	UK National Archives	Selection: thematic	http://www.nationalarchives.gov.uk/preservation/webarchive/default.htm
Germany	Bundestag Archive	Selection: thematic	
MINERVA	Library of Congress	Selection: thematic	http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html
UKWAC	UK national libraries	Selection: domain	http://www.webarchive.org.uk/
Czech Web Archive	Czech National Library	Selection: domain	http://en.webarchiv.cz/
Tomba	University of Lisbon	Selection: domain	http://tomba.tumba.pt/
Pandora archive	National Library of Australia	Selection: domain	http://pandora.nla.gov.au/index.html
Israeli Internet Sites Archive	Jewish National and University Library	Selection: domain	http://www.jnul.huji.ac.il/iae06/firstpage.html
Web Archive Singapore	National Library of Singapore	Selection: domain	http://was.nlb.gov.sg/wera/
Bibliothèque national de France	Bibliothèque national de France	Bulk and selection	http://www.bnf.fr/default.htm
Netarchive.dk	National Library of Denmark	Bulk and selection	http://netarchive.dk/index-en.php
Kulturarw3	National Library of Sweden	Bulk	http://www.kb.se/kw3/ENG/
Finland	National Library of Finland	Bulk	
Iceland	National Library	Bulk	

Appendix 2: KB user scenarios

Each user scenario indicates what target group is involved and what the members of that group can use the web archive for. Some user scenarios include an example from a news report.

I. Researcher

Who:

Researcher/Student

What:

Today's websites are snappy and offer the user many different personal options. Websites were not always like this, however. By comparing websites a student can analyse trends in web design, paying attention to functionality and appearance.

Requirements:

- Annual archiving
- Searching via the Wayback Machine
- Broad selection (personal/official/heritage/online merchants)

Is this function being offered elsewhere:

Partially. The archiving in some foreign websites can be very comprehensive (differs per country; see Pandora, Our digital island). This is not yet true of Dutch websites, however.

Who:

Researcher

What:

A researcher reads a colleague's research report and would like to track down the cited sources so he can check them and possibly use them for further study. It appears that the used website no longer exists. It can still be retrieved from the web archive, however, and the quotation can be checked.

Requirements:

- Annual archiving
- Searching via the Wayback Machine
- Broad selection (personal/official/heritage/weblog)

Is this function being offered elsewhere:

Partially, via the Internet Archive. The archiving there is not completely reliable, however, and does not cover the entire website. This means there is a good chance that the necessary source is not archived.

Practical example:

Diomidis Spinellis, *The decay and failure of web reference* (http://www.spinellis.gr/pubs/jrnl/2003-CACM-URLcite/html/urlcite.html)

Who:

Researcher (Sociologist)

What:

A sociologist wants to do research on youth culture on the internet. In addition to current websites he also wants to study 'old' sites so he can make a comparison between then and now.

Requirements:

- Annual archiving
- Searching via the Wayback Machine and full-text
- Selection of personal sites/weblogs

Is this function being offered elsewhere:

Partially, via the Internet Archive. The archiving is probably not deep enough, however, and this archive cannot be searched by text.

Who:

Researcher (Historian)

What:

A historian is researching the effects of the attacks of September 11 in the US on heritage institutions in the Netherlands. His assumption is not only that the security is more widespread but also that certain sensitive exhibitions that had been planned were cancelled. By comparing websites from before and after the attacks he can obtain an impression of the validity of his assumption.

Requirements:

- Biannual archiving
- Searching via the Wayback Machine
- Selection of heritage websites

Is this function being offered elsewhere:

Partially, via the Internet Archive. The archiving is not deep enough, however, and the heritage websites are often highly dependent on illustrations and Flash.

Who:

Researcher (Linguist)

What:

The linguists of the Netherlands Lexicology Institute (INL) do research on the Dutch language. The materials they use include corpora, large collections of texts. A selection is made from the web archive. The research tools of the INL are unleashed on this web corpus.

Requirements:

- Annual archiving
- Searching is not necessary
- Selection of Dutch websites
- Subset must be provided

Is this function being offered elsewhere:

No

II. Journalist

Who:

Journalist

What:

Rumours are spreading that the PBB, a political party, has provided support to extremist activities via an open letter to the owner of an extremist website. According to the rumour this person had published the letter on her page. When the journalist, J. Jansen, searches the website, however, the declaration of support cannot be found.
The journalist uses the Wayback Machine to search the web archive for earlier versions of the website. He finds an archived version of the website, including the letter, and is now able to write an article.

Requirements:

- Monthly/weekly archiving
- Searching via the Wayback Machine is sufficient
- Selection of weblogs/personal websites/forums

Is this function being offered elsewhere:

Yes, the Internet Archive also archives Dutch websites that can be searched by URL.

Who:

Journalist

What:

For an overview entitled '21st century, the first five years', a journalist is putting together a survey of 2005. Part of this is about life online, in which he describes the things that happened on the internet in 2005. To do this the journalist looks at archives of the websites themselves as well as the internet archive.

Requirements:

- Monthly archiving
- Searching via the Wayback Machine (with timeline!)
- Selection weblogs/personal websites/forums/news sites/government/heritage

Is this function being offered elsewhere:

Partially. The Internet Archive offers the necessary access and also archives these kinds of websites, but the frequency of archiving is too low to be useful.

III. Lawyer

Who:

Lawyer

What:

Jansen B.V. has discovered that a former employee is publishing inaccurate and damaging information about Jansen B.V. on his website. As preparations for a libel suit are being made, it is discovered that the website has been removed. Jansen B.V. wants to institute proceedings anyway on account of the damages caused by the information.

The lawyer of Jansen B.V. searches the website in the web archive and finds the exact text against which proceedings are then instituted.

Requirements:

- Monthly archiving
- Searching via the Wayback Machine is sufficient
- Selection of weblogs/personal websites

Is this function being offered elsewhere:

Yes, the Internet Archive also archives Dutch websites that can be searched by URL.

Who:

Lawyer/Public Prosecution Service

What:

Erik J. is arrested for selling medicines without a prescription on the internet. Because he only does business via his website, the website is used as evidence in the action against him. The web archive makes it possible to determine when the website was available on the internet. Reports can also be found on other websites on which Erik J. advertised his medicines.

Requirements:

- Biannual archiving
- Searching via the Wayback Machine is sufficient for Erik J.'s website. The reports are found after a full-text search (WERA/NutchWax)
- Selection of weblogs/personal websites/forums/online merchants.

Is this function being offered elsewhere:

Partially. Biannual archiving accessible through the Wayback Machine is available via the Internet Archive, but full-text searching is not a possibility here.

Practical example:

Webwereld: Duitser opgepakt na online verkoop zelfmoordpillen (German arrested for selling suicide pills online)
http://www.webwereld.nl/articles/44218/duitser-opgepakt-na-online-verkoop-zelfmoordpillen.html

---

Who:

Lawyer (Patent agent)

What:

The patent agent is approached by J. Jansen, who has come up with an invention. Jansen says he has not yet published anything about it. To check the feasibility of the patent request the patent agent visits the web archive, where he finds not only an announcement of the invention on J. Jansen's website but also a virtually identical invention on the website of a company that has gone out of business.

Requirements

- Biannual archiving
- Searching via the Wayback Machine for J. Jansen's website. Prior inventions were found via the full-text search possibilities of WERA/NutchWax
- Selection of personal/companies

Is this function being offered elsewhere:

Partially. Biannual archiving accessible through the Wayback Machine is available via the Internet Archive, but full-text searching is not a possibility here.

IV. Consumer

Who:

Consumer/Lawyer

What:

A consumer orders an item on the internet. Upon delivery it is discovered that this is not the correct item. The consumer sends the item back under the assumption that he will get his money back. Referring to the general terms and conditions on the website, the web shop refuses to refund the full amount. The consumer says that before making his purchase he read in the general terms and conditions that if his order is returned within seven working days he will be refunded the full amount. In the intervening period the general terms appear to have changed. The consumer searches the web archive for the old general terms and conditions, and it turns out that he is correct.

Requirements:

- Biannual archiving
- Searching via the Wayback Machine
- Selection of online merchants

Is this function being offered elsewhere:

Partially, websites like Wehkamp and Otto – including the General Terms and Conditions – are partly accessible via the Internet Archive.

Practical example:

Webwereld: 'Algemene voorwaarden Otto niet van toepassing' (Otto's general terms and conditions not applicable)
http://www.webwereld.nl/articles/43564/ictrecht

V. Other target groups

Who:

Writer

What:

A writer discovers that his work is being published in its entirety on the internet on websites other than his own and without his permission. After investigating the matter it seems that the website he has now found is not the only website that has taken over his work. He would like to know which websites published earlier work and have now removed it.

Requirements:

- Biannual archiving
- Searching by full text (WERA/NutchWax)
- Selection of weblogs/personal websites

Is this function being offered elsewhere:

No, Dutch websites are not yet fully archived and accessible via a full-text search engine.

Who:

Web designer

What:

The web designer S. de Jong is commissioned to design a website for a particular company. To make sure the corporate identity of the company is preserved in the website, he looks in the web archive for earlier versions of the sites. He bases his design on earlier websites.

Requirements

- Biannual archiving
- Searching via the Wayback Machine
- Selection of companies
- High-quality archiving (including illustrations)

Is this function being offered elsewhere:

No, high-quality archiving of Dutch web sites is not yet being done (except for a few big companies, which can be found on the Internet Archive)

Who:

Genealogist

What:

The genealogist W. Steen regularly uses the website of one of his distant relatives. This website contains a great deal of information on the history of the Steen family. One day Mr Steen goes to visit the website again but cannot find it. After some time he realises that the website has disappeared. Mr Steen searches the website in the web archive and stores the information from the site.

Requirements

- Biannual archiving
- Searching via the Wayback Machine
- Selection of personal websites

Is this function being offered elsewhere:

Partially. Personal website can be found in the Internet Archive via the Wayback Machine. The archiving is severely backlogged, however, especially personal websites.

Who:

Culturally interested person

What:

Julia, an office worker, is very interested in the music of the late '90s. She knows that several small bands started out by promoting themselves via their own websites, so she would like to search the web archive for these special, lost sites.

Requirements

- Biannual archiving

- Searching via WERA/NutchWax/Wayback Machine
- Selection of cultural websites

Is this function being offered elsewhere:

Partially. If the URL is known it may be possible to find these websites in the Internet Archives.

Who:

Web archive manager

What:

To make sure the web archive continues to be a good collection, the people who work on it must also use it. They should check to see if the archiving is being done frequently enough, if the selection is comprehensive and if the archive is running well (if the links work, illustrations are visible, etc.)

Requirements

- No requirements for frequency, searching or selection

Is this function being offered elsewhere:

No, because this has to do with the management and maintenance of a particular archive.

Appendix 3: User survey observation questions

I. Experience with the internet and web archives

1. Where do you use the internet?
    a. Private use
    b. School
    c. Work

2. How often do you use the internet?
    a. 1 – 2 times a week
    b. 3 – 4 times a week
    c. Every day

3. How long are your internet sessions?
    a. One hour per session
    b. A few hours per session
    c. The whole day with breaks

4. What do you use the internet for?
    a. Searching for information
    b. E-mailing
    c. Chatting
    d. Maintaining my own weblog/website
    e. Entertainment (watching films, reading weblogs, reading the news)
    f. Administrative (banking, library, etc.)

5. Were you aware of the existence of web archives?
    a. Yes
    b. No

6. Have you ever visited a web archive?
    a. Yes (which ones)
    b. No

7. What did you look up in the web archive/what kind of information were you looking for?

8. Do you have any comments or criticisms regarding your earlier visit to a web archive (was there something you missed or something you really liked?)

II. Test tasks

1. On the main page of www.kb.nl the Gruuthuse manuscript is now featured as the focus point. What was on the website of 20 February 2007?

2. The appearance of the Dubbelde Palmboom and TNO websites has changed completely since March 2007. Both versions of the websites can be found in our web archive. Can you find them?

3. Can you find a photograph of Maxima in the web archive?

4. Cam you find a PDF document in the web archive with the file name ‘nieuwe woorden’?

5. Searching the web archive, can you find information about web archiving within the KB website, making sure you only see results from the KB website?

6. Say it’s 2017 and you want to write an article about the PvdA (Labour Party) in the build-up to the Provincial State elections of March 2007. How would you search for information in the web archive?

7. Go to the website of the Internet Archive and search for a website that is interesting to you. As you do so, compare your experiences with the KB web archive with those of the Internet Archive.

III. Follow-up discussion

Observation

1. What did you think of this?
2. Did you run into any problems?
3. Are there options you missed?
4. Were some things unclear?
5. Do you have any ideas for improvement?

Web archive

1. Do you think you will use a web archive in the future?
2. If this web archive existed, what would you like it to contain (what kinds of websites?)
3. If this web archive could only be accessed in the Koninklijke Bibliotheek would you make use of it?
4. If you could access web archive on the internet with a KB year pass, would you make use of it?
5. If you could access this web archive on the internet after registering your name and e-mail address, would you make use of it?
6. If this web archive was freely accessible on the internet, would you make use of it?

Appendix 4: Literature used

Beumer, Nanet, *Klantentevredenheidsonderzoek Koninklijke Bibliotheek, Onderzoek onder pashouders en webbezoekers* (Amsterdam 2006)

Beunen, Annemarie and Tjeerd Schiphof, *Juridische aspecten van webarchivering* (Leiden 2006)

Booij-Sleutel, Christien and Saskia van der Linden, *Vier artikelen over Usability* (n.p. 2006)

> Booij-Sleutel, Christien and Saskia van der Linden, *De kluwen ontward, hoe maak je een analyse en plan voor een ontwerp of herontwerp van een website?* (n.p. 2006)
>
> Booij-Sleutel, Christien and Saskia van der Linden, *Lessons learned, de bijlage: Voorbeelden van usability-problematiek uit de archiefpraktijk* (n.p. 2006)
>
> Booij-Sleutel, Christien and Saskia van der Linden, *Schatgraven zonder zweetdruppels, welke evaluatie technieken zijn er en in welke fase kunnen deze ingezet worden?* (n.p. 2006)
>
> Booij-Sleutel, Christien and Saskia van der Linden, *Voor Antonius en Johanna tot Anouk en Jaimie, gebruiksvriendelijk ontwerpen, wat houdt dat in?* (n.p. 2006)

Foot, Kristen et al., *The internet and elections, an international project for the comparative study of the role of the internet in the electoral process* (2003)

Foot, Kristen and Steven M. Schneider, 'The web as an object study', *New Media and society* 6 (1) 114-122

Gomes, Daniel, Sérgio Freitas and Mário J. Silva, *Design and selection criteria for a national web archive* (Lissabon 2006)

Hokke, Erika, 'Een toekomst voor het verleden van het web', *Informatie Professional* 7/8 (2006) 26-31

IIPC about members, < http://www.netpreserve.org/about/members.php>, consulted on 12 June 2007

IIPC, *Use Cases for Access to Internet Archives*, <http://netpreserve.org/publications/reports.php?id=003>, consulted on 1 February 2007

Koninklijke Bibliotheek Collectieplan 2006-2009, Nederland in de wereld - de wereld in Nederland (The Hague 2005)

Lyman, Peter, *Archiving the world wide web* <http://www.clir.org/pubs/reports/pub106/web.html>, consulted on 27 February 2007

PADI webarchiving, *secties 1 (purpose of page), 2 (web archiving models), 3 (case studies of archiving approaches)* <http://www.nla.gov.au/padi/topics/92.html>, consulted on 18-12-2006

Pandora user survey, <http://pandora.nla.gov.au/usage_survey_form.html>, consulted on 18-12-2006

Ras, Marcel, *Projectplan Eerste Fase Webarchivering* (The Hague 2006)

Rauber, Andreas, et al., 'Uncovering information hidden in web archives', *D-Lib magazine* Dec. 2002 vol. 8 no. 12

SIDN, *Jaarverslag 2006* (2006)

Sieverts, Eric et al., 'IP webtest VIII, Wetenschappelijke bibliotheken: meer' *Informatie professional* 6/12 (2006) 16-33

Stack, Michael, *Full text search of web archiving collections* (San Francisco n.d.)

UKWAC evaluation report, appendix 2: End user survey form for the UKWAC web site - Summary of responses (2006)

Verisign, *The domain industry brief* 1 (2007)

Voorburg, R.J.J. and J.L.E. Goutier, *Web sphere analysis: an approach to studying online action* (n.p. 2004)