These Crawls can talk

Context Information for Web Collections

|  |  |
| --- | --- |
| **Susanne van den Eijkel** | **Daniel Steinmeier** |
| *National Library of the Netherlands**Susanne.vandenEijkel@KB.nl**0000-0002-8221-869X* | *National Library of the Netherlands**Daniel.Steinmeier@KB.nl* |
|  |  |

**Abstract – The National Library of the Netherlands has been harvesting the web since 2007. As is well known, an archived website is fundamentally different from a website on the live web. As a digital repository we create archival objects in the process of archiving, especially in the case of web harvesting since a website does not have clear boundaries. This means we define the limits ourselves. So how do we provide researchers with information on the choices we made during the process of selection, harvesting and ingest? How can we prove the integrity and authenticity of our web collections to our designated community? Based on the findings of Maemura et al. in 2018, three categories of information should be available for researchers of web collections: scope elements, process elements and context elements. This will help researchers understand what is present in a collection, what curatorial decisions have been made in the process and the reason behind the creation of the collection. In this article we will describe three documentation initiatives related to our web collections that we think could be seen as implementations of these three types of documentation elements.**

**Keywords – Preservation, Metadata, Context Information, Web Collections**

**Conference Topics – Exchange; Community**

# Introduction

As National Library of the Netherlands (KBNL) we are determined to guarantee long-term accessibility to our cultural heritage. The first thing people think of when they hear ‘library’ is books and indeed we have a lot of those, but for fifteen years now it has also been part of our mission to preserve the web. KBNL intends to collect a copy of every publication that was made in the Netherlands or is about the Netherlands, as defined in our content strategy [1]. In 2019, a ‘digital first’-principle was added. This means that we only save a physical copy of the content if there is no digital one available. Websites are considered publications just as much as books or articles. Therefore, selecting and archiving websites has the same priority as archiving books, newspapers and journals from our regular publishers.

In the field of web archiving, the focus is often on tools for harvesting certain types of content, or innovative technical solutions for providing access to web collections. However, access also means keeping material understandable for our users [2]. An important target group within our user community consists of researchers. How do we provide for this group when we want to keep material understandable? A lot of research about the web is based on statistical information extracted from data present in web archives. However, for these statistics to be reliable researchers should be able to understand the choices made by archivists during the selection and creation of the web collection. This is information that cannot simply be extracted or generated by tools but should preferably take the form of documentation written by the people making the day-to-day business decisions necessary in the process of archiving the web.

In our preservation plan we outlined three key themes important for long-term preservation: integrity, authenticity and long-term accessibility [3]. For us, integrity revolves around all measures necessary for ensuring completeness of the objects and the collection itself. The concept of authenticity requires information about provenance, the producer and the intention of the archival object. Both integrity and authenticity are important for considering archival material long-term accessible. In digital archiving a fair amount of trust is placed in technical measures that are supposed to prove authenticity. However often non-technical factors like policy documents play an important role in proving authenticity in actual practice [4]. It is important to consider that many guidelines in digital preservation can be implemented using non-technical solutions. An example of this would be documentation meant to enhance understandability of the objects and the collection, described in the Open Archival Information System (OAIS) and the Trustworthy Digital Repository (ISO 16363) standards as representation information. After all, business decisions made by operators during archiving are best explained in written documentation and cannot be simply generated.

So, what would be important documentation to capture in the case of web collections? For this paper, we based our examples on the three element types described by Maemura et al. in the article ‘If these crawls could talk’, namely: scope elements, process elements and context elements [5]. All information on what is and is not included in the web collection is part of the scope elements. Scope is also part of the configuration of the web harvesting software we use. Harvesting is the process of collecting or crawling websites. Without scope settings the crawler would go on crawling the web indefinitely. The scope of the harvester defines boundaries, but this also has impact on the integrity of the resulting collection, so the choices made in the scope need to be explained and documented in detail so researchers will understand to what extent the collection can be considered ‘complete’. Not only scope is important in this regard however, also the process of archiving needs to be mapped. This entails the process of harvesting, and the process elements present in the configuration of the harvester, but also the curatorial decisions made after harvesting. For instance, what actions are taken within the scope of doing quality analysis or in case of errors that occur in the process. Finally, an important piece of documentation concerns the context of the collection. Why is a collection harvested and how are targets for harvesting selected? So yes, these crawls can talk, if we make this context information available for our users. Taken together these three categories of documentation will provide researchers with information on the social factors impacting the integrity and authenticity of archived websites in heritage collections.

# Process Elements: Quality Assurance on Selective Harvests

In 2006, KBNL started a small collection of websites based on manual selection. This part of the collection is what we call the selective harvest, since it is a curated collection of websites that is deemed important from a cultural perspective, as defined in our collection policy. Every website in this collection has been selected because it represents in some way a part of Dutch language, culture and history. This selective approach is in line with the remit of KBNL, the resources available for web harvesting and legal considerations [6]. KBNL uses the Web Curator Tool (WCT) [7] to collect the selective harvests. This is an open-source tool for managing selective web harvesting for non-technical users. It provides a graphical interface for changing settings and adding basic descriptive metadata, but the underlying crawler software is Heritrix [8].

The web collection of KBNL currently consists of 20,000 websites. Among this number are the special collections. Since 2013, our curators started assembling special collections for the web archive in a similar fashion as the UK web archive. A few examples of these special collections are websites about the commemoration of 200 years Kingdom of the Netherlands, the First World War and the Covid-19 pandemic. This collection continues to grow, as our curators are still actively searching for popular websites or websites that reflect topics currently relevant in Dutch society [9]. What distinguishes a selective harvest from other forms of harvesting, like a domain crawl, is that a selective harvest consists of a relatively small number of websites. This means we are better able to analyse the results of the harvest thoroughly to see whether content is missing. Based on our findings we can alter the settings, in order to harvest a complete collection as possible.

We have documented in an internal manual how quality assurance (QA) is done and how the results of this process are added to the metadata as annotations via the WCT interface. By adding the outcome of the QA process in these notes, it will stay available as context information for internal users and researchers. Every two weeks employees check the content that was harvested in that period. In order to execute QA as effectively as possible, the new harvests are divided into three groups: websites smaller than 1MB, websites bigger than 1GB or with a runtime longer than twenty-four hours and websites that have a divergent schedule or ran into a limit that we did not expect. Each target should have a schedule that indicates how often a website will be harvested, for example every half year. If a scheduled website contains less content than expected, the website will appear in the selection for QA. We try to find out if and why there is content missing and how to handle this. In the notes in WCT we document the findings and whether the problem was solved.

Our documentation also provides rules for URLs that must be included or excluded, defined as regular expressions. These rules are often linked to problems that were found during QA. Shopping carts, for example, are filtered out, because the crawler gets stuck on add-to-shopping-cart-buttons, resulting in unwanted content and a long runtime. Sometimes the homepages of websites refer to different URLs, even though they are part of the same website. Depending on how deep we wanted to harvest a specific website, we would add the other URL for the same website as include or as a secondary seed and documented it in the notes. In this way it is also clear to future users why earlier harvests were incomplete and differ from later harvests.

These annotations about quality assurance give users insight into the status of websites in our repository. Part of the harvest history as well as the decisions made during QA can be reconstructed with the help of these notes. For example, ‘QA 2019 PC OK’ means that during a QA-analysis in 2019 a website was picked out for error handling, that the problem has been fixed and that the website can be harvested again without errors. Often a specification follows the annotation in the notes, describing the problem and the solution briefly. It can also happen that a website no longer exists. In that case the annotations will mention ‘QA 2021 PC cancelled’, followed by an explanation in Dutch and the exact date that we found out that the website was no longer accessible. So, with the help of the documentation on how we handle quality assurance and notes that were added to specific targets, we can better understand the choices that were made in the harvest process such as why settings have been changed or why a website is no longer harvested.

# Scope Elements: Settings of Crawler Software for Web Archiving

The second documentation initiative we want to highlight in this paper is connected to a collection started as preparation for a national domain crawl. The Dutch national domain that we intend to harvest consists of 6.5 million websites, as was identified by KBNLs curator of digital collections. This is a vastly different scale from our selective harvest, which means different tooling is needed and manual quality analysis is ruled out. To prepare for this undertaking, we have been testing crawler tools and appropriate settings on a smaller domain: Frisia. This is a province in the Northern Netherlands, with its own history, culture and language. This province got its own top-level domain, namely .frl, in 2014. This domain is much smaller than the .nl-domain which makes it more suitable for testing crawler tools. On the Frisian websites we see the Frisian identity of the 21st century very clearly, for example in GIFs with Frisian puns. KBNL has selected approximately 10,000 Frisian websites and websites about Frisia. For example, Frisian Wikipedia (.org) and a Frisian news site on the .nl domain, are also included. In order to define the Frisian domain, researchers were asked which websites should be part of this collection, resulting in the current selection. It was a collaboration with researchers of digital humanities and the biggest cultural and heritage institution of Frisia: Tresoar. In a way this collection is still a selective crawl, but the websites will not undergo rigorous quality control afterwards because of the high number of websites harvested. Lessons learned from this regional domain crawl will be used for the Dutch national domain crawl.

For the first tests we decided to use NetarchiveSuite (NaS) [10]. This is an open-source tool, developed by the National Library of Denmark to harvest the Danish web. One of the reasons we decided to use NaS is because this also has a web interface on top of a Heritrix crawler, just like WCT. Testing a new tool and defining a new process of harvesting gave us the chance of rethinking the rationale behind our current harvesting settings and testing out variables to better understand how these have an impact on what is harvested. This resulted in better knowledge of the different settings available in Heritrix. We decided that this information would need to be stored for future reference, because it helps to better understand the collection.

During the harvest, multiple websites can be collected, by providing the crawler with more than one seed as starting point. The settings of the crawler define the limits for the harvest. They determine, for example, how deep the website is crawled in terms of slashes in the URL (path-depth) and how far the crawler moves (hops) from the starting point (seed) when counting the number of links followed. The whole path of link-hops from the seed up till the current URL is called the discovery path and can be found in the log file after the harvest is done. By looking at the log file and finding patterns between unwanted URLs and types of hop in the discovery path, conclusions can be drawn on how many of the different types of hops should be allowed in the scope [2].

Our test strategy was to try and harvest websites as complete as possible at least insofar as they were still identifiably related to the seed. By using the default settings, we allowed five trans-hops (hops based on things like embed-links) and one speculative hop (hops based on links extracted from Javascript). This resulted in too much unwanted content, like login pages for social media in various languages. We concluded that we could filter out these pages by allowing less hops. We tested this theory by allowing only two trans-hops and no speculative hops. This limit was too strict. Relevant PDF-files and stylesheets for example, on the same domain as the original seed, were missing. To mitigate this, we ran a third test with adjusted settings. Once again, we permitted two trans-hops and a speculative hop, but only if the URL was from the same domain as the seed. This time we saw that the relevant PDF-files were indeed harvested, but with minimal unwanted content.

Now why is it important to document these choices and provide them to users as context information? Determining the correct settings is an intensive process and documenting the decision process means manual work. So why invest in this? It is important for archivists to understand the material in the archives, in order to preserve digital objects and define appropriate preservation strategies. On the other hand, context information is also vital for researchers. With access to context information, they can reconstruct the life cycle of a digital object. Insight into the settings will help them understand the choices that were made in the process. This includes understanding how the original website has been transformed into an archival object. In this way, researchers will be able to determine completeness and authenticity in much the same way as they would do with physical historical sources.

# Context Elements: Collection Description XS4ALL

Context information, however, can be more than only technical details about the settings of the crawler software. Collection descriptions can also give us important historical context information, such as how the collection was created and why curators thought this collection was important to acquire [11]. A special subcollection within our selective harvest collection are the XS4ALL homepages. XS4ALL was one of the first Internet providers in the Netherlands to provide services that allowed individuals to create their own homepages from 1994 onwards. In 2019, it was announced that the brand name XS4ALL would no longer be used. This meant that original homepages would not be available anymore on the original URLs with the risk that they might disappear from the web during migration, and with that a great online source of early historical web content. This was reason enough for KBNL to take immediate action. Our curators compiled a list of the most important webpages to preserve, based on criteria like historical value, authenticity, rarity and technical and copyright considerations. A total of 19,000 websites was identified as XS4ALL homepages from the period between 1994 and 2001 [12]. The selection was harvested using WCT between 2019 and 2021. Effort not only went into harvesting the material, but also into describing the creation of this collection. This description contains information on the origins, the content and the sources that were used for discovering potential XS4ALL homepages. The collection of archived websites represents a cross section of the homepages that were hosted by XS4ALL and contains a wide variety of homepages about hobbies, animals, sports, music and personal online diaries.

Since the homepages of users were hosted on subdomains of the XS4ALL domain name and there was not a complete inventory from XS4ALL itself, tracking down all existing homepages was not a straightforward task. The collection description describes exactly which sources curators have used for arriving at the current inventory of XS4ALL homepages. It turned out that there were still old collections of links available on the web. These sources formed the groundwork for our selection. Found links were crawled and availability status was documented since many links were not available anymore at the time of harvesting. In the end not all websites that were still online have been added to the collection. A selection has been made based on criteria such as age or cultural importance. These selection criteria have been documented, as well as the methods used for dating the webpage. These range from content elements within the page to last-modified date of images. This information will help researchers understand the rationale behind the selection and the reason why websites in the collection are deemed to be representative of a certain period in the history of the web.

Without a list of all existing items that are supposed to be in a collection, it is difficult to determine collection completeness. However, for research purposes it is important to be able to gauge whether an archive has a representative collection. Say a researcher discovers an XS4ALL homepage that is not present in our collection. Using the collection description, source material and other documentation mentioned, it will be possible to determine why this homepage was not archived. For instance, because it was not present in one of the existing link sources, or because the homepage was not available anymore at the time of archiving. This information is important from the perspective of source criticism and functions as an implementation of the concept of context elements. It is also valuable as an independent mechanism for establishing how complete the collection is as required by the guidelines for trustworthy digital repositories (ISO-16363).

# Conclusion

In our opinion, context information can provide insight into the process of creating an archival object from files on the live web. The technology used, the process of quality control as well as the curatorial context of a web collection all impact the integrity and authenticity of the archival objects and the collection itself. The result of harvesting is what is stored in the archive and is therefore also the source material for research on the history and topography of the web. As a trustworthy digital repository, we have the responsibility to provide our users with enough information to keep our collection items understandable. For web archives we think it is important to provide information on the settings of the harvester, on the process of harvesting and quality control and on the curatorial decisions taken in order to acquire the content. Together these three types of information provide ways for our crawls to ‘talk’. They tell us more about how the digital objects came into being and which steps have been taken to ensure quality. In this way, context information will provide researchers with the necessary means for evaluating integrity and authenticity of the web collections that are part of our digital heritage.

#### ACKNOWLEDGMENT

The authors would like to express their thanks to Iris Geldermans, Johan van der Knijff and Kees Teszelszky for reviewing the first draft of this paper.

# REFERENCES

1. National Library of the Netherlands, “Content Strategy”, 2019. https://www.kb.nl/sites/default/files/documents/content\_strategy\_eng%20%282%29.pdf
2. S.C. van den Eijkel and D. Steinmeier, “What is an archived website?”, unpublished
3. National Library of the Netherlands, “Preservation Plan”, 2019. https://www.kb.nl/en/about-us/expertise/preservation-policy
4. C. Rogers, Authenticity of Digital Records: “A Survey of Professional Practice”, in *Canadian Journal of Information and Library Science*, vol. 39, no. 2, pp. 97-113, June 2015. https://muse.jhu.edu/article/590936
5. E. Maemura, N. Worby, I. Milligan and C. Becker, “If These Crawls Could Talk: Studying and Documenting Web Archives Provenance,” in *Journal of the AAOXIATION FOR Information Science and Technology*, vol. 69, no. 10, pp. 1223-1233, October 2018. https://doi.org/10.1002/asi.24048
6. B. Sierman and K. Teszelszky, “How can we improve our web collection? An evaluation of webarchiving at the KB National Library of the Netherlands (2007-2017),” in *Alexandria: The Journal of National and International Library and Information Issues*, vol. 27, no. 2, pp. 94-107, August 2017. https://doi.org/10.1177/0955749017725930
7. Web Curator Tool, https://webcuratortool.org/
8. Heritrix, https://github.com/internetarchive/heritrix3
9. I. Geldermans, “Historical growth of the KB web archive,” online available at KB Lab, The Hague, 2021. https://lab.kb.nl/dataset/historical-growth-kb-web-archive
10. NetarchiveSuite, https://sbforge.org/display/NAS/NetarchiveSuite
11. K. Teszelszky, Chapter: “The historic context of web archiving and the web archive”, in: N. Brügger and D. Laursen (eds.) *The Historical Web and Digital Humanities. The Case of National Web Domains*, 2019.

https://www.taylorfrancis.com/chapters/edit/10.4324/9781315231662-2/historic-context-web-archiving-web-archive-kees-teszelszky

1. P. de Bode and K. Teszelszky, Collection description XS4ALL homepages, 2021.

https://zenodo.org/record/5055571#.YVGivH1cKUk (Full text only available in Dutch, summary provided in English).