Chapter 3: A Collaborative Approach to Research Data Management in a Web Archive Context

Hugo C. Huurdeman, University of Oslo Library
Jaap Kamps, University of Amsterdam

Abstract

In our times, researchers create and gather their datasets, irrevocably changing from solely 'small' to increasingly 'big' data. Institutions provide more and more services to manage, store and access this essential research data. However, despite the invaluable services provided by repositories, there is still a mismatch between the myriad of operations carried out by researchers, and the storage services offered by generalized repositories: often, only the end products of the research process are stored. This may mean that valuable operations and transformations of data during the research process are lost, in effect also making it harder for future researchers to interpret this data. To address this issue, and to bring researchers and institutions closer together, we propose a more collaborative approach, potentially involving researchers in their entire workflow, from data creation to use and potential reuse. This proposition is contextualized via the concrete use case of the web archive. Organizations and individuals across the globe archive web content, assembled into vast web archives. These web archives could potentially be used as research datasets in various settings, ranging from computer science to the humanities. However, they have scarcely been used for research thus far, due to a number of limitations, including suboptimal access services for research. This chapter discusses the experiences and lessons learned in the concrete case of the web archive, and their implications for research data management at large.

3.1 Introduction

Web archives document the web by archiving its contents, and provide large datasets, which could potentially be used for research in different settings, ranging from computer science, to social sciences and the humanities. However, they have scarcely been used for research thus far, due to various limitations, including data deficiencies and suboptimal access services for research. This has similarities with research datasets available in repositories, which have great potential: there is an unprecedented amount of data

generated in the research process, which is increasingly shared via these repositories. Borgman (2012) distinguished four reasons for data sharing: "to reproduce or to verify research", to make "results of publicly funded research" public, "to enable others to ask new questions of extant data", and "to advance the state of research and innovation". However, despite the obvious opportunities inherent in research data sharing, a myriad of challenges exists in the management of this data. At a broad level, research data management "involves all the process that information from research inputs undergoes as it is manipulated and analysed *en route* to becoming a research output" (Wilson et al. 2010). Capturing this process, as well as making it available for future researchers, is no straightforward task.

This chapter examines the complicated case of the web archive, which involves many barriers to successful research use. Using this case as a basis, the chapter discusses the larger implications of making archives and datasets searchable for a research community. This chapter consists of five parts. In the next section (4.2), we introduce the concept of a web archive, characterize what constitutes the archive, and discuss the actors involved in web archiving. Second, section 4.3 looks at the properties of web archives as research datasets and their potential deficiencies and limitations. Then, we introduce a concrete case in section 4.4, in which we performed experiments towards making web archives available as research datasets. 4.5 describes further ways to overcome limitations of web archive access interfaces. Finally, section 4.6 indicates the implications for research data management at large.

3.2 An introduction to web archives

First, we introduce the concepts utilized in the remainder of this chapter. We discuss the rationale behind web archiving, various definitions, and the variety of actors which are involved in archiving.

3.2.1 Why archive the web?

The ever-growing web takes up a pivotal role in our everyday lives. We use the web to lookup information, to communicate, and for our daily entertainment and leisure. However, the web is of a highly ephemeral nature: if a server disappears, the content is lost (Masanés 2006, 7), and if a website is renewed, content may be moved, changed or deleted altogether. Hence, "the content and structure of the web are constantly in flux", and proactive steps have to be taken to ensure that web content will be preserved (Dougherty and

Meyer 2014). Thus, as Kahle has indicated[1], through web archiving, we may enable a "memory" of the web and avoid to get stuck in a "perpetual present". Various individuals and institutions at local, national and international scales have taken up this challenge, together harvesting Petabytes of valuable web material. In the complex and volatile environment of the current web, however, archiving institutions have a hard time keeping up with the technological developments, but also with the web's massive scale. Estimates based on different samples taken in 2012 indicated that about 35-90% of the web was at least archived once (Ainsworth et al. 2012), but this does not even take into
account information in the Deep Web, unreachable for web archive harvesting tools. Further hindrances are formed by privacy issues, intellectual property and copyrights (Masanés 2006). Before delving deeper into these issues, we first must arrive at a definition of web archiving[2].

3.2.2 Defining web archiving

Web archiving has been defined by the International Internet Preservation Consortium (IIPC)[3] as "the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use." Another definition by Ball (2010) focuses on more specific procedural aspects, characterizing web archiving as "the selection, collection, storage, retrieval, and maintenance of the integrity of web resources". Complementary to this more institutional perspective, Niels Brügger, a web historian, focuses on the intention and rationale behind archiving: "Web archiving means any form of deliberate and purposive preserving of web material" (Brügger 2009). Brügger elaborates that this definition implies that archiving is a *conscious* act: the act of preserving the material itself, but also the conscious reasoning about why the material is collected and preserved. Different actors may be involved in this process, discussed in the next
section.

[1] http://brewster.kahle.org/2015/08/11/locking-the-web-open-a-call-for-a-distributed-web-2/ (accessed: 29/02/16)

[2] Despite the term 'web archiving', web archives paradoxically often do not yet adhere to long-term digital preservation standards, such as the ISO's OAIS Reference Model (available at: https://www.iso.org/standard/57284.html) (accessed: 29/02/16)

[3] This definition is available via: http://www.netpreserve.org/web-archiving/overview/ (accessed: 29/02/16)

3.2.3 Classifying web archiving actors

An increasing number of institutions, companies, groups and individuals are collecting web material[4]. As this very diverse group of actors implies, web archiving may be done for a great variety of purposes (see Brown (2006); Masanés (2006); Brügger (2009)). The way the archived web is formed differs based on who does the archiving, when, and for what purpose (Brügger 2005 as cited by Rogers 2013, 64). These purposes may include *collection building and preservation* (for example by libraries and archives), *research* (for example in the context of a research institution), or applicable *legislation*. The latter be may obliging web archiving due to *legal deposit* laws, requiring institutions to document all published documents in a country (e.g. national libraries in the UK or Denmark)[5], or due to *archival laws*, obliging government entities to archive their own website. In terms of their funding, the initiatives can further be divided in *state-funded*, *nonprofit* and *commercial* web archives (Masanés 2006, 41).

Brügger (2009) provides a broad division of web archiving efforts based on their scale: on the one hand, *macro* archiving entails the archiving of web material by professionals, often in the context of national or local institutions. On the other hand, *micro* archiving involves small-scale archiving carried out by researchers and other individuals, based on a "here-and-now" need to preserve an object of study. The prime example of an institution applying a macro perspective to web archiving is the Internet Archive, a non-profit institution which began archiving the web in 1996 on a massive and transnational scale. Other institutional initiatives, often state-funded, may range from local-level (e.g. a Municipal Archive), to regional and national scales (for instance the UK Web Archive or the Netarkivet in Denmark). Also, an increasing number of commercial initiatives provide web archiving as a service[6].

The micro-level of archiving is for instance reflected by individual researchers, who may gather their own collections of web material in the context of their research. Finally, a more blurred category exists of bottom-up 'crowdsourced' initiatives like ArchiveTeam, that jointly archive web

[4] An updated list of web archiving initiatives is available at: https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives (accessed: 29/02/16)

[5] To be more precise, legal deposit legislation "defines a legal obligation for publishers to deposit copies of all published works with designated libraries, in order to maintain a comprehensive collection of a nation's published output" (Brown, 2006)

[6] For instance, Archive-It (https://www.archive-it.org/ (accessed: 29/02/16)), a spin-off of the Internet Archive, provides subscription-based web archiving services, for instance used by cultural institutions. The Canadian-based Pagefreezer (https://www.pagefreezer.com/ (accessed: 29/02/16)) provides web archiving, also for digital evidence purposes, while Archiefweb (http://www.archiefweb.eu/ (accessed: 29/02/16)) captures many Dutch government websites.

material, using shared methods and highly collaborative approaches[7]. In this chapter, we mainly focus on the macro approach to web archiving, i.e. the larger web archives assembled by cultural heritage institutions.

3.2.4 A transition from building infrastructure to supporting use

In the early years of web archiving, the archives predominantly focused on preservation and creating the infrastructure to harvest internet pages – in itself no easy task, due to the voluminous scale of the web. As Thomas et al. (2010) have argued, much of the preliminary archiving efforts have been done "from the point of view of archiving for its own sake", but less work has been carried out towards the actual use of these archives by researchers. As Rogers (2013, 72) has put it, "Web archiving infrastructure receives scholarly and nonscholarly attention; the archived materials –the primary source materials– gain less notice." A related issue, as indicated by Thomas et al. (2010) is that "the traditional practices of the field of Library and Information Science" have dominated web archive development, not necessarily providing the right handles for humanities and social sciences' researchers. Dougherty and Meyer (2014) suggest that there is a "wide gap between the researchers who need archival data sets to support their studies of online phenomena, and the archivists and other practitioners who have the expertise to build such collections and the tools to manage and access them". Moreover, they indicate that these efforts have "so far not yet provided reliable methodological solutions for researchers who wish to use archived web materials". In the next section we zoom in on the issues that may arise in the use of web archives as research datasets.

Summarizing, the past two decades have shown an imminent rise of web archiving initiatives, preserving and providing access to our online past. A gradual move from building infrastructure to supporting use has emerged, but we are still in the beginning of that transition. The myriad of ongoing initiatives, ranging from bottom-up initiatives to large-scale institutional archiving, shows the importance and multidimensional aspects of web archiving. However, it also results in datasets that vary in multiple ways, due to differences in the purpose of the archives, their approaches and selection criteria. To what extent these aspects influence the suitability of web archives as research datasets is the topic of the next section.

---

[7] To give a practical example, the Dutch pre-Facebook social network Hyves was going offline on a short notice, and via ArchiveTeam a joint collective of individuals managed to harvest around 9M public profile pages in time (http://www.archiveteam.org/index.php?title=Hyves (accessed: 29/02/16))

3.3 Web archives as research datasets

Web archives, due to their vast and diverse contents, can provide a valuable resource for scholars in various disciplines, for instance computer science, the humanities and the social sciences. In theory, web archives may allow for novel research questions and methods, but in practice many carefully crafted archives have remained underused (Dougherty and Meyer 2014). Utilizing web data in general, and archived web data in particular, introduces various challenges when performing research. This section focuses on these challenges, and looks at various limitations of web archives in research context. These limitations may be divided in three categories: limitations in *data quantity*, limitations in *data quality* and limitations in *access*.

3.2.1 Limitations in data quantity

Web archiving is predominantly performed by web crawlers, which "harvest content from remote web servers" (Brown 2006, 50). Similar to the crawlers used by common search engines, these crawlers iteratively follow hyperlinks within webpages to capture content. This can be done using three main strategies (Brügger 2011), which influence the data which is captured in archives. Broad *domain* crawls are delimited by the boundary of the national top-level domain (e.g. *.uk*), *selective* snapshots focus on a predefined selection of websites, and *event* harvests focus on important ongoing events. These three strategies feature distinct trade-offs in terms of the breadth and depth of captured content.

First, the *broad domain* strategy entails taking a snapshot of web documents at one or more points in time. This approach usually implies a 'breadth-first' strategy, meaning that the web crawlers focus on capturing the breadth of web material as opposed to the depth. Hence, a wide range of content may be captured for posterity, but it also means that material located deep in a web site may not be captured. Domain crawls also result in very large datasets, making quality assurance hard to manage, and can take a very long time to complete. For instance, the full .uk domain crawl of the British Library in 2013 took almost eleven weeks to complete, leading to a sizable set of material totaling in 31TB[8]. A *selective* strategy, on the other hand, may result in a more manageable set of material. This strategy is based on a (finite) selection of websites, based on certain properties, such as subject, creator, genre or domain (Brown 2006, 31). The selective nature implies that material outside the selection lists is excluded from archiving, although using this approach, the amount of captured material per site may be higher: crawlers are

[8] As stated on: http://britishlibrary.typepad.co.uk/webarchive/2013/09/domaincrawl.html (accessed: 29/02/16)

usually configured to follow links deeper into a domain. Hence, as Brown (2006, 32) has argued, the selective approach "is likely to facilitate a more detailed understanding of the properties and qualities of the individual resources collected" – it may be feasible for archiving entities to perform quality control, or to adjust crawl settings for individual websites. A third common strategy is *event harvesting*, i.e. the harvesting of material related to events on a local, national or international scale. For instance, in the context of the International Internet Preservation Consortium (IIPC), institutions perform collaborative harvests. Covered events may consist of anticipated and planned events (e.g. the the 2014 Winter Olympics in Sochi) or unplanned events (e.g. the 2005 Katrina hurricane).

Regardless of the approach, temporal omissions exist, and, depending on the webpage captured, these may influence the types of research which can be performed with a web archive. In effect, as Masanés (2006, 17) has argued, archiving "always implies some selectivity, even if it is not always in the sense of manual, site-by-site, selection", and "the archived portion of the web will always only be a slice in space and time of the original web". Moreover, for individual researchers, limitations may exist in the large multipurpose *macro*-level archives, as they often apply a one-sizefits-all approach to web archiving, which may use generalized crawl settings, crawling schemes, and selection policies. For instance, a researcher's interest may involve sources that lie outside the scope of a web archive's selection criteria (for example highly controversial websites), or a researcher may need more frequent harvests for a certain website. In addition, certain popular websites, such as Facebook, are difficult or impossible to crawl using regular crawling techniques.

3.2.2 Limitations in data quality

There are inevitable limitations in the data quality, which we define as the extent to which captured web content resembles the original content on the 'live' web. First of all, there are technical issues related to the 'archivability' of a website: some data formats and certain types of interactive websites cannot be archived, for instance form-based pages, or dynamic web content based on HTML5 techniques. This leads to an incompleteness at several levels: at the level of the website (individual pages may be missing), but also at the level of the page (page elements, such as embedded material may not be included). Rogers (2013, 64) summarized it as such: "In a sense, the 'new media' elements (cookies, embedded material, recommendations, comments, etc.) are eliminated for posterity, and a traditional content container, looking somewhat broken for its missing pieces, remains as the 'archived website.'" Furthermore, the "interconnectedness", i.e. the unique hyperlinked-based nature of the web may get lost (Masanés 2006, 17; Rogers 2013, 63); and in

effect, the archived website, an assembled object, becomes detached from its larger context (Helmond 2015, 118). Thus, we may arrive at something different from the 'live' web in a multitude of ways.

In addition, temporal inconsistencies may occur (Brügger 2005, 2009, 2011). For instance, capturing a large website, such as *www.cnn.com*, may take a long time, during which some contents have already changed. For instance, news items have been added to the homepage. However, the harvested pages may both reflect the initial state (when the harvesting started) and other states later in time. This has led Brügger to argue that an archived page is a "version" and not a "copy" of a website[9]. Moreover, what is the *right* version of content is often unknown: web servers may adapt content to each request, for instance based on the device that a user utilizes for accessing the web. Hence, the web may be seen as "a black box with resources, of which users only get instantiations" (Krishnamurthy and Rexford 2001, as cited by Masanés 2006, 13). Thus, the captured material may in many cases be different than the original resources.

3.3.3 Limitations in access

As evidenced in this chapter so far, web archiving institutions across the globe are spending substantial efforts on *collecting* and *preserving* our valuable web heritage, but another crucial issue is *providing access*. Several factors influence access to archives.

First of all, legal reasons may impede archive access. While some archives are fully accessible online (e.g. the Portuguese web archive), for the majority of archives this is not the case. Some web archives are only accessible from the library or archive's premises (e.g. the National Library of the Netherlands), other archives are partially accessible online (e.g. the UK Web Archive), and some archives, so-called dark archives, do not provide access to end-users at all.

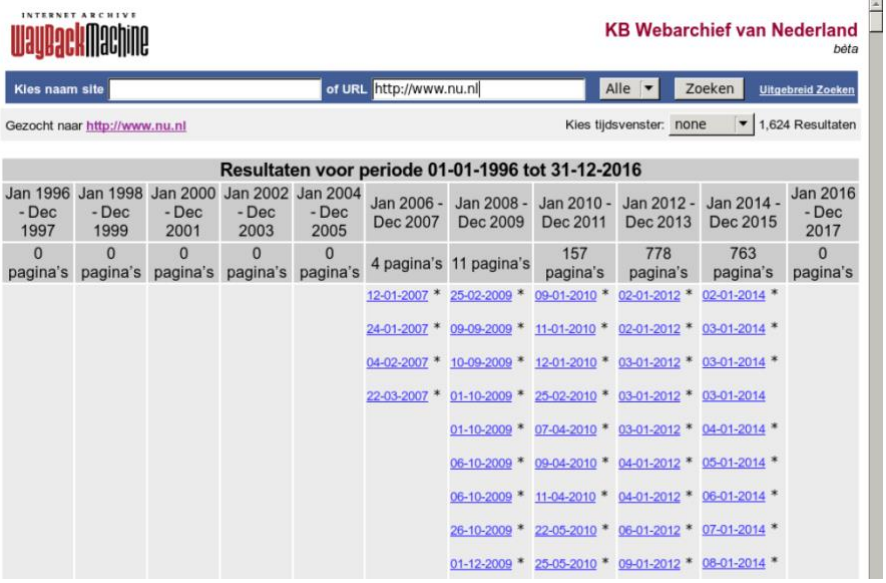
The second limitation, the main focus of the remainder of this chapter, lies in the access systems and interfaces, the intermediary between the data in the archive and the potential user of this data. Web archives have taken different approaches to provide access, including *URL-based*, *browsing* and *search-based* access options. The most common way of accessing content is through the Wayback Machine, which "allows users to locate archived website snapshots, to differentiate between multiple snapshots of the same site collected on different dates and to navigate across all content collected at a

[9] For these reasons, Brügger has classified the content of web archives as "re-born digital material" instead of "born digital material"

certain point in time, effectively recreating the original context of that content" (Brown 2006, p.135). This interface provides URL-based access to web pages in web archives, but evidently, this necessitates the knowledge of the URL and date of a certain page or site. Some archives, such as the UK Web Archive, also provide ways to browse the contents of the archive through subjects and collections[10]. However, for archives, these hierarchical classifications may be difficult to create and maintain; and a user's navigation may be "limited by the classification decisions made by the archive" (Brown 2006, 129). Both URL and browse-based access approaches are still 'document-centric' methods (Hockx-Yu 2014), focusing on specific documents. In effect, as Hockx-Yu (2014) has indicated, the current user interface for web archives may work "well with small, curated collections but does not scale up and provide the users with a functional way to use larger collections". The 'single URL' approach implied in the current tools (Ben-David and Huurdeman 2014) facilitates 'close reading' of individual text, but precludes 'distant reading', which implies a move from just single documents to providing the broader picture (Moretti 2013).

To a certain extent, the addition of search-based access to web archives has allowed for this broader view, substantially enhancing access, and "scaling the analysis from the single URL to the full archive" (Ben-David and Huurdeman 2014). Search-based access has many advantages, and overcomes the necessity to know URLs in advance, but the next section will show that also a number of issues are involved.

[10] http://www.webarchive.org.uk/ukwa/browse/ (accessed: 29/02/16)



In sum, web archives potentially provide numerous opportunities for research, but their potential for reuse as research datasets has not been fully harnessed yet. In part, this is caused by issues in *data* and *access*, only corroborated by the variety of actors involved in web archives discussed in the previous section, which take different approaches to archiving. To better understand these limitations, and to potentially amend them, we examine the actual use of the web archive in a practical setting. Taking a large Dutch research project about web archives as its basis, the next section discusses the pitfalls and opportunities of using web archive data for research, as well as ways to improve scholarly access.

3.4 Experiments towards making web archives available as research datasets

After discussing the limitations of web archives as research datasets in the previous section, we now discuss experiments towards making web archives more useful as research datasets. Within a project setting, we focused on improving access to web archives in a scholarly context.

3.4.1 Introduction to WebART

A highly influential and long-running research program in the Netherlands was the CATCH (Continuous Access To Cultural Heritage) program[11], which aims at making "the collections of museums, archives and historical associations more accessible." Between 2005 and 2016, the program has funded 18 multidisciplinary projects, in which researchers and heritage institutes collaborated to improve access to Dutch cultural heritage collections.

[11] http://www.nwo.nl/catch (accessed: 29/02/16)

The WebART project[12] (2012-2016), part of CATCH, has looked at ways to evaluate the current use of web archives and to design novel access methods, both from theoretical and practical perspectives. In the WebART project, the University of Amsterdam[13] and Centrum Wiskunde & Informatica[14] (CWI) joined forces with the National Library of the Netherlands[15] (KB). The interdisciplinary
project involved researchers with backgrounds in computer science, information science and new media & digital culture. The main collection studied in the project was the KB's web archive. The KB initiated their web archiving program in 2007, employing a selective policy. As of 2016, over 10,000 Dutch websites were harvested on a regular basis, the full archive amounting to over 18 Terabytes. Initially, this large amount of data was only accessible via the Wayback Machine, severely limiting the research opportunities. Therefore, we set about to explore additional and novel access methods in a collaborative setting.

3.4.2 Exploring researchers' needs related to web archives

The WebART project organized and participated in a series of events in 2012 and 2013 (see Table 1). These events shed more light on the needs of researchers that use web data to perform their research, and that take the web as their object of study. The participants in the events ranged from Master students to PhD-level researchers and renowned senior scholars, reflecting a wide range of potential web archive users and use cases. Many of the participating scholars were affiliated with the Digital Methods Initiative (DMI), which is "one of Europe's leading Internet Studies research groups"[16]. It "designs methods and tools for repurposing online devices and platforms (such as Twitter, Facebook and Google) for research into social and political issues". One of their research foci is the website as an archived object (Rogers 2013, 61). To facilitate these types of research, the Digital Methods Initiative provides a comprehensive set of tools to study the web (seventy at the time of writing), including tools for data extraction, scraping, processing, analysis and visualization[17]. Some of these tools can also be used in conjunction with web archives. Based on the insights from the events listed in Table 1, and an intense collaboration within the project, a set of web archive retrieval tools (dubbed WebARTist) was designed to accommodate for web researchers' needs (Huurdeman et al. 2013; Ben-David and Huurdeman 2014; Huurdeman 2015). This was done in in an action research setting, a "collaborative

[12] 'Web Archive Retrieval Tools', http://www.webarchiving.nl/ (accessed: 29/02/16)
[13] http://www.uva.nl (accessed: 29/02/16)
[14] http://www.cwi.nl/ (accessed: 29/02/16)
[15] https://www.kb.nl/ (accessed: 29/02/16)
[16] https://digitalmethods.net/ (accessed: 29/02/16)
[17] https://tools.digitalmethods.net/ (accessed: 29/02/16)

approach to investigation", in which the researcher engages in "examining current processes, taking action to improve those processes, then analyzing the results of the action" (Pickard 2007, 134). This led to three phases of tool development: first, an analysis of the use of existing web archive access tools ("problem identification", section 4.4.2.1), followed by the development of search-based access ("implementation", 4.4.2.2), and finally the evaluation thereof (4.4.2.3).

Table 1: WebART events and event participation. More information about each event can be found via http://www.webarchiving.nl/events/ (accessed: 29/02/16)

Event	Date	Description
(1) DMI Summer School	08/12	Participation in the Digital Methods (DMI) Summer School, developing research scenarios for the Dutch Web archive
(2) Web Archiving: Theorized Practices	12/12	Organization of an ACHI seminar involving renowned scholars using the Web as a corpus
(3) DMI Winter School	01/13	Participation in DMI Winter School, developing research scenarios for the Dutch Web archive
(4) WebART CATCH Event	04/13	Symposium on Web archives with speakers from the British Library, Library of Congress and the University of Amsterdam
(5) Exploring Israeli Politics Online	05/13	Workshop at Bar-Ilan University, Israel, aimed at analyzing political Web archive data
(6) DMI Web Archiving Day	09/13	Workshop and focus group, evaluating all WebARTist tools up to that point
(7) New Media Research Masters	11/13	Seminar for new media Master student, creating proposals for research using the Dutch Web archive

3.4.2.1 Phase I: Existing web archive access tools

In this initial phase in Summer 2012, the main issues in web archive research and web archive access were explored via a literature review and by active participation in a summer school. The WebART team participated in the Digital Methods Initiative's (DMI) Summer School (Table 4.1 [1]), a yearly summer school in which motivated scholars "learn and develop research techniques for studying societal conditions and cultural change with the Internet".[18] In particular, the selection policies and content of the web archive of the Dutch KB were explored, as well as the possibilities of doing research using existing web archive access tools, such as the Internet Archive's

[18] https://wiki.digitalmethods.net/Dmi/DmiSummerSchool/ (accessed: 29/02/16)

Wayback Machine. Gained insights could be used in later development of solutions
for improving web archive access.

Thus, the WebART team members collaborated with participants of the summer school in a project-based setting to develop research scenarios using the Dutch web archive. At this point, activities consisted of analyzing the KB archive's selection list, comparing the coverage of the Dutch web archive with the Internet Archive's web archive, and network analysis of different categories of websites in the archive's selection list. Furthermore, illustrating the wide range of research questions which can be asked to web archives, one project during the summer school analyzed 'trackers' in the Internet Archive, i.e. objects embedded in the source code of webpages which can track user behavior, often for advertising purposes (see Helmond (2015, 124)).

This research was enabled by combining DMI tools[19] with the Internet Archive's Wayback Machine. The activities in the Summer School provided more insights into the KB's web archive, its selection policies and the potential research opportunities of web archives in general.
The explorations in this phase resulted in an initial understanding of the KB's web archive, its selection policies and the potential research opportunities of web archives in general. Also, a number of limitations of available access tools were confirmed. In particular, the document-centric and 'single site' approach to web archive research of the Wayback Machine interface posed problems, impeding analysis beyond the page level (Hockx-Yu 2014; Rogers 2013; Ben-David and Huurdeman 2014). Without resorting to external tools, the interface of the Wayback Machine predominantly facilitates qualitative inspections of web archive content, as opposed to analyzing broader patterns and underlying structure. Even though specific tools allowed researchers to analyze multiple URLs[20], and the source code of pages[21], previous knowledge of URLs of online resources was still required. As web archives contain pages from the web of the past, the consequence is that a substantial amount of resources cannot be located. In order to support scholarly use of web archives within WebART, the natural next step was plan the development of search-based access tools, allowing researchers to dynamically search content in the web archive.

[19] https://tools.digitalmethods.net/ (accessed: 29/02/16)

[20] The DMI 'Internet Archive Wayback Machine Link Ripper' and 'Network Per Year' tools, at https://tools.digitalmethods.net/ (accessed: 29/02/16)

[21] E.g. tracker 'fingerprints': https://tools.digitalmethods.net/beta/trackerTracker/ (accessed: 29/02/16)

3.4.2.2 Phase II: Development of search-based access

In the next *implementation* phase, a full-text search system for the Dutch web archive was introduced, potentially offering additional support to new media scholars in their research process. To design the system, the thesis author collaborated with a new media researcher and a computer science researcher on a day-to-day basis. Additionally, further insights were gained via various workshops performed with other new media researchers (Table 4.1 [3,4,5]). Hence, the actual functionality of the search tools (both back-end and front-end) was developed in a bottom-up way (Huurdeman et al. 2013), meaning that researchers in- and outside the WebART project were consulted for building the system's functionality.

The data of the Dutch web archive is stored in the ARC-format, which aggregates web resources and their metadata, as well as crawl-related metadata[22]. After experimentation with different information retrieval solutions, the Terrier Information Retrieval platform (Ounis et al. 2006) was used to create a search environment. Terrier is a highly scalable and customizable search engine written in Java. It functioned with the Hadoop computing cluster which we utilized[23]. To informally evaluate the initial search tools, the WebART team participated in the DMI Winter School in January 2013 (see Table 1 [3]), and a group of seven researchers started to use the developed tools in combination with an extracted dataset of nu.nl[24], a leading Dutch news aggregator. As it turned out, the most urgent functionality request by researchers consisted of a feature to export resultsets. For instance, when querying for 'eurocrisis' in the Dutch web archive, researchers wanted to export all results into a structured format, which could be imported in their own analysis and visualization tools (e.g. Excel, or Gephi). Hence, they wished to perform their analyses outside of the system. In addition, various enrichments to the data were requested and subsequently implemented, including news locations and information about pages' outlinks, which facilitated performing temporal hyperlink analysis, considered "an important way to reconstruct views of the past" (Huurdeman et al. 2013). In sum, the tools created in this collaborative setting facilitated new possibilities in the exploration and use of archived material, allowing for both content-based and structural analysis types, as distinguished by Schneider and Foot (2004).

Suggestions for improvement were taken into account during subsequent iterations of search tool development. In addition, based on the fact that researchers performed a large part of their analysis outside of the system, we

[22] http://archive.org/web/researcher/ArcFileFormat.php (accessed: 29/02/16)

[23] Due to the sheer size of the KB's dataset, amounting to over seven Terabytes at the time, the extraction and indexing had to be carried out via a Hadoop computer cluster at SURF's Dutch national e-infrastructure: http://www.surf.nl/ (accessed: 29/02/16).

[24] http://www.nu.nl/ (accessed: 29/02/16)

intended to also provide more analysis functionality within the system. While it is certainly important to integrate export functionality to facilitate research using scholars' familiar tools, analysis functionality within the system may provide a useful extension for performing exploratory research using the system. Various interfaces were created on top of the full-text search system, which included specific functionality, ranging from full-text search to statistical visualizations and aggregated results. The visualizations and aggregations (e.g. results grouped by site names, thematic categories, temporal occurrences, outlinks and occurring words) now allowed for viewing the big picture, i.e. 'distant reading' (Moretti 2013). At the same time, the more regular search results page allowed for closer inspection of individual results, thus still facilitating 'close reading', i.e. studying the webpages and websites involved. The extent to which this improved scholarly access to web archives is discussed next.

3.4.2.3 Phase III: Evaluation of search-based access

In September, 2013, the search tools created in the context of WebART were more formally evaluated, and six media researchers from the Department of Media Studies participated (Table 4.1 [6]). This day included presentations by researchers, a survey and focus group. After exploring the possibilities of the WebARTist toolset, participants in the focus group indicated that the WebARTist system supported "looking at data rather than sites", and that it supported "the shift of studying a web archive through queries"; a big step forward as compared to earlier URL-based Wayback Machine interfaces. Furthermore, the "aggregate views and bar graphs" were found "extremely useful" by researchers, as they allowed them to view a summary of thousands of results in one page.

When discussing the possibilities of this system, the researchers indicated various topics of interest: first of all, in the context of web history, it allowed them to "conjure up past states of the web" using daterange queries, and to "derive periodizations of the web". For instance, a mentioned research topic to study via the archive would be the rise of social media on the web. Other mentioned possibilities were creating "source hierarchies", i.e. looking at the "dominant sources in the archive". For instance, WebARTist would allow a researcher to query for 'financial crisis' and find out which are the key sources for this topic in the archive. Thirdly, researchers could look at the "keyword uptake", the occurrence of keywords in the archive over time, also aided by the aggregation and visualization possibilities of the tools. For instance, a researcher could search for 'climate change' and look how various categories of websites (e.g. news or government websites) evolve over time, including the language used on these websites. Finally, a selection-based archive like the Dutch web archive is by nature incomplete. However, using a feature in

WebARTist which showed whether a site is on the original seedlist or not, it is possible to study a phenomenon labeled "accidental" or "incidental" archiving, i.e. the occurrence of certain, unselected, sites in the web archive.

The focus group in September 2013 indicated that the experimental prototypes, however useful, also needed more functionality to be usable for research. In particular, the researchers requested additional support for selection methods, analysis, collection making, and more transparence. First of all, it was indicated that improved selection methods were needed, in concordance with researchers participating in a workshop in Israel (Table 1 [5]). One scholar argued that the limitation of the Wayback Machine is that one always
has to start with a URL, while the limitation of WebARTist was that the starting point has to be a query. Suggested expansion included the possibility to start with selecting a site or category of sites, or using lists including "historical web directories, blogrolls and link lists". Also, web archives generally include a massive amount of content. Hence, sampling (using a subset of the large amount of data for initial analysis), was deemed useful for future versions of WebARTist. As a researcher put it, it should be possible to "first view a sample to get a sense of what's in, and if it's worth pursuing, get the full data". Furthermore, in terms of further analysis functionality, a comparison feature was deemed useful, to directly compare differences in resultsets without manually opening multiple browser windows. In particular, a feature to compute and visualize the differences between resultsets was suggested. Moreover, researchers mentioned the possibilities to create custom collections[25] and to add annotations as an important addition: to "make the collection you build accessible and annotate it for other users". Finally, *transparence* was a key issue to be addressed in future systems. Researchers would like to know more about the archive's selection procedures, the (in)completeness of the archive and algorithms. Clearly, an archive's selection policies and the completeness of harvests have a direct influence on the items which are retrieved, leading a researcher to argue that "data is still a crucial factor". Moreover, ranking and retrieval algorithms in access interfaces may excert a profound influence on the degree of relevant items appearing in results lists[26]. Evidently, these issues may influence subsequent analysis. This suggests a need for more contextualization and transparence in future access interfaces.

[25] Here, we define 'collection' as an assembled set of materials around a certain research theme.

[26] To take a practical example, in the DMI Winter School (Table 4.1[1]), researchers looked at news coverage about the former Egyptian president, Hosni Mubarak. Initial indexing settings resulted in high *recall*, but low *precision* of retrieved items when querying for "Mubarak", impeding temporal analysis of the found articles. Customized indexing settings, ignoring the text surrounding articles (e.g. other news items), resolved this issue.

Summarizing, in the co-design setting of the WebART project, access tools for web archives have been iteratively refined, thus arriving at tools which provide enhanced research support. In the following section, we discuss the suggested future steps in the process of tool development.

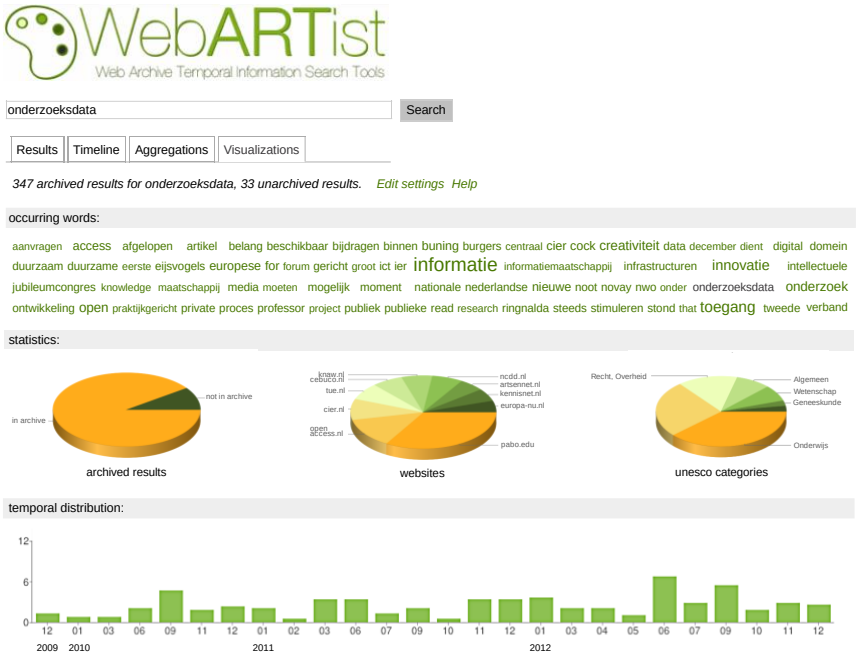


Figure 2: Screenshot of WebARTist for the query ‘onderzoeksdata’ (research data).

3.5 Towards ‘research engines’

The previous section has shown that to move beyond mere search engines, and to create true ‘research engines’ for web archives, several limitations of access interfaces have to be overcome. In this section, we suggest two ways to achieve this: increasing transparency, and providing process support for scholars.

3.5.1 Increasing transparency

In a *macro* archiving context, Figure 3 summarizes various influences on the eventual results which a researcher retrieves via a search-based web archive access interface. Due to the curatorial decisions (1), crawler limitations (2), but also the limits of access systems (3) a researcher may actually miss a substantial amount of data, which may be potentially crucial for analysis.

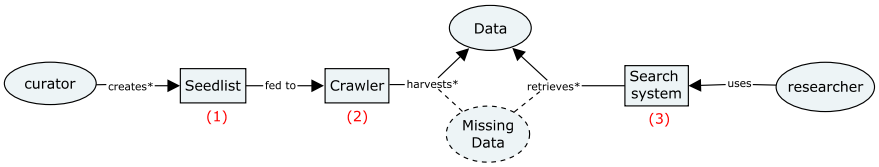


Figure 3: Schematic overview of actors and interactions in a "macro" web archiving context. Due to curatorial decisions (1) and crawler limitations (2), data may be missing, while technical and algorithmic properties of search engines may influence retrieved material by a researcher (3).

For point (1), the curatorial decisions, insights can be provided to prospective researchers by making selection policies, criteria and selected websites available at an institutional level. For instance, the Dutch KB provides this information via its website[27]. Besides this issue, however, web crawling involves various technical issues, indicated by (2) in the diagram. These impede some websites from being harvested correctly. Usually, though, documentation about crawler settings and their influence on obtained data are not provided. For enabling a better understanding of archived content, this documentation should be made available to researchers as well.

Finally, at point (3), standard full-text search systems may hide the particularities of underlying data to their users. As discussed in Section 4.4.2.3, ranking and retrieval algorithms may influence which "relevant" items a researcher retrieves. Moreover, search systems typically hide which data is *missing* from the archive. In a recent study, we have explored solutions for the latter issue. By harnessing the link structure and anchor text (the textual descriptions of links) of the web material contained in the archive, we have analyzed the inherent incompleteness of the archive, and looked at ways to improve upon this situation (Huurdeman et al. 2015). In our research, we uncovered the *aura* of the web archive, i.e. "the web documents which were not included in the archived collection, but are known to have existed", since "references to these unarchived web documents appear in the archived pages". A high number of unarchived pages was found, almost the size of the original KB archive, potentially dramatically increasing the coverage of the archive.

[27] In Dutch, via https://www.kb.nl/webarchief/ (accessed: 29/02/16)

Due to the unique interlinked structure of the web, we were also able to create representations of unarchived content. For each of the discovered pages, we generated representations based on anchor text and URL words. These representations were generally succinct in nature, but rich enough to identify pages in a known-item search setting. Connecting this work to the lack of transparence issue pinpointed by researchers in the previous section, we may contextualize web archive search by providing information about the material which is included and excluded from the archive at retrieval time. Search results in the archive can be combined with found representations of unarchived search results, thereby increasing transparence of web archive search tools. In
recent work, we have demonstrated the possibilities for contextualization of search results by integrating our approach into prototypes of the WebARTist toolset (as presented in Figure 2).

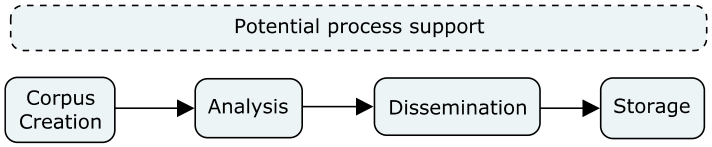


Figure 4: Phases of research, adapted from Brügger; and proposed process support.

3.5.2 Building process support

On a broader level, we propose an approach to more directly support the research process of scholars. Brügger has distinguished various research phases in web archive research[28]. These phases firstly includes *corpus creation*, during which a researcher identifies and isolates a corpus[29]. This is followed by *analysis* of the created corpus, using analytical tools and visualizations. After this phase, *dissemination* follows, which involves dissemination of the analysis, for instance in scholarly papers. Finally, *storage* is needed, which involves long-time preservation of corpora and tools.

In our view, more support for these research phases should be offered in-situ, i.e. within web archive access systems. This way, researchers can define their corpora, analyze, disseminate and store them within one system. So far, this has been achieved in specialized fields, such as bioinformatics (Zoubarev et al.

[28] Summarized in a presentation from 2015 available at: http://alexandria-project.eu/wpcontent/uploads/2015/11/2nd_alex_ws_niels_bruegger.pdf (accessed: 29/02/16)

[29] In this chapter, we define a ‘corpus’ as a set of data sources assembled by a researcher for the direct purpose of studying her research questions

2012), and genomics research (Goecks et al. 2010), where systems allowing for collaborative analysis, sharing and reuse have attracted a substantial number of researchers. Naturally, creating such a system is a profound challenge in the context of the web archive, a data source which can potentially be used in a wide variety of research settings.

The first step is to determine the functionality needed for such an approach. While the collaborative setting described in the previous section provides ample insights, we also performed an exploratory analysis based on published research papers using web data (Huurdeman 2015). This analysis showed a strong need for more fine-grained selection methods in the context of corpus building, allowing researchers to *iteratively* select material (i.e. to initially select material, and to later refine and extend these selections). Subsequently, researchers may analyze their selections. Additionally, the granularity, or analytical level (Brügger 2009) of selections are of importance, ranging from page elements, webpages, websites and web *spheres*[30] to the full web.

To achieve this aim, more supportive user interfaces are essential. This calls for additional research in the field of information seeking and retrieval as well as human-computer interaction. To avoid overly complex 'dashboard'-style interfaces, we have looked at a 'multistage' approach to search (Huurdeman and Kamps 2014; Huurdeman et al. 2016). This approach is based on existing models and theories covering different *stages* in complex searches. By understanding the needs of users in at different moments in the search process, we may provide customized functionality per information seeking stage, such as exploration and the formulation of a focus.

In various prototypes providing access to the Dutch web archive, we have explored ways to support different stages of research in distinct interfaces. In the corpus creation interface, researchers may iteratively build their complex queries on data, and store these complex queries. Subsequently, this corpus can be analyzed and annotated, before creating visualizations. In further prototypes, we add support for storage and sharing. This way, researchers are able to share the workflows which they used to derive at their final analyses. An important advantage of this approach is that analyses and derived datasets can be kept 'inside the system'. This would improve support for key rationales of research data sharing introduced by Borgman (2012), in particular reproducibility, enabling possibilities for reuse and advancing the state of research.

[30] Defined by Schneider and Foot (2004) as a "hyperlinked set of dynamically-defined digitalresources that span multiple websites and are deemed relevant, or related, to a central theme or 'object' "

Summarizing, we suggested various steps to increase transparency and to improve process support, which may transform web archive access tools from mere search tools to ‘research engines’.

3.6 Discussion and conclusions

In this chapter, we have introduced the concept of the web archive, the feasibility to use web archives as research datasets, and we described a concrete use case. This culminated in a suggested move towards ‘research engines’, overcoming limitations of access interfaces. The concrete use case of the web archive has shown that the choices made during the creation of these archives have a profound influence on the research which can performed with them. Different actors and policies profoundly influence the nature of the data ending up in the archive. On a technical level, intricate dependencies exist between the harvesting settings and capabilities, and the quality and quantity of the captured content. In effect, performing research using one of these archives is no easy task, let alone performing cross-comparison analysis using different archives; an issue only corroborated by the scattered nature of archives across institutions and countries. This issue also applies, in some ways more gravely, to research datasets available in research data repositories. A continuum exist from loosely to highly structured research data, and properties of this data differ per research domain. For instance, as Wilson et al. (2010) have indicated, “certain characteristics of humanities data make its re-use particularly difficult. The data is often messy and incomplete, being derived from diverse sources that were never intended to provide information in a regular or comparable format”. Different researchers may create different research datasets, containing various assumptions and limitations. Hence, it can be hard to reuse datasets, and to do comparisons and further analysis, hindering the possibility to “ask new questions to extant data” (Borgman 2012). This leads us to conclude that transparency in these issues is a key prerequisite, and *documentation* is essential. However, keeping documentation during the research process is arguably a time-consuming task.

Besides transparence, we identified various other limitations of web archives in a scholarly setting based on the concrete case of the WebART project, such as a lack of support for selection methods, analysis and ‘collection making’. Hence, there is a need for increased *process support* for researchers. We emphasized that future systems should support researchers in the creation of corpora, analysis and storage of derived results. In effect, the system could allow researchers to store and annotate corpora and derived datasets inside the system, instead of taking them ‘out’, but also enable researchers to keep track of their workflows. This is a double-edged sword: supporting *in-situ* analysis may come as a solution for various *transparence* issues stated previously, and

the automatic tracking of workflows may may taper the requirements for elaborate manual documentation of the research process.

Hence, we propose a more collaborative approach between institutions archiving research data and researchers, theoretically supporting scholars in their entire workflow, from corpus creation to dissemination and storage. At the moment, this may seem an ambitious view, but the emergence and popularity of web platforms for supporting collaborative research, for example in bioinformatics and genomics research (Zoubarev et al. 2012; Goecks et al. 2010), has proved that this is no distant dream. To allow for these types of functionalities in more generalized research data repositories is certainly no trivial task, but may constitue a fruitful point of embarkment for future inquiry into research data management support.

Acknowledgments

The authors wish to thank all collaborators in the WebART project, as well as the reviewers of this book chapter. This research was supported by the Netherlands Organization for Scientific Research (NWO, WebART project, # 640.005.001).

References

Ainsworth, S., Alsum, A., SalahEldeen, H., Weigle, M.C. and Nelson, M.L. 2011. "How much of the web is archived?" *Proceedings of the 11th annual international ACM/IEEE joint conference on digital libraries*:133-136.

Ben-David, A. and Huurdeman, H. C. 2014. Web archive search as research: Methodological and theoretical implications. *Alexandria* 25:1.

Borgman, C. L. 2012. The conundrum of sharing research data. *J Am Soc Inf Sci Tec* 63(6):1059–1078.

Brown, A. 2006. *Archiving websites: a practical guide for information management professionals*. Facet Publishing, London.

Brügger, N. 2009. "Website history and the website as an object of study." *New Media & Society*, 11 (115).

Brügger, N. 2005. *Archiving Websites - General Considerations and Strategies*. Århus:The Centre for Internet Research.

Brügger, N. 2011. "Web Archiving – Between Past, Present, and Future". In *The Handbook of Internet Studies*, edited by Mia Consalvo and Charles Ess:24–42. Chichester:Blackwell Publishing Ltd.

Dougherty, M. and Meyer, E.T. 2014. *Community, tools, and practices in web archiving: The state-ofthe-art in relation to social science and humanities research needs*. J Am Soc Inf Sci Tec:65(11):2195–2209.

Goecks J., Nekrutenko A., Taylor J. and Galaxy Team 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11(8):R86.

Helmond, A. 2015. “The web as platform: Data flows in social media.” PhD diss., University of Amsterdam.

Hockx-Yu, H. 2014. “Access and Scholarly Use of Web Archives”. *Alexandria* 25(1-2):113–127.

Huurdeman, H.C. 2015. “Towards Research Engines: Supporting Search Stages in Web archives”. Paper presented at Web Archives as Scholarly Sources: Issues, Practices and Perspectives conference, Aarhus University, Aarhus, June 8–10.

Huurdeman, H.C. and Kamps, J. 2014. From multistage information-seeking models to multistage search systems. In *Proceedings of the 5th Information Interaction in Context Symposium, IIiX ’14*:145–154.

Huurdeman, H.C., Ben-David, A. and Sammar, T. 2013. Sprint methods for web archive research. In *Proceedings of the 5th Annual ACM Web Science Conference, WebSci ’13*:182–190.

Huurdeman, H.C., Kamps, J., Samar, T., de Vries, A.P., Ben-David, A. and Rogers, R.A. 2015. Lost but not forgotten: finding pages on the unarchived web. *Int J Digit Libr* 16: 247–265.

Huurdeman, H.C., Wilson, M. and Kamps, J. Active and passive utility of search interface features in different information seeking task stages. In *Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016*:3-12.

Masanés, J. 2006. “Web Archiving”. In *Web Archiving: Issues and Methods*, edited by J. Masanés, 1–53. Berlin, Heidelberg:Springer.

Moretti, F. 2013. *Distant reading*. New York:Verso Books,.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., and Lioma, C. 2006. Terrier: A High Performance and Scalable Information Retrieval Platform. *Proceedings SIGIR’06 Workshop on Open Source Information Retrieval, OSIR 2006*.

Pickard, A. Research methods in information. London: Facet Publishing, 2007.

Rogers, R.A. 2013. Digital methods. Cambridge:MIT Press.

Schneider, S.M. and Foot, K.A. 2004. “The web as an object of study.” *New media & society* 6(1): 114–122.

Thomas, A., Meyer, E., Dougherty, M., Van den Heuvel, C., Madsen, C. and Wyatt, S. 2010. “Researcher engagement with web archives: Challenges and opportunities for investment.” *Technical report*, London:JISC.

Wilson, J. A. J., Fraser, M. A., Martinez-Uribe, L., Jeffreys, P., Patrick, M., Akram, A., and Mansoori, T. 2010. “Developing Infrastructure for Research Data Management at the University of Oxford.” *Ariadne* 65.

Zoubarev, A., Hamer, K.M., Keshav, K.D., McCarthy, E.L., Santos, J.R.C., Rossum, T.V., McDonald, C., Hall, A., Wan, X., Lim, R., Gillis, J., and Pavlidis, P. 2012. “Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data.” *Bioinformatics* 28(17):2272–2273.