This article was published in
Alexandria: The Journal of National and International Library and Information Issues, published 29-08-2017, DOI https://doi.org/10.1177/0955749017725930

How can we improve our web collection? An evaluation of webarchiving at the KB National Library of the Netherlands (2007-2017)

Barbara Sierman and Kees Teszelszky

Abstract

The Koninklijke Bibliotheek, the Dutch National Library (KB-NL), started in 2007 the project "web archiving" based on a selection of Dutch websites. The initial selection of 1,000 websites has currently grown into over 12,000 selected web sites, crawled on different intervals. Although due to legal restrictions the current use is limited to the KB-NL reading room, it is important that the KB-NL includes the requirements of the (future) users in her approach of creating a web collection. With respect to the long term preservation of the collection, we also need to incorporate the requirements for long term archiving in our approach, as described in the Open Archival Information Model (OAIS)[1].

This article describes the results of a research project on web archiving and the web collection of archived sites in the KB-NL, investigating the following questions. What is web archiving in the Netherlands? What are the selection criteria of KB-NL and how are these related to what can be found on the Dutch web by the contemporary user? What is the influence of the choice of tools we use to harvest on the final archived website? Do we know enough of the value of the web collection and the potential usage of it by researchers and how can we improve this value? This article will describe the outcomes of the research, the conclusions and advice that can be drawn from it and will hopefully inspire broader discussions about the essence of creating web collections for long term preservation as part of cultural heritage.

Keywords

web archiving; long term preservation; web history; library science; OAIS

0. Introduction

The importance of websites and web archives as cultural heritage has been acknowledged worldwide in the UNESCO Charter on the Preservation of the Digital Heritage, published in 2003.[2] KB-NL started with web archiving a selection of the Dutch web as a research project ten years ago in March 2007.[3] Over the years,

web archiving activity evolved from a pilot project initiated by the Research Department into a regular activity of the Collections Department. The KB-NL collection of archived websites contains now more than 12,000 websites preserved in almost 26TB data and comprises around 211 million URL's.[4] As there is no legal deposit in the Netherlands and due to copyright reasons, the collection can only be studied in the reading rooms of the library and attracts around hundred users every year.

I. Prepare for Preservation

OAIS

The aim of the KB-NL web collection is to select, to preserve and to make accessible a representative set of Dutch websites. The KB-NLs policy is to be conform with the Open Archival Information System standard (OAIS) which describes on a conceptual level the approach for long term preservation, both in an functional model as well as in an information model. Currently the KB-NL web collection is stored on a dedicated server at an external location as .ARC files, and not yet in a preservation system. This means that the web collection currently is not compliant with the OAIS standard (as are many web collections). But the KB-NL need to prepare herself for the ingest of the web collection into the preservation system in a few years, whereby the OAIS requirements will be taken into account. So the various concepts of OAIS need to be taken into account now, although the actual ingest might be at a later stage.

The information model of OAIS distinguishes several topics in the Information Model, related to what the so called "Designated Community" need to know in order to be able to make "Information" from the Content Data Object (the preserved website) and the related Representation Information, together the Content Information. As the custodian, the KB-NL is to a certain degree responsible for delivering this information. The Content Data Object has clear boundaries, as this is the "original target of preservation". The Representation Information not only encompasses the information needed to render in this case the instance of the web harvest faithfully, but also need to contain relevant information in relation to the so called Knowledge Base of the Designated Community.

Designated Community

In OAIS terminology the Designated Community is the target community of your archive: the future users of your archive. The primary Designated Community of the KB-NL web collection is the academic and non-academic research community. Researchers that will base their research on the KB-NL web collection need to know their topic of study, so they might be expected to have at least a basic knowledge of the original historic context of the collection, the process of web harvesting and how the web was organised during the period of their investigation. But they will also be interested in several aspects of the collection, of which only the KB-NL can provide information.

In order to better understand the KB-NL web collection, context information about the broader environment of which the web collection was part of is important. The KB-NL is doing selective harvesting of websites of the Dutch national domain, but which criteria were applied during the selection process? And did these selection criteria change over the years? Did the KB-NL harvest all the sources within the scope of the crawl or were there technical or legal barriers that prevented achieving these goals? How much was selected in comparison with what was available at a certain time? Did ethical questions play a role during the process of selection or harvest? In short: which aspects that will be relevant for researchers influenced the KB-NL collection during the process of selection and harvest? Where can a researcher find this so-called context information if not available in the web collection?

Preservation Description Information

Apart from this context information, the OAIS model also requires extra information in the Information Model for each web site that is preserved, especially in the Preservation Description Information (PDI), as described in paragraph 4.2.1.4.2 of the OAIS reference model and figure 4.16. It will contain Reference Information, Provenance Information, Context Information, Fixity Information and Access Rights Information. Some of this information that can be captured in metadata.

In a survey done in 2013 by the IIPC Preservation Working Group amongst the IIPC members a question was asked whether the web collection was already integrated in their preservation system, assuming that for this preservation system the requirements of the OAIS model were realized.[5] Out of 25 completely received surveys, 37% answered that they had it integrated, while 63% answered that they were either planning it or had not done so. We can conclude that the majority of the respondents did not archive their web collection conform the OAIS Reference Model into a preservation system. One of the exceptions is the Bibliotheque National de France, who gave a description of their Premis/METS model in which part of the PDI information is incorporated.[6] In short: context information is crucial to understand a web collection, both for current and for future users.

Although the KB-NL web collection is not yet ingested into a preservation system that is organized in compliance with the OAIS Data Model (including the above mentioned Representation Information and Preservation Description Information), we started this research to investigate what we could improve in the KB-NL to prepare ourselves for an OAIS compliant web collection, with a focus on the three main steps of the workflow: selection, harvesting and presentation.

II. Selection

The Netherlands has no legal deposit law prescribing what the National Library should collect. The collections of the KB-NL are based on the collection plans, including the web collection. This should be a representative set of Dutch web publications. The lack of legal deposit also requires that the KB-NL must ask permission of each web site owner before the web site can be part of the collection. Although the KB-NL has a set of selection criteria, this does not mean that the web archiving team can collect as it pleases. In practice this approach resulted also in some exemplary websites that the KB-NL did not get permission to harvest. It also hampers rescue archiving: some sites went offline while KB-NL awaited permission to harvest. What we could not harvest or were not allowed to is also context information which lacks now.

When the web archiving activities started in 2007, the KB-NL based its seed list on a selection of Dutch websites, with the intention to create a representative set. Looking back we need to conclude that more effort into determining what encompassed the Dutch web in 2007 might have helped to make a more representative selection from the beginning. The knowledge of the web had to be build.

The OAIS model also describes that an OAIS archive should be clear in what is collected and what is not in the collection and to be transparent in this aspect to the Designated Community. In this paragraph we will briefly describe the Dutch national web, the background information that could be helpful in determining the size of a national web and extra sources that could be explored in shaping the overview of the Dutch web.

The Dutch National Web

If we consider the national Dutch web as the way the Dutch shaped their virtual space on the web, this development started early. The .nl domain of The Netherlands was the world's second country code Top Level Domain (ccTLD) outside of the USA. The first .nl domain name cwi.nl was registered already in 1986.[7] The first Dutch website was published online six years later in 1992 as the third website in the world, after that of CERN in Switzerland and SLAC in the USA.[8] The .nl ccTLD comprises of 5,76 million domain names in June 2017 according to the figures of the Dutch national domain registrar SIDN.[9] Given its size and the amount of websites per inhabitant, the Dutch national domain is one of the biggest national domains to be crawled in the world and has a daily growth of 5,500 domain names. If we take the non .nl domain names into account which are used to publish Dutch sites, it is even bigger. To preserve even a representative part of this national web is an enormous task for a national library with limited resources.

Mapping the Dutch web and the start of web archiving

Even before web archiving started, the KB-NL National Library was aware of the potential information value of the Dutch part of the World Wide Web for its patrons. It swiftly followed digital developments: it started mapping the Dutch web already in 1992 by compiling web directories or web lists of relevant websites for their users. This "NL-menu" was a classified overview of information services of Dutch organisations on the web.[10] This activity was done as a part of the traditional library task regarding the collection of Dutch publications, be analogue or digital. Another KB-NL project, Duchess (Dutch Electronic Subject Service) started in 1997 and was a service to disclose internet sources for researchers and add to them the Dutch Basic Classification. Only the URL's were collected and described: the websites were not harvested yet. These are only two examples of various initiatives that the KB-NL started to provide their audience information about new information sources. For years, the web directory of the Dutch national library was one of the most important providers of information on the geography and content of the Dutch web. This activity has only been finished in 2004 as it seemed no longer useful, due to the popularity of commercial web directory sites and the growth of search engines as Google.

When web archiving started in 2007, a selection of 400 websites taken from the last Dutch web directory DUCHESS was used as a seed list for the first experiments with selective crawls. Yet this context information of the start of KB-NL web archiving seemed to be lost and forgotten after ten years. These KB-NL initiatives from the 90ties were part of the collective memory in the library, but their traces were hidden and detailed information was not easy to find. NL menu was handed over to a Public Library organisation and DUTCHESS was stopped. But, the limited information that still existed in the KB-NL (sometimes on an old CD in a drawer of an employee which we were able to rescue) gave us important context information about the historic development of the Dutch web and what was seen as important web sites from the KB-NL perspective in those days, although it required some "web archaeology activities" to find and rescue these sources. The future publication of these web lists will place the current web collection into context, which is one of the essentials for long term preservation. The collected information by the KB-NL was seen as only of actual use and was not preserved yet. But in the end, this kind of information, which is already available in the institutional memory, can be important background information and an useful source of information for researchers.

Selection policy of KB-NL

It is important to stress that although the KB-NL is the national library of The Netherlands, the collection of archived websites was neither created as a national web archive, nor it aimed to collect all material from the Dutch national web. The selection criteria of the KB-NL are defined by technical, financial, legal and human possibilities, web space, content, theme and period of time. As the web collection is the end result of

this range of criteria, this is important context information for a researcher, and need to be added to the preserved web collection in some way. We have not yet discussed how it would fit in the OAIS data model. The general criteria for selecting material were written down in the KB-NL collection plan.[11] The collection plan states therefore that the KB-NL collection covers 'everything published in and about the Netherlands'.[12] This policy was refined and limited at the start of web archiving in 2007 to websites about Dutch language, history and culture. Moreover, the KB-NL limited itself to the selection of websites from the Dutch national domain: other Dutch material from the internet, as emails, programs or apps are not preserved. Furthermore, a website must be a separate publication of a certain size: tweets or other microblogs and social media items were excluded from selection.

As web archiving became a more regular activity, the KB-NL selection criteria were adapted to technical or other limits and web trends. Due to copyright reasons, the lack of a deposit law and resources, the original ambitions became more modest and were adapted to the goal of archiving 'a representative selection of the Dutch national domain'. The legal limits of Dutch web archiving were described in several publications which were written in collaboration with legal experts of the University of Leiden.[13]

As the web archiving team began to know the Dutch national domain better, the selection policy was also extended and further refined. The original restriction to sites of the Dutch ccTLD .nl domain turned out to be too narrow for making a representative selection, as many sites with valuable content did have other extensions. Therefore the selection policy was extended to sites with other extensions as well. Another criterion did exclude commercial sites at first. As many old and established Dutch companies went bankrupt last years and their websites went offline, the KB-NL selection policy became more flexible to include endangered sites which do not meet the selection criteria, but are important to preserve from a Dutch digital web heritage point of view. Finally, websites were selected based on popularity on the Dutch web, using various sources as Alexa, Wikipedia and Similarweb.

As the selection policy of the KB-NL did evolve over time, these changes were neither always recorded, nor communicated with present and future users through the website. One could say that the changing context of the web collection was not recorded. The future goal is to better inform users, our Designated Community, what is preserved, included and excluded from selection. Therefore we need also to provide information on the development of the Dutch web and the content of past web directories: even if most of the sites present are neither online anymore, nor preserved. We even plan to make visible what was not archived due to legal, technical or other issues to publish these URL's. It is important for our Designated Community to realise the library did its best to preserve as much digital heritage as possible and to make them aware of the limits of web archiving by providing context information of what was not preserved.

Future developments in the selection policy

There are still three issues related to the selection policy which will be solved in future. First, the selection policy focuses since 2007 on actual websites and does include only a small amount of websites published before 2007. We plan now to select relevant Dutch web heritage which is still online by conducting web archaeology and to harvest sites which are important to study the origin of the Dutch web. Already one ancient web directory, the tenth website ever published in the Netherlands, was rescued from a server on an attic, reconnected to the web and finally harvested.[14]Also, the web archiving team focuses now on the selection of more online news sites, as newspapers become less important and online news more influential. Third, the selection policy excluded websites which despicable or untrue content. As abject content on the web becomes more influential in society and fake news is thriving, it is necessary to preserve this source material for future research.

The selection policy of KB-NL from a national perspective

Apart from the KB-NL, several other Dutch organizations are creating web collections based on a specific portion the Dutch web or a specific Dutch theme. All of them have a different selection policy, crawl strategy and sometimes even use another web harvesting technology as well. The KB-NL selection policy takes the selection policies and the harvest activities of other institutions into account, even if a different crawl technology is used. What other organisations crawl, the KB-NL does exclude from its policy in principle. This principle is also contextual information about our web collection and thus of importance for our Designated Community.

As far as we know, web archiving started in The Netherlands by the Dutch Documentation Centre of Political Parties (DNPP) in 2000, using HTTrack web crawler tool.[15] This organisation harvests almost all the websites of Dutch political parties, politicians and political movements on a monthly or yearly base. Many local or municipal archives run also web archiving projects, like the Frisian Treasor collection (the repository of the history of Fryslân), the county of Groningen and the cities of Rotterdam and Dordrecht.[16] Besides the Dutch National Archive harvests websites from an archival point of view. Finally, the Netherlands Institute for Sound and Vision collects the websites of the Dutch broadcast organisations. All the organisations together collect around 15,000 websites, but the KB-NL collection is the largest of all. Future researchers must take this national context information into account when studying the Dutch national web and the KB-NL web collection. They will be able to do so, as most of the above mentioned organisations have a preservation task and will preserve their web collections according to the OAIS model.

If we consider the selection criteria of the different web initiatives in the Netherlands including KB-NL, we can observe a bias towards politics, local sites, media and cultural history and heritage and a lack of archived sites before 2007. Due to this, a national expert group was launched at the end of 2016 that focuses on web archiving on a national scale.[17] Its purpose is to promote cooperation between all the different institutions and the professionalization of web archiving in the Netherlands. Another goal is to make an inventory of all the different web archiving initiatives in The Netherlands and to make a list of all the websites which are harvested by various organisations. In this way, a researcher of the Dutch web can get a better insight of what is archived where and what technique is used. The combined efforts of all small organisations together will enrich the value of the separate web collections and provide context information about the national Dutch domain for future users.

III. Harvest

Web archive or web collection?

A harvest policy is a key issues to the development of a collection of archived websites. As web archiving is a relatively new activity for libraries, the definition of the goal driving the collection development is still under discussion. The collection's original target in 2007 was defined as "to harvest a selection of the Dutch web with a maximum of 3,000 websites". Therefore, the collection of archived websites at the KB-NL National Library can be regarded as a special web collection of a scientific library, rather than a general or even a national web archive. According to Helen Hockx-Yu, a collection of archived websites was described by Brewster Kahle as a web archive when he founded the Internet Archive in 1996. In his opinion, a digital collection of websites must be considered as a web archive and not as a web library, as its collection can never be complete.[18] Still, a web archive is not an archive, in the sense of a place in which public records or historical documents are preserved. The archived websites which are collected and preserved by the Dutch National Archives can be considered as a true web archive from this point of view.[19] The KB-NL owns a collection of more or less similar archived websites which have been selected for a reason with a specific goal in mind. The term "web collection" is therefore more suitable in the Dutch KB-NL case.

Harvest strategy of the KB-NL and web sources

What is harvested by the KB-NL for its web collection and how is this actually done? The mission of the KB-NL National Library is clear: to collect and preserve everything published in and about the Netherlands and the Dutch culture in order that researchers, students and other users will be able to consult this now and in the future. A problem arises when trying to apply this policy to web material, as it is not clear how to define the scope, size, content and even value of the digital object which we want to collect and to preserve.

The Danish web historian Niels Brügger has described five analytical layers of the web to identify digital web objects.[20] These objects can be identified, harvested and archived on the following layers:

1. The individual textual elements of a web page: source code, text, images, style sheets, etc.;
2. The individual web page: the layer where all above described elements can be found under a certain URL and which are linked to it;
3. The individual website: the level where all linked web material which can be found under a certain domain name;
4. The web sphere: the layer where all sites which are linked together with one certain website;
5. The web as a whole: the level where all websites are online at a certain moment.

As the KB-NL harvests from a web collection point of view, the focus of its harvest strategy is on the third analytical layer of the web. This means that the KB-NL preserves the individual website, which is considered as a separate digital object to be collected as a single unite in web space and time. The KB-NL web collection as a whole is therefore described in an amount of selected websites with a separate time stamp and presented as a list of URL's accompanied with the date of selection. The contextual information we want to offer our Designated Community will focus on 1,2 and 3 in the metadata in the Archival Information Package, while contextual information for 4 and 5 need to be described separately.

The harvest strategy of KB-NL is to make a snapshot of all the elements of one website at a certain moment or period of time. An online website is a dynamic object linked to the live web. The goal is to harvest the selected live website as complete as possible and to collect as much web material as can be done by the harvester from one URL within the shortest possible time. The purpose of this is that it can be studied by the user as an object in the KB-NL Wayback Machine like it was live at a certain moment.[21] During the harvest, the website is cut off from the live web, harvested on the level of the site and the individual web pages by following and harvesting links and web elements. Afterwards, the harvested web material is reconstructed as an archived version of the site in the web collection and made accessible through the Wayback Machine. The context information of the KB-NL web collection which is presented to the user is about the third analytical level of the web.

Due to the harvesting by the KB-NL from a collection perspective, we can state that its activity of web archiving is not the preservation of a historic source, but the creation of a new one on the third level of web analysis out of harvested elements. The result which can be viewed in the Wayback Machine must resemble the live version as much as possible to be considered as an authentic source of our digital age. The authenticity which is missing of the archived instance is the dynamics, as it is a snapshot made of a dynamic

website on a certain moment. Still, the harvest of the website also can have its own dynamic characteristics. The bigger and more dynamic the website is at the moment of harvest, the longer is the "shutter time" of the harvest, which is also due to the used web archiving technology. One link, web page or web element can be harvested at a different time than the other.

KB-NL and the Internet Archive

The general harvest strategy of the Internet Archive is focused on the first analytical layer of the web.[22] The IA describes its collection present in the Wayback Machine for the general public in the amount of time-stamped web objects or web captures, which means archived web elements.[23]

The harvest strategy of the KB-NL differs from that of the Internet Archive with regard to the analytical level of the web, as the KB-NL harvests from a collection point of view. The KB-NL focuses on preserving an authentic archived website. The IA focuses on broad harvesting at the first analytical level of selected individual web elements through domain harvests of web spheres, not on snapshots of selected websites.[24] When viewing a specific website in the Wayback Machine, the researcher navigates through snapshots of websites with elements which were harvested on different moments of time and were brought together later in the Wayback Machine.[25]

The difference between harvesting methods has important implications for research on websites. If we research web material at the analytic level of the websites, selective harvests offer us a more authentic source. But if we conduct research on websites on the level of the web sphere, selective harvests offer less authenticity, as the instances of separate websites were harvests within different time intervals and together cannot be treated as one source of a certain moment. It is therefore important that the user is aware of this context information when studying the archived web.

A domain crawl of a national domain or national web in addition to selective crawls can provide valuable context information about the individual archived websites in the web collection for future users. Due to legal issues, the conducting of a domain crawl is not yet possible for the KB-NL. If the KB-NL had been able to conduct domain harvests of the Dutch national web, like the British Library and the National Danish Library are able to do, it could have provided an authentic snapshot of the Dutch web sphere as well for future researchers.[26]

Heritrix web crawler tool

As web harvesting technology results in the creation of new sources in the web collection of archive, it is important to understand the working of the web harvester. The KB-NL uses Heritrix version 1.14.1 for web

archiving, as most of the national libraries and large heritage institutions in the world like the British Library, the National Library of New Zealand, the Biblioteca Nazionale Centrale Firenze, Netarkivet in Denmark and the Bibliothèque Nationale de France do. The tool Heritrix is responsible for the majority of archived websites and the content of web collections in the world. Therefore background information on the working of this program will be crucial for future users of archived web material to judge the value and authenticity of the sources and to understand the context in which this material was created.

The core setting of Heritrix is to focus on or the first, or the third analytical level of the web. The main difference between the crawl strategies of IA and KB-NL as described above has its root in a different settings of the web crawler Heritrix. The IA conducts broad domain harvests, which means that the crawler harvests as much web material as possible from the first layer of the web, but only one of two levels of a website and therefore only scratches on the surface of the third analytical level of the web. The KB-NL does selective harvests: focused crawls of selected websites. Heritrix is therefore instructed to keep in scope of the selected website and crawls as much web material as possible from a certain website. If we compare IA and KB-NL on the fourth level of the web sphere, it means that the IA harvests more from this, but superficial, and KB-NL harvests less, but very thorough.

It is therefore not possible to state that the IA harvests "everything" and KB-NL possesses only a small probe of the Dutch national web which is already present in the collection of the Internet Archive. What collection is most useful for research of the past web depends on the need of a researcher what analytical level of the archived web he wants to study and how authentic the archived resource must be for his research goal.

We can state that it is necessary for web archivists and researchers to understand the working of Heritrix and its and outcome. At the moment, there is a lack of information on the web about Heritrix ,and the difference between the versions. Unfortunately, even developer documentation for Heritrix is largely out of date and scattered around the web.[27] No person or organisation in the world takes the responsibility to keep this information up to date or checks the content of it. There exists a developer community of Heritrix, but its relatively small and no organisation takes the lead of the further development.[28] This poses serious limits on the availability of context information about the harvest.

As we have stated above, many institutions do still use an old version of Heritrix for different reasons, including KB-NL. Still, this version has serious flaws, of which most researchers are not aware of. Sites with https cannot be harvested anymore for example. Another serious issue is the crawler trap, through which tons of unwanted data is harvested which is useless for analysis. At last, dynamic websites are hard to crawl.

There are new tools like Brozzler and Webrecorder under development to deal with these issues, but it takes a serious investment in time and money to implement these in the regular work flow of institutions.[29]

Our Designated Community of researchers need to become familiar with the details of harvest techniques when doing research on web collections. Now, only a few researchers understand the technical aspects of web archiving and most of them are web archivists themselves. Researchers are scarce now who take the working of a web crawler into account when analysing archived web resources. It is not enough anymore for a researcher to understand the first layer of the archived web and be able to analyse texts and images: to be a serious researcher of the digital age, knowledge of all web layers must be a prerequisite for doing research of web collections.

Source criticism of archived websites is thus scarcely out of the egg, but this knowledge is very necessary to make web collections useful for researchers and to increase the overall value of the web collections. When researchers understand the working of Heritrix and its outcome better, they are also able to judge the value of the web collection better. It is therefore important not only to collect and preserve context information of the selection policy and the collection, but also preserve the data and other background information on the harvesting tools which are used to build the web collections.

Conclusion

The KB-NL has archived websites for more than ten years and has built up an unique web collection of the Dutch digital web culture since 2007. Still, this digital collection is not ready for long term preservation yet. If we want to preserve this collection for the future in a responsible way, we need to incorporate the requirements for long term archiving as described in the Open Archival Information Model. The most important requirement is to provide context information about the Dutch national web domain, the national web collection of Dutch web archives, the KB-NL selection policy and the harvest strategy and technology. This can be done by mapping the past and present Dutch national domain by using old and forgotten data, to draw up a national list of all archived websites in the Netherlands by Dutch web archiving institutions, including those of the KB-NL. Also the selection policy and policy changes must be recorded and this information made available for future researchers. Besides, understanding the harvesting techniques and the outcome of these is crucial to value the authenticity of a preserved digital source of the past web. Web archiving institutions should make researchers more aware of the possible limits of their objects and the difference between the various collections due to different harvesting strategies and used tools. Finally: the web does not have national borders, neither does a national web collection have. In order to be fully prepared for the future, national libraries must secure national context information by international cooperation with other institutions.

Barbara Sierman MA is digital preservation manager at the Research Department of the Koninklijke Bibliotheek, National Library of the Netherlands. She participated in several European projects like Planets, SCAPE and APARSEN and participated in the development of ISO standards related to Audit and Certification of Trustworthy Repositories. She is a member of the IIPC Steering Committee and Chair of the Board of the Open Preservation Foundation. Blogs on https://digitalpreservation.nl

Kees Teszelszky (Voorburg, 1972) graduated in Leiden (Political Science, 1999) and Amsterdam (East European Studies, 1998), and obtained his PhD in Groningen (Cultural History, 2006). He does research on the web archive of KB National Library of the Netherlands

---

[1] CCSDS (2012) Reference Model for an Open Archival Information System. https://public.ccsds.org/pubs/650x0m2.pdf

[2] UNESCO (2003) Charter on the Preservation of Digital Heritage, 15 October. http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html (accessed 5 April 2017); UNESCO / PERSIST Content Task Force (2016) The UNESCO/PERSIST Guidelines for the selection of digital heritage for long-term preservation, March. https://www.unesco.nl/sites/default/files/uploads/Comm_Info/persistcontentguidelinesfinal1march2016.pdf (accessed 5 April 2017)

[3] Available at: https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving (accessed 5 April 2017)

[4] Available at: https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving (accessed 5 April 2017)

[5] Pearson D, Oury C, Goethals A, Sierman B, Steinke T (2015) Facing the Challenge of Web Archives Preservation Collaboratively: The Role and Work of the IIPC Preservation Working Group. *D-Lib Magazine* 21(5/6). Available at: http://www.dlib.org/dlib/may15/goethals/05goethals.html (accessed 5 April 2017)

[6] Dappert A, Guenther R and Peyrard S (2010) *Digital Preservation Metadata for Practitioners: Implementing PREMIS.* Chem: Springer: 59-82.

[7] Available at: https://www.sidn.nl/a/over-sidn/ons-verhaal (accessed 5 April 2017)

[8] Available at: https://www.nikhef.nl//nikhef-en-het-www/ (accessed 5 April 2017)

[9] Available at: https://www.sidn.nl/ (accessed 5 April 2017)

[10] Available at: http://www.nl-menu.nl/over (accessed 5 April 2017)

[11] KB (2014) Strategic plan 2015-2018 - The power of our network. Available at: https://www.kb.nl/sites/default/files/docs/strategicplan-2015-2018.pdf (accessed 5 April 2017)

[12] KB (2012) Collection development programme 2010-2013. Available at: https://www.kb.nl/en/organisation/organization-and-policy/collection-development-programme-2010-2013 (accessed 5 April 2017)

[13] Beunen A and Schiphof T (2006) Legal aspects of web archiving from a Dutch perspective. The Hague: KB; Beunen A (2008) Webarchivering en auteursrecht. In: Van Trier G (Ed.) *Handboek Informatiewetenschap*. Deventer: Kluwer. pp III 120-1-III 120-19; Beunen A, Ras M, Cameron E. and Schiphof T. (2007) Juridische haken en ogen van webarchivering. In: *InformatieProfessional* (10) pp 22-27. Available at: https://www.kb.nl/sites/default/files/docs/webarchivering_informatie_professional_2007_oktober.pdf (accessed 5 April 2017)

[14] Dutch Home Page. Available at: http://dhp.overmeer.net/ (accessed 5 April 2017)

[15] http://www.archipol.nl/ (accessed 5 April 2017)

[16] https://www.tresoar.nl/; www.regionaalarchiefdordrecht.nl/; https://www.groningerarchieven.nl/onderzoek/webarchief-groningen/; http://www.stadsarchief.rotterdam.nl/ (all accessed 5 April 2017)

[17] Available at: http://www.ncdd.nl/en/knowledge-and-advice/expert-groups/expert-group-web-archiving/ (accessed 5 April 2017)

[18] Available at: https://twitter.com/hhockx/status/799278725569257472 (accessed 5 April 2017)

[19] Available at: http://www.nationaalarchief.nl/kennisbank/onderwerp/web-archivering/ (accessed 5 April 2017)

[20] Brügger N (2010) Web History, an Emerging Field of Study. In: Brügger N (ed) Web history. New York: Peter Lang, p 3.

[21] Available at: http://webaccess.kb.nl/ (Accessible only in the KB-NL reading rooms, accessed 5 April 2017)

[22] Available at: https://archive.org/ (accessed 5 April 2017)

[23] Available at: https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/ (accessed 5 April 2017)

[24] Available at: https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/ (accessed 5 April 2017)

[25] Kalev Leetaru, How Much Of The Internet Does The Wayback Machine Really Archive? Available at: https://www.forbes.com/sites/kalevleetaru/2015/11/16/how-much-of-the-internet-does-the-wayback-machine-really-archive/#1020a34a9446 (accessed 5 April 2017)

[26] Available at: http://www.bl.uk/aboutus/stratpolprog/digi/webarch/; http://netarkivet.dk/in-english/faq/ (accessed 5 April 2017)

[27] IIPC Heritrix Task Force (2012) Report from the Heritrix Workshop held in London, September. Available at: https://sbforge.org/download/attachments/11338002/IIPC%20Heritrix%20Task%20Force%20-%20London%20Meeting%20report.docx?version=1&modificationDate=1352110933248&api=v2 (accessed 5 April 2017); See also: http://crawler.archive.org/An%20Introduction%20to%20Heritrix.pdf; http://crawler.archive.org/articles/user_manual/; http://crawler.archive.org/articles/developer_manual/index.html; (all accessed 5 April 2017)

[28] Available at: https://github.com/internetarchive/heritrix3/network (accessed 5 April 2017)

[29] Available at: https://github.com/internetarchive/brozzler; https://webrecorder.io/ (all accessed 5 April 2017)