

Long-term preservation and access of the Dutch Web

Marcel Ras, Barbara Sierman

National Library of the Netherlands (KB)
Digital Preservation Department, Research & Development Division,
P.O. Box 90407 2509 LK The Hague, the Netherlands

marcel.ras@kb.nl, barbara.sierman@kb.nl

Phone: +31 70 3140 1180

www.kb.nl

Abstract

For the past years the National Library of the Netherlands (KB) concentrated its efforts on developing and optimizing its system for long-term preservation. Now that the e-Depot is fully operational, it is possible to concentrate on the development of new services. The KB started a web archiving project, focusing on the Dutch web space in 2005. Unlike many other web archiving projects, the KB project is based on securing long-term storage and permanent access for websites. The main goal of the project is to develop a strategy for long-term preservation and permanent access of Dutch websites based on the existing e-Depot system and infrastructure.

Introduction

In 2005 the Koninklijke Bibliotheek, the National Library of the Netherlands (KB), started a project to investigate the organisational and financial consequences of archiving the Dutch web. This project is connected with other activities of the KB in the field of long term preservation of electronic publications.

The KB is a medium sized national library, with 3 million paper volumes and over 6 million electronic items in deposit. It has a staff of 275 full time equivalent people and is structurally funded by the Ministry of Education, Culture and Science. Since 1974 the KB acts as a deposit library, although the Netherlands does not have a legal deposit obligation. As National Library of the Netherlands the Koninklijke Bibliotheek (KB) is responsible for the collection and preservation of all Dutch publications. To ensure the long-term preservation of the increasing number of electronic publications, the KB developed the e-Depot, the world's first long-term digital archiving system for academic publications.

The e-Depot

In 1994 the KB extended its depository task to include digital publications. Consequently, the library needed to think about the consequences of this new task. Activities were

started to develop an electronic archive to store this electronic material for the long term. From 1998 to 2000, the KB was project leader of the European project NEDLIB (Networked European Deposit Library). This project has meant a lot for the enhancement of our knowledge of digital preservation.

When a market survey in 2000 made it clear that no ready to use systems existed at that time, a European call for tender was issued to find a partner to build a deposit system for e-publications, based on the OAIS model and the NEDLIB Guidelines. In 2002 the selected partner IBM delivered the DIAS system (Digital Information Archival System), based on standard software and hardware components of IBM. DIAS is the technical heart of the e-Depot. Its functional design is based on the OAIS Reference Model, an ISO certified system for digital archiving. As a result, DIAS's high-quality technical and organizational infrastructure meets criteria of durability, flexibility and scalability.

Currently the e-Depot contains more than six million publications, a number which is expected to double in a few years. Initially archiving agreements were only concluded with publishers of Dutch origin. However, it soon emerged that the traditional concept of national deposits is less relevant for electronic publications without geographical ties. Furthermore, various international publishers showed an active interest in the e-Depot. Consequently, an international approach was adopted.

At present, agreements have been signed with a number of large scientific publishers. More than 3500 journals, around seventy per cent of the total number of academic journals in these disciplines, are stored in the e-Depot now. The KB hopes to expand this number in the coming years. Details on deposit conditions and terms of access are agreed upon with every individual publisher.

The KB keeps developing new services for the e-Depot. Besides academic journals the KB will focus on other types of publications such as e-books, images created in mass-digitization programs, audio books and websites. Furthermore, studies are carried out into the potential role of the e-Depot as a digital archive for academic and heritage institutions in the Netherlands.

Research in the KB

To succeed in maintaining permanent access to the objects in the e-Depot, digital preservation research has been made a key priority. Information and communication technology are characterized by rapid technological change. New versions of hardware and software succeed each other, causing digital material to become obsolete fast. The KB carries out innovative research, often in collaboration with fellow libraries, archives and research institutes, both in the Netherlands and abroad.

Digital preservation research in the KB mainly concentrates on Preservation Planning. To keep digital objects accessible over the years, tools and services are needed that will enable us to make sound decisions. For this purpose a Digital Preservation department, with a staff of seven people, was established. The web archiving project is part of the activities of this department. Next to this, the KB participates in the European Planets project, which will deliver tools and services for Preservation Planning.

Day-to-day maintenance of the e-Depot system is executed by the ICT department. Operation of the system, as loading of publications and processing of metadata, is executed by the e-Depot department.

The Dutch approach to web archiving

The World Wide Web has quickly developed into an important medium of communication. It plays an important role in social, cultural and academic exchange and also in business transactions. Governments are using the Web for communication purposes as well, which makes the Internet the main source of information for a growing number of people. Increasingly, information is published exclusively on the Internet and no longer in print.

At the same time, the Web is changing constantly. Millions of new pages appear every month, while others are altered or disappear altogether. Although it is easy to publish information online, its origins and quality are often much more difficult to establish. Furthermore, it is short-lived: it has been estimated that the average lifespan of a web page is only 75 days. Some historians have already suggested that the internet will cause a gap in history. If nothing is done to prevent it, this type of digital heritage¹ will disappear. As a consequence, valuable information about the development of the Web and our present-day society is lost for future researchers.

The first efforts towards creating national web archives were made in 1996 by the Australian and Swedish national libraries and the Internet Archive². Since then, a large number of national libraries, as well as other institutions like universities have been involved in web archiving projects.

Until the start of the KB web archiving project at the end of 2005, only small scale projects had been carried out in the Netherlands, mainly by universities, municipal archives and governmental bodies.

As national library the KB is responsible for collecting, describing and preserving all Dutch publications. As websites can be seen as publications, the KB sees it as its task to collect them. For this reason, the KB started a web archiving project, focusing on the Dutch web space. The main goal of the first phase of the project is to develop a strategy for long-term preservation and permanent access of Dutch websites based on a selective collection approach.

The KB project has a set of special characteristics that distinguishes it from most other web archiving projects:

- It aims to create a system for the long-term preservation of Dutch websites
- The system is based on the existing ingest process and workflow of the e-Depot
- It will also fit in the existing KB infrastructure
- It will ingest a different type of content in the e-Depot

The fact that the KB already possesses a reliable electronic depot can be considered an advantage. On the other hand, this could well turn into a disadvantage as the intended web archive is committed to the existing system. This project will show whether the existing system and workflow meet the requirements of a national web archive or whether extensions will be necessary.

Dutch web space

No exact figures are available on the size of the Dutch web space. Based on surveys of the OCLC Web characterization project and Netcraft it is possible to get a broad idea however³. Firstly, the size depends on how the Dutch Web is defined exactly. Apart from all websites hosted under the .nl domain, does it also include all .com, .net, .org, .eu domains that are written in Dutch or hosted in the Netherlands? In May 2006 about 1.9 million .nl domains had been registered by the central administration of domain names (SIDN). Nevertheless, not all of these registered names refer to unique web sites. For instance, www.kb.nl, www.koninklijkebibliotheek.nl, www.koninklijke-bibliotheek.nl and www.konbib.nl all refer to the same site, the site of the KB. Based on surveys it is estimated that the Dutch web consists of approximately 1,5 million unique websites, covering at least 80 million web pages. However, these figures only refer to static websites and web pages that can be indexed by search engines. Websites that are constructed dynamically out of databases and content management systems cannot be indexed by search engines and are therefore not included in these numbers, whereas research has suggested that these types of pages, usually referred to as the “deep web”, exceed the static pages by a factor 400.

The project

For the past years the KB concentrated its efforts on developing and optimizing its system for long-term preservation. Now that the e-Depot is fully operational, it is possible to focus on developing new services such as web archiving. For the KB, web archiving is not an isolated activity. The KB intends to incorporate web archiving into the existing workflow and infrastructure of the organization in general and the e-Depot in particular. The project team that has been set up therefore consists of staff of all divisions of the organization: Research & Development, IT, Acquisitions & Processing, User Services, Selection & Information Services. As this project will be the KB’s first step in web archiving, it was decided to take control of the entire process and take up developments in-house, in order to build-up knowledge and experience.

The five key deliverables for the project are:

1. A selection and harvesting workflow based on a selective approach, by using the IIPC toolset
2. A workflow and strategy for ingest based on the existing e-Depot workflow and infrastructure
3. A strategy for long-term preservation and permanent access of websites
4. A searchable archive of websites collected and catalogued for the benefit of a designated community
5. A strategy on how to deal with IPR and a common permissions form for archiving and preserving websites
6. An evaluation report providing a set of recommendations and a strategy on how to proceed with long-term preservation of Dutch websites and a vision on preservation planning

Selection

As for gathering Dutch websites, the KB has decided to adopt a selective approach. Selection policy is based on the general collection policy of the KB and will contain

information on Dutch history, language and society. For practical reasons selection within the first phase has been limited to websites that offer us as many technical challenges as possible. By first making a relatively small selection of websites to harvest, the KB will have the opportunity to learn without having to collect a vast amount of data.

Nevertheless, at an early stage a more detailed selection policy will be developed, based on the general collection policy of the KB. For this, we rely on the knowledge of specialists as they have to monitor the Web within their field of expertise.

Harvesting

One advantage of not being an early adapter is that we can benefit from experience of past and current international initiatives. Choosing a suitable crawler was not very difficult; the Open Source crawler Heritrix will be used as it is a widely used crawler and part of the IIPC toolset. There is an active community of Heritrix users, so developments within the KB project can easily be shared with others. Within the limited selection only the seed URL's will be gathered. In this crawl we shall collect all possible MIME types without limitations to the internal depth of the selected sites and a broad download time. External links will not be followed from the seed list. Decisions relating to the depth, width and frequency of future crawls have to be made based on the experiences of the first crawl. The next step will be to implement a structural harvest strategy for a selection of websites and to set up an event-driven crawl in 2007.

Access

Searching the Internet can be a painstaking experience as it will provide you with a long list of results. Only if you are lucky you will find a relevant hit on page ten. The same problem will occur in a web archive that contains a vast amount of data. We propose a two way solution to solve this problem. A high-quality index of the web archive and a usable interface is part of this. As for indexing and providing access, we consider to use the tools recommended by the IIPC as these have proven to be standard solutions used by many web archives⁴.

In addition to this, a standard description, even on a minimum level, can help the user. We are considering a basic classification for archived Websites only for the selective crawls. As the amount of collected data in the web archive is continuously growing, manual descriptions must be kept to an absolute minimum for practical and economic reasons. Automatic creation of metadata will be our main focus.

In general, the use of metadata standards like Dublin Core and general guidelines for website development within the cultural heritage field will have to be promoted. The use of general guidelines and standards will make the archiving task easier⁵.

Legal aspects

As for the legal aspects, we have to deal with two difficulties: the lack of a legal deposit act in the Netherlands, and the existing copyright legislation.

The current copyright law⁶, legislation on databases and privacy regulations pose problems for archiving the Web⁷. The question is whether, according to copyright legislation, we are legally allowed to make archival copies of websites and provide access

to the archive. Current Dutch copyright law has an explicit provision for libraries and archives copying works for preservation purposes⁸. They are permitted to make a copy from an item of their own collection for preservation and replacement without explicit permission of the rights holder. As this exception is only applicable to their own collections, it will not be of much use in archiving the Web, because Websites are not part of the permanent collection of the KB. Another exception in Dutch legislation is the possibility to provide access for scientific use within the perimeters of the Library without having permission of the rights holder⁹. As one of the objectives of a web archive is to gather resources for future research, this is feasible. According to legislation, a restricted network and restricted access cannot be regarded as the perimeters of the library. As digitization and networks allow users to have access to resources independent of time and place, copyright restrictions, by contrast, will complicate future access for researchers.

National legislation on the deposit of publications may provide a solution. This would provide us with explicit permission to make copies of websites with the aim of preservation and archiving. However, the Netherlands does not have a deposit act in which a free copy of every printed work can be claimed. Deposit of publications is based on voluntary agreements with publishers. This happens to be a good solution as the KB does not have problems in collecting publications. However, the vast amount of data we will collect from the Web and the fact that in theory everyone can be a publisher shall be an administrative burden. Obtaining permissions from the right holders to archive websites and provide access to the archive will be difficult and time consuming.

A survey on the possibilities within copyright legislation for archiving the Dutch Web is part of the KB project. Especially the exceptions within the law can, to a certain extent, prove to be helpful. We will have to focus on a licensing approach and develop a model for deposit of websites. A strategy for dealing with Intellectual Property Rights in web archiving will be developed, based on this research.

Ingest for long-term preservation in the e-Depot

As mentioned above, the web archive will use the existing e-Depot system and infrastructure. The main challenge is to define the best approach for ingesting a completely different type of material into the e-Depot.

As the e-Depot system has been developed and optimized for long-term storage of publications, access is not the main focus of the system. Storage and preservation planning are the basic principles of the e-Depot as one of the few safe places worldwide. Nevertheless, the web archive aims to enable research on the harvested websites, and prefers to provide access to this data within a reasonable time. As a consequence, we will store two copies of the harvested websites: a preservation copy and an access copy. This will give us the “best of both worlds”: the data is stored in a system dedicated to long-term preservation and we are able to provide access and to create an interface with search facilities based on the needs of the target audience of the archive.

Based upon the current e-Depot process, a workflow for processing websites, from selection to access, has been outlined, which is fully OAIS compliant. The basic assumption is not to change the existing process and infrastructure, but to add functionality that is specific for the web archive. As mentioned above, the type of material to be ingested for long-term storage is what makes the difference. The Web

ARCiving file format (ARC) is the format which is currently used as the most appropriate format for storing harvested data out of websites. ARC may be considered as a standard within the field of web archiving. As the ARC format is an open document standard and will become an ISO-standard, it will become a standard for collecting, storage and retrieval of websites¹⁰.

As the main focus of the KB is long-term storage and permanent access of websites, we need to examine how suitable the ARC format is for long-term preservation and storage in the e-Depot.

The ARC format consists of a sequence of document records. Each record starts with a header containing basic metadata, followed by the actual file that was gathered by the crawler. One of the main advantages of the format is its efficiency. It collects and stores a large amount of single files from a website in one container, so we need not to store many single files in a file-system¹¹. One of the disadvantages of ARC is the limited set of metadata that can be stored in the document headers. Because of this, it is not possible to store administrative and preservation metadata in the file. The upcoming extension of the ARC format, WARC, will offer more possibilities for incorporating additional metadata, which solves this first disadvantage. Another disadvantage is the compression of the single files that are stored in the ARC container. Although this will reduce necessary volume and costs of storage, the basic policy of long-term preservation systems like the e-Depot is not to compress the original data. However, it is possible to gather files in an ARC file without compressing them, which solves this problem.

The e-Depot system is set-up primarily to store the output of large publishers and contains almost exclusively single files or a collection of files of the same type. By taking up web archiving, we will now have to deal with a content type that contains a large collection of interconnected files. Without a doubt this will affect the workflow and requirements of the e-Depot system. It depends on the results of the project whether the W/ARC format will be used to store websites in the e-Depot, or whether we will look for another way of packaging for the long-term preservation of Dutch websites.

Preservation planning

Storing digital objects in the e-Depot is only a first step towards digital preservation. Apart from being stored in a safe storage system, the material needs to remain accessible in the long term as well. The usability of digital objects is threatened by the rapid innovations in computer technology. New systems and software supersede each other faster every year, making existing technology obsolete. This is becoming a problem in everyday life, but is especially visible within cultural heritage and research because there the main focus is long-term preservation.

Long-term preservation of websites presents us with some specific challenges. The vast amount of data is one of these as we have to store and render a continuously growing archive containing many versions of the same websites. Even though the costs of storage are decreasing and storage techniques are developing quickly, the performance of access to such a large amount of data will remain a challenge.

Another important issue is the large number of different file formats that are used in websites. Large domain crawls identify huge lists of MIME types¹². But if we examine

these MIME types in closer detail, it appears that only 20 types make up for 98% of all documents. Of these 20 types, html, jpeg, gif and pdf are the preferred formats, taking into account that the jpeg file covers most of the volume of collected data. For long-term preservation planning it is crucial to preserve information about the formats that are stored. If we incorporate the single files in W/ARC containers that act as a collection of interconnected files we need to concentrate on both the incorporated files as on the W/ARC format. Given the predominance of a selective number of formats, we can keep the majority of the Web Archive readable when we focus on these file types. They will provide us with a major challenge anyway. In addition to focusing on the preferred formats, we have to pay attention to the appearance of new formats on the Web. Permanent technology watch is needed to take the right measures for newly developed file formats. International cooperation on long-term preservation, research on preservation strategies, and the development of tools for validation and identification are necessary.

To support the long-term storage provided by the e-Depot, the KB carries out research on preservation strategies. Permanent access strategies can roughly be divided into two groups: migration and emulation. Migration is focused at the digital object itself, and aims to change the object in such a way that software and hardware developments will not affect its availability. By changing or updating the format of an object, it is possible to render these objects on current systems. Emulation does not focus on the digital object, but on the hard- and software environment in which the object is rendered. It aims to (re)create an environment in which the digital object can be rendered in its authentic form. The choice for either strategy is determined by the demands of future users and the types of digital objects that need to be preserved. Currently, the KB is developing migration and emulation, as well as combinations of these methods, such as the Universal Virtual Computer for images, a project carried out in 2004¹³.

As for emulation, the KB joined forces with the Dutch National Archive. Both organisations are convinced that emulation is a viable strategy and are building a modular emulator based on an existing standard configuration. This will be the first step; the next phase will be to make a universal emulator, suitable for a virtual PC and different object types¹⁴. As for migration, the KB started a research project to investigate the consequences of migration for different types of digital objects and will evaluate various migration tools¹⁵.

Besides research on these main preservation strategies, the KB carries out research into the use of preservation metadata and the development of tools for preservation planning. The KB also participates in a large scale preservation project funded by the European Commission, PLANETS. This project will deliver tools and services to facilitate preservation planning. It is carried out by a group of national libraries and archives, technical universities and software developers.

To be able to plan preservation strategies for websites, we have to define the essential characteristics of a website. This will help us decide what to preserve. As it is impossible to define all the needs of future users, we have to make decisions based on user scenarios. When future users want to browse through an old website, the representation of the site shall play an important role, depending on the objectives or research questions of the user. If they are only interested in the content, migration of the object will suffice. However, if they wish to experience the website in its original look-and-feel, the representation needs to be reconstructed as accurately as possible. As websites are

complex objects and we do not know which aspects will be considered important by future users, we most likely will need a combination of strategies.

Future developments

Other national libraries and archives follow the strategy of a combined approach: selective harvesting, added with event-based crawls and a periodical overview of the whole national domain. The KB has chosen to adopt this strategy as well to be incorporated in the day-to-day acquisition and long-term preservation workflow. In addition to this we may carry out one event-crawl and a broad scale crawl of the Dutch domain in 2007.

As information is created more and more exclusively on the Web, websites are becoming crucial in all types of research. Web data are clearly important to understand the development of these fields. Data collected on a regular basis by thematic and selective harvesting are more useful for over-time analysis than large scale domain crawls in which only a snapshot of the domain on a certain date is captured¹⁶. However, a broad scale crawl can be very useful for providing the context of the Web sphere, for research after genres of sites and for providing a portrait of an era. The web archive we create must be able to serve all of those questions of future researchers. Presently there is no clear vision on how to make a web archive available for scholarly research. Libraries and archives are only beginning to develop methods and techniques to conceptualise web archives in such a way that they can produce datasets for social science and humanities research. Therefore the KB intends to search for cooperation with designated communities to learn more about their needs.

In the near future the KB plans to set up a national collaboration structure for archiving the Dutch Web. Heritage and research organisations in the Netherlands have a common interest in web archiving. There is a common need to exchange knowledge and experience and probably a common need to exchange and share software. The manner in which this collaboration will take shape needs to be discussed further, perhaps with the model chosen by the UK Web Archiving Consortium as an example.

As for the KB, long-term storage in the e-Depot and permanent access of a selection of Dutch websites will be the main principles for its web archiving activities.

References

-
- ¹ Website are explicitly mentioned as cultural heritage in the UNESCO Charter on the Preservation of the Digital Heritage (17 October 2003)
- ² Pandora, Australia's web archive: <http://pandora.nla.gov.au/index.html>
Kulturarw, National Library of Sweden: <http://www.kb.se/kw3/ENG/Default.aspx>
Internet Archive: <http://www.archive.org>
- ³ OCLC Web Characterization Project: <http://www.oclc.org/research/projects/archive/wcp/>.
Netcraft Web Server Survey: <http://news.netcraft.com/>
- ⁴ We are planning to test Nutch/WAX for indexing and WERA for access.
- ⁵ Overheid.nl developed Web Guidelines for Dutch Governmental organizations:
<http://webrichtlijnen.overheid.nl/>
- ⁶ Dutch Copyright law dates from 1912, but has been adapted to the European Copyright Directive in September 2004.
- ⁷ According to the Database Directive, websites are to be considered to be databases.
<http://www.ivir.nl/wetten/nl/databankenwet.pdf>
- ⁸ As stated in Article 16n of the Dutch Copyright Law, consolidated version of 01-09-2004.
- ⁹ As stated in Article 15h of the Dutch Copyright Law, consolidated version of 01-09-2004.
- ¹⁰ The W/ARC format will be the extended version of the ARC format, being standardized by the International Internet Preservation Consortium ([IIPC]). Heritrix is expected to implement the revised format.
- ¹¹ Considering the best approach for storage of a web archive, the ARC format has demonstrated its strength as it meets most requirements. See: S.S. Christensen, *Archival Data Format Requirements*. The Royal Library, Copenhagen & The State and University Library Arhus, Denmark (July 2004)
- ¹² In the broad crawl of the Danish domain carried out in July 2005 over 600 different MIME-types were collected. In the crawl of the Australian domain conducted in 2005 over 900 different MIME-types were collected.
- ¹³ More information on the UVC research project:
http://www.kb.nl/hrd/dd/dd_onderzoek/uvc_voor_images.html
- ¹⁴ Hoeven, J.R. van der en Wijngaarden, H.N. van, *Modular emulation as a long-term preservation strategy for digital objects*, International Web Archiving Workshop 2005 (IWAW'05), Vienna, Austria, 2005; Hoeven, J.R. van der en Wijngaarden, H.N. van, *Modular emulation as a viable preservation strategy*, 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005), Vienna, Austria, 2005. In: *Proceedings Research and Advanced Technology for Digital Libraries: 9th European Conference*, ECDL 2005.
- ¹⁵ More information on the Migration research in the KB:
http://www.kb.nl/hrd/dd/dd_projecten/projecten_migratie.html
- ¹⁶ S.M. Schneider, K. Foot, M. Kimpton, G. Jones, *Building Thematic Collections: Challenges and Experiences from the September 11 Web Archive and the Election 2002 Web Archive*.