Finding Pages on the Unarchived Web

Hugo C. Huurdeman[1] Anat Ben-David[1] Jaap Kamps[1] Thaer Samar[2] Arjen P. de Vries[2]
[1]University of Amsterdam, Amsterdam, The Netherlands
[2]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
{huurdeman|a.ben-david|kamps}@uva.nl {samar|arjen}@cwi.nl

ABSTRACT

Web archives preserve the fast changing Web, yet are highly incomplete due to crawling restrictions, crawling depth and frequency, or restrictive selection policies—most of the Web is unarchived and therefore lost to posterity. In this paper, we propose an approach to recover significant parts of the unarchived Web, by reconstructing descriptions of these pages based on links and anchors in the set of crawled pages, and experiment with this approach on the Dutch Web archive.

Our main findings are threefold. First, the crawled Web contains evidence of a remarkable number of unarchived pages and websites, potentially dramatically increasing the coverage of the Web archive. Second, the link and anchor descriptions have a highly skewed distribution: popular pages such as home pages have more terms, but the richness tapers off quickly. Third, the succinct representation is generally rich enough to uniquely identify pages on the unarchived Web: in a known-item search setting we can retrieve these pages within the first ranks on average.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process, Selection process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation (efficiency and effectiveness)*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*collection*

General Terms

Experimentation, Measurement, Performance

Keywords

Web archives, Web archiving, Web crawlers, Anchor text, Link evidence, Information retrieval

X-XXXXX-XX-X/XX/XX.

1. INTRODUCTION

The advent of the Web has had a revolutionary impact on how we acquire, share and publish information. The vast amount of digital born content is rapidly taking over other forms of publishing, and the overwhelming majority of online publications has no parallel in a material format. Memory and heritage institutions increasingly recognize that such digital born data are as easily deleted as they are published, thereby introducing unprecedented risks to the world's digital cultural heritage [27]. Web archives address this problem by systematically preserving parts of the Web for future generations. It involves a "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" [12]. Pioneered by the Internet Archive and later joined by many national libraries, Web archiving initiatives have archived petabytes of Web data. Despite the important attempts to preserve parts of the Web by archiving, a large part of the Web's content is unarchived and hence lost forever. It is impossible to archive the entire Web due to its ever increasing size and rapidly changing content. However, even the parts that have been preserved are incomplete at several levels.

There are two basic strategies for Web archiving, performed by Web crawlers. The first strategy focuses on the automatic harvesting of websites in large quantities (usually a national domain), also known as 'breadth-first crawls'. The second strategy is based on a specific selection policy, where the crawler settings are intended to ensure the complete preservation of specific websites, also known as 'deep crawls' [6, 11, 19]. On the one hand, consider a breadth-first crawl intended to harvest a top-level domain of a country such as the Netherlands. Being the fifth largest top-level domain in terms of registered domains [3], such a crawl may take several months to complete. Additionally, since its settings are designed to discover as many new links as possible, the crawl may not preserve all internal pages within hosts. On the other hand, selective archives might capture more deep levels of harvested websites, since they are focused on crawling specific websites. However, a large degree of linked pages will not be preserved, since the applied crawler settings typically exclude encountered links outside the seed list, even if relevant to a country's cultural heritage.

The overall consequence is that our Web archives are highly incomplete, and researchers and other users treating the archive to reflect the Web as it once was, may draw false conclusions due to unarchived content. The main research question of this paper is: can we recover parts of the unar-

chived Web? This may seem like a daunting challenge or a mission impossible: how can we go back in time and recover pages that were never preserved? Our approach is to exploit the hyperlinked structure of the Web, and collect evidence of uncrawled pages from the pages that were crawled and are part of the archive. We show empirically that it is possible to recover significant parts of the unarchived Web, by reconstructing descriptions of these pages based on links and anchors in the crawled pages. We refer to the recovered Web documents as the Web archive's *aura*: the Web documents which were not included in the archived collection, but are known to have existed—references to these unarchived Web documents appear in the archived pages.

Specifically the paper is investigating the following research questions:

RQ1 Can we recover a significant fraction of unarchived pages and hostnames from references to them in the Web archive?

We exploit the link structure of the crawled content to derive evidence of the existence of unarchived pages, and investigate their number of pages and of domains or hostnames.

RQ2 How rich are the representations that can be created for unarchived URLs?

We build implicit representations of unarchived Web pages and domains, based on link evidence and anchor text, and investigate the richness (or sparseness) of the descriptions in the number of incoming links and the aggregated anchor text, and break this down over unarchived home pages and other pages.

RQ3 Are the resulting derived representations of unarchived pages useful in practice? Do they capture enough of the unique page content to make them retrievable amongst millions of other pages?

As a critical test, we study the effectiveness of the derived representations of unarchived home pages and deep pages in a known-item search setting. Only if the derived representation characterizes the unique page's content, we have a chance to retrieve the page within the first ranks.

The remainder of the paper is organized as follows: we first introduce related work (Section 2), followed by a description of the experimental setup (Section 3). Next, we look at the results of our analysis, characterizing the actual contents of the Dutch Web archive and the *aura* of unarchived pages around the archive (Section 4). Furthermore, we look into the potential richness of generated representations (Section 5). The generated representations are evaluated using known-item search topics (Section 6). We end by discussing our findings and drawing conclusions (Section 7).

2. BACKGROUND

In this section, we discuss related work, which falls in two broad areas. First, we discuss related research in Web archiving and Web preservation. Second, we discuss previous work in search based on link evidence and anchor text.

2.1 Web Archives and Web Preservation

Experts in the Web archiving community discuss the shortcomings of Web archiving crawlers in terms of the content they fail to capture [19]. Some websites are intentionally excluded, breadth-first crawls might not capture deeper pages of a website, and selective crawlers exclude sites beyond the scope of the selection policy. However, as argued by Day [6], in most cases, even the sites that meet selection guidelines on other criteria may include errors, be incomplete or have broken links. Moreover, Web archiving crawlers often times fail to capture specific content elements such as JavaScript, Flash, and database-driven sites [6, 11, 19]. This prompts Web historian Brügger [2] to argue that almost every Web archive is incomplete to the extent that it is hard to determine what is missing.

The limits of Web archives' crawlers may result in partial and incomplete Web archives. However, crawlers do encounter and register additional information about a page they encounter, such as its outlinks, anchor text, and crawl and page timestamps. Rauber et al. [24] have recognized the wealth of additional information contained in Web archives which can be used for analytical purposes. Gomes and Silva [9] used data obtained from the domain crawl of the Portugese Web archive to develop criteria for characterizing the Portugese Web. More recently, researchers from the LiWA project have developed a prototype for an analytical user interface designed to use these elements for analyzing large scale Web archives [26]. The Memento project has expanded the scope of analysis of archived web data beyond the boundaries of a single archive, in order to profile and analyze coverage of archived websites across different web archives. Memento [28] is an HTTP-based framework which makes it possible to locate past versions of a given Web resource through an aggregator of resources from multiple Web archives. In a recent study, Alsum et al. [1] queried the Memento aggregator to profile and evaluate the coverage of twelve public Web archives. They found that the number of queries can be reduced by 75% by only sending queries to the top three Web archives. Here, coverage (i.e. whether a resource is archived and in which archive its past versions are located) was calculated based on the HTTP header of host level URLs.

We take a radically different approach and try to uncover significant parts of the unarchived Web, by not only uncovering missing (unarchived) pages but also recovering a representation of these by using URL and anchor text representations.

2.2 Link Evidence and Anchor Text

One of the defining properties of the Internet is its hyperlink-based structure. The Web's graph structure is well studied, and also methods to use this structure have widely been applied, especially in the context of Web retrieval (for example PageRank [18]). The links which weave the structure of the Web consist of destination URLs, and are described by anchor text. Aggregating anchor text of links makes it for example possible to create representations of target pages. Techniques based on the graph structure of the Web, and anchor text have widely been used in Web retrieval. In this paper, we mainly focus on the use of anchor text.

Craswell et al. [4] explored the effectiveness of anchor text in the context of site finding. Aggregated anchor texts for a link target were used as surrogate documents, instead of the actual content of the target pages. Their experimental results show that anchor texts can be more effective than con-

tent words for navigational queries (i.e. site finding). Work in this area led to advanced models that combine various representations of page content, anchor text, and link evidence [13]. Fujii [8] presented a method for classifying queries into navigational and informational. Their retrieval system used content-based or anchor-based retrieval methods, depending on the query type. Based on their experimental results, they concluded that content of webpages is useful for informational query types, while anchor text information and links are useful for navigational query types. Contrary to previous work, Koolen and Kamps [16] concluded that anchor text can also be beneficial for ad hoc informational search, and their findings show that anchor text can lead to significant improvements in retrieval effectiveness. They also analyze the factors influencing this effectiveness, such as link density and collection size. In the context of Web archiving, link evidence and anchor text could be used to locate missing webpages, of which the original URL is not accessible anymore. Klein and Nelson [14] computed lexical signatures of lost webpages, using the top n words of link anchors, and used these and other methods to retrieve alternative URLs for lost webpages.

Following Kleinberg [15], Dou et al. [7] took the relationships between source pages of anchor texts into account. Their proposed models distinguish between links from the same website and links from related sites, to better estimate the importance of anchor text. Similarly, Metzler et al. [20] smoothed the influence of anchor text which originates from within the same domain, using the 'external' anchor text: the aggregated anchor text from all pages that link to a page in the same domain as the target page. Another aspect of anchor text is its development over time: often single snapshots of sites are used to extract links and anchor text, neglecting historical trends. Dai and Davison [5] determined anchor text importance by differentiating pages' inlink context and creation rates over time. They concluded that ranking performance is improved by differentiating pages with different in-link creation rates, but they also point to the lack of available archived resources (few encountered links were actually available in the Internet Archive).

Our approach is inspired by the previous results on various Web centric document representations based on URL and incoming anchor text, typically used in addition to representations of the page's content [4, 13, 17, 21]. We focus on the use case of the Web archive, which is different from the live Web given that we cannot go back and crawl the unarchived page, hence have to rely on these implicit representations exclusively. It is an open question whether the resulting derived representations—based on scant evidence of the pages—is a rich enough characterization to be of practical use.

3. EXPERIMENTAL SETUP

This section describes our experimental setup: the approach, the dataset, the link extraction methods and the way the links were aggregated for analysis.

3.1 Data

This study uses data from the Dutch Web archive at the National Library of the Netherlands (KB). The KB currently archives a pre-selected (seed) list of more than 5,000 websites [23]. Websites for preservation are selected by the library based on categories related to Dutch historical, social and cultural heritage. Each website in the seed list has been categorized using a UNESCO classification code.

Table 1: Number of documents per year

year	number of docs
2009	17,014,067
2010	38,157,308
2011	53,604,464
2012	38,865,673
	147,641,512

Our snapshot of the Dutch Web archive consists of 76,828 ARC files, which contain aggregated Web content. A total number of 148M documents has been harvested between February 2009 and December 2012, resulting in more than 7 Terabytes of data (see Table 1). Basic harvest metadata is available (crawl dates, page modification dates, etc.). Additional metadata is available in separate documentation, which includes the KB's selection list, dates of selection and the manually assigned UNESCO codes by the curators of the KB. In our study, we focus on the documents crawled in 2012.

In our extraction, we differentiate between four different types of URLs found in the Dutch Web archive:

1. URLs that have been archived intentionally as they are included in the seedlist,
2. URLs that have been unintentionally archived due to the crawler's configuration,
3. unarchived URLs, of which the parent domain is included in the seedlist, and
4. unarchived URLs, which do not have a parent domain that is on the seedlist.

3.2 Link Extraction

We created our dataset by implementing a specific processing pipeline. This pipeline uses Hadoop MapReduce and Apache Pig for data extraction and processing. The first MapReduce job processed all archived webpages contained in the archive's ARC files, and used JSoup to extract links from their contents. For each link, the source URL, target URL, crawldate, anchor text and (MD5) hashcode of the source page were kept. Subsequently, this file was matched against the KB's list of seed domains and assigned UNESCO codes, to create a set with an indication if a specific URL is on the seedlist at the moment of crawling, and if it has a UNESCO classification code. A second MapReduce job built a temporary index of all URLs (with their associated crawldate) that occur in the Dutch Web archive, allowing lookups to validate if a given URL exists in the archive or not. Subsequently, the processed files have been joined to create the following list:

(sourceURL, sourceUnesco, sourceInSeedProperty, targetURL, targetUnesco, targetInSeedProperty, anchorText, crawlDate, targetInArchiveProperty, sourceHash)

In our study, we look at the content per year. Therefore, additional steps in our data preparation included deduplication of links per year, to correct for different harvesting frequencies of sites in the archive. While some sites are harvested yearly, other sites are captured biannually, quarterly

or even daily. This could result in a large number of links from duplicate pages. To prevent this from influencing our dataset, we deduplicated the links based on their values for year, anchor text, source, target, and (MD5) hashcode. The hashcode is a unique value representing a page's content, and is used to detect if a source has changed between crawls. We keep only links to the same target URLs if it originates from a unique source URL.

In our dataset, we include both inter-server links, which are links between different servers (external links), and intra-server links, which occur within a server (site internal links). We also performed basic data cleaning and processing: removing non-alphanumerical characters from the anchor text, converting the source and target URLs to the canonicalized SURTURL format, removing double and trailing slashes, and removing *http(s)* prefixes (see `http://crawler.archive.org/apidocs/org/archive/util/SURT.html`).

3.3 Link Aggregation

Our next step consisted of aggregating the extracted links by target URL, retaining the captured metadata. In this process, we create a representation that includes the target URL and properties, and grouped data elements with source URLs, anchor texts and other associated properties. Using another Apache Pig script, we counted different elements, for example the unique source sites and hosts, unique anchor words, and the number of links from seed and non-seed source URLs. We also split each URL to obtain separate fields for TLD, domain, host and filetype. To retrieve correct values for the TLD field, we matched the TLD extension from the URL with a list of all TLDs, while we matched extracted filetype extensions of each URL with a list of common Web file formats.

This aggregated representation containing target URLs, source properties and value counts was subsequently inserted into a MySQL database (13M rows), to provide easier access for analysis.

4. EXPANDING THE WEB ARCHIVE

In this section, we study RQ1: Can we recover a significant fraction of unarchived pages and hostnames from references to them in the Web archive? We investigate the contents of the Dutch Web archive and quantify the unarchived material that can be uncovered via the archive. Our finding is that the crawled Web contains evidence of a remarkable number of unarchived pages and websites, potentially dramatically increasing the coverage of the Web archive.

4.1 Archived Content

We begin by introducing the actual archived content of the Dutch Web archive in 2012, before characterizing the unarchived contents in the next subsection. Here, we look at the unique text-based webpages (based on MD5 hash) in the archive, totaling in 11,041,113 pages. Of these pages, 10,158,586 were crawled in 2012 as part of the KB's seedlist (92%). An additional 882,527 pages are not in the seedlist but included in the archive (see Table 2). As discussed in section 2, each 'deep' crawl of a website included in the seedlist also results in additional ('out of scope') material being harvested, due to crawler settings. For example, to correctly include all embedded elements of a certain page, the crawler might need to harvest pages beyond the predefined seed domains. These unintentionally archived contents amount to 8% of the full Web archive in 2012.

Table 2: Unique archived pages (2012)

	on seedlist	%	not on seedlist	%	total
pages	10,158,586	92.0	882,527	8.0	11,041,113

Table 3: Unique archived hosts, domains & TLDs

	on seedlist	%	not on seedlist	%	total
hosts	6,157	14.2	37,166	85.8	43,323
domains	3,413	10.1	30,367	89.9	33,780
TLDs[1]	16	8.8	181	100	181

Table 4: Coverage in archive

mean page count	on seedlist	not on seedlist
per host	1,650	24
per domain	2,976	29
per TLD	634,912	4,876

We can take a closer look at the contents of the archive by calculating the diversity of hosts, domains and TLDs contained in it. Table 3 summarizes these numbers, in which the selection-based policy of the Dutch KB is reflected. The number of hosts and domains is indicative of the 3,876 selected websites on the seedlist in the beginning of 2012: there are 6,157 unique hosts (e.g. *papierenman.blogspot.com*) and 3,413 unique domains (e.g. *okkn.nl*).

The unintentionally archived items reflect a much larger variety of hostnames and domains than the items from the seedlist, accounting for 37,166 unique hosts (85.8%), and 30,367 unique domains (89.9% of all domains). The higher diversity of the non-seedlist items also results in a lower coverage in terms of number of archived pages per domain and per host (see Table 4). The mean number of pages per domain is 2,976 for the sites included in the seedlist, while the average number of pages for the items outside of the seedlist is only 29.

According to the KB's selection policies, sites that have value for Dutch cultural heritage are included in the archive. A more precise indication of the categories of websites on the seedlist can be obtained by looking at their assigned UNESCO classification codes. In the archive, the main categories are Art and Architecture (1.3M harvested pages), History and Biography (1.2M pages) and Law and Government Administration (0.9M pages). For the sites harvested outside of the selection lists, no UNESCO codes have been assigned. A manual inspection of the top 10 domains in this category (35% of all unintentionally harvested pages) shows that these are heterogeneous: 3 sites are related to Dutch cultural heritage, 2 are international social networks, 2 sites are related to the European Commission and 3 are various other international sites.

4.2 Unarchived Content

To uncover the unarchived material, we used the link evidence and structure of crawled contents of the Dutch Web archive. We refer to these contents as the Web archive's

[1]Since the values for the TLDs overlap for both categories, percentages add up to more than 100% (same for Table 6).

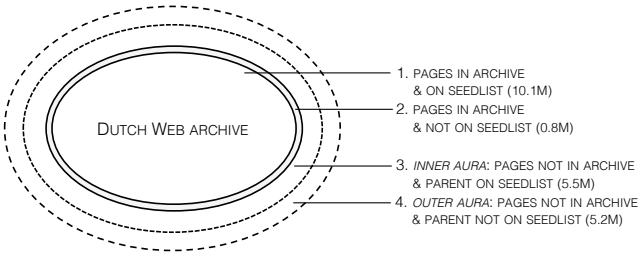

Figure 1: 'Layers' of contents of the Dutch Web Archive (2012)

Table 5: Unarchived *aura* unique pages (2012)

	inner aura	%	outer aura	%	Total
pages	5,505,975	51.5	5,191,515	48.5	10,697,490

Table 6: Unarchived unique hosts, domains & TLDs

	inner aura	%	outer aura	%	total
hosts	9,039	1.8	481,797	98.2	490,836
domains	3,019	0.8	369,721	99.2	372,740
TLDs	17	6.6	259	100	259

aura: the pages that are not in the archive, but which existence can be derived from evidence in the archive.

The unarchived *aura* has a substantial size: there are 11M unique pages in the archive, but we have evidence of 10.7M additional link targets that are not in the archive. In the following sections, we will focus on this aura, and differentiate between the *inner aura* (unarchived pages of which the parent domain is on the seedlist) and the *outer aura* (unarchived pages of which the parent domain is not on the seedlist). The inner aura has 5.5M (51.5%) unique link targets, while the outer aura has 5.2M (48.5%) unique target pages (see Figure 1 and Table 5).

Like the number of pages, also the number of unique unarchived hosts is quite substantial: while *in* the archive there are 43,323 unique hosts, we can reveal a total number of 490,836 hosts in the unarchived aura. There is also a considerable number of unique domains and TLDs in the unarchived contents (see Table 6).

The tables above also show the difference between the *inner* and *outer* aura. The outer aura has a much larger variety of hosts, domains and TLDs compared to the inner aura (Table 6). On the other hand, the coverage in terms of the mean number of pages per host, domain and TLD is much greater in the inner aura than the outer aura (see Table 7). This can be explained by the fact that the pages in the inner aura are closely related to the smaller set included in Web archive's seedlist, since they have a parent domain which is on the seedlist.

Finally, to get an overview of the nature of the unarchived resources, we have matched the link targets with a list of common Web file extensions. From this data, we can derive that the majority of references to the unarchived aura points to textual Web content. Table 8 shows the filetype distribution: the majority consists of URLs without an extension

Table 7: Unarchived *aura* coverage (2012)

mean page count	inner aura	outer aura
per host	609	10
per domain	1,823	14
per TLD	323,881	20,044

Table 8: Unarchived *aura* filetypes

inner aura	count	%	outer aura	count	%
http	4,281,750	77.77	http	3,721,059	71.68
html	351,940	6.39	php	585,024	11.27
php	321,095	5.83	html	582,043	11.21
asp	38,0964	6.92	asp	181,963	3.51
pdf	70,371	1.28	jpg	30,205	0.58

Table 9: TLD distribution

inner aura	count	%	outer aura	count	%
1 nl	5,268,772	95.7	1 com	1,803,106	34.7
2 com	130,465	2.4	2 nl	1,613,739	31.1
3 org	52,309	1.0	3 jp	941,045	18.1
4 net	44,348	0.8	4 org	243,947	4.7
5 int	8,127	0.2	5 net	99,378	1.9
6 *other*	1,954	<0.1	6 eu	80,417	1.6
			7 uk	58,228	1.1
			8 de	44,564	0.9
			9 be	43,609	0.8
			10 edu	29,958	0.6

(http), html, asp and php pages for both the inner and outer aura. Only a minority of references are other formats, like pdfs and non-textual contents (e.g. jpg files in the outer aura).

4.3 Characterizing the "Aura"

Here, we characterize unarchived contents of the archive based on the top-level domain distribution and the domain coverage.

From the top-level domains (TLDs) we derive the origins of the unarchived pages surrounding the Dutch Web archive. Table 9 shows that the majority of unarchived pages in the inner aura (95.69%) have Dutch origins. The degree of .nl domains in the outer aura is lower, albeit still considerable, with 31.08% of all 1.8M pages. The distribution of TLDs in the outer aura seems to resemble the TLD distribution of the open Web. Even though the regional focus of the selection policy of the Dutch Web archive is apparent in the distribution of the top 10, the comparison does provide indications that the outer aura is more comparable to the full Web. The prominence of the .jp TLD can be explained by the fact that some Japanese social networks are included in the unintentionally harvested pages of the Dutch archive.

Another way to characterize the unarchived contents of the Dutch Web is by studying the distribution of the target domain names. This distribution is quite distinct in the two subsets of the aura: while the inner aura contains many specific Dutch sites, as selected by the KB (e.g. *noordhollandsarchief.nl* and *archievenwo2.nl*), the outer aura contains a much more varied selection of sites, which include both popular international and Dutch sites (e.g. *facebook.com* and *hyves.nl*), and very specific Dutch sites potentially re-

Table 10: Coverage of most popular Dutch sites (*Alexa position*)

inner aura	count	outer aura	count
nu.nl (*6*)	74.2K	twitter.com (*9*)	266.7K
wikipedia.org (*8*)	17.4K	facebook.com (*3*)	227.0K
blogspot.com (*15*)	3.5K	linkedin.com (7)	184.9K
kvk.nl (*90*)	2.2K	hyves.nl (*11*)	125.6K
anwb.nl (*83*)	1.7K	google.com (*2*)	106.4K

lated to Dutch heritage (e.g *badmintoncentraal.nl*).

To get more insights into the degree of popular sites in the unarchived aura, we compare the domains occurring in the aura against publicly available statistics of websites' popularity. Alexa, a provider of free Web metrics, publishes online lists of the top 500 ranking sites per country, on the basis of traffic information. Via the Internet Archive, we retrieved a contemporary Alexa top 500 list for sites in the Netherlands (specifically, `http://web.archive.org/web/20110923151640/alexa.com/topsites/countries/NL`). We counted the number of sites in Alexa's top 100 that occur in the inner and outer aura of the Dutch archive (summarized in Table 10). The inner aura covers 7 sites of the top 100 Alexa sites (including Dutch news aggregator *nu.nl* and *wikipedia.org*), while the outer aura covers as much as 90 of the top 100 Alexa sites, with a considerable number of unique target pages. For these 90 sites, we have in total 1,227,690 URL references, which is 23.65% of all unarchived URLs in the outer aura of the archive. This means that we have potentially many representations of the most popular websites in the Netherlands, even though they have not been captured in the selection-based archive itself.

Summarizing, in this section we have quantified the size and diversity of the unarchived sites surrounding the selection-based Dutch Web archive. We found it to be substantial, with almost as many references to unarchived URLs as pages in the archive. These sites complement the sites collected based on the selection policies, and provide context from the Web at large, including the most popular sites in the country. The answer to our first research question is resoundingly positive: the indirect evidence of lost Web pages holds the potential to significantly expand the coverage of the Web archive. However, the resulting Web page representations are different in nature from the usual representations based on Web page content. We will characterize the Web page representations based on derived descriptions in the next section.

5. REPRESENTATIONS OF UNARCHIVED CONTENT

In this section, we study RQ2: How rich are the representations that can be created for unarchived URLs? We build implicit representations of unarchived Web pages and domains, based on link evidence and anchor text, and investigate the richness (or sparseness) of the resulting descriptions in the number of incoming links and the aggregated anchor text, and break this down over unarchived home pages and other pages. Our finding is that the link and anchor descriptions have a highly skewed distribution: popular pages such as home pages have more terms, but the richness tapers off quickly.

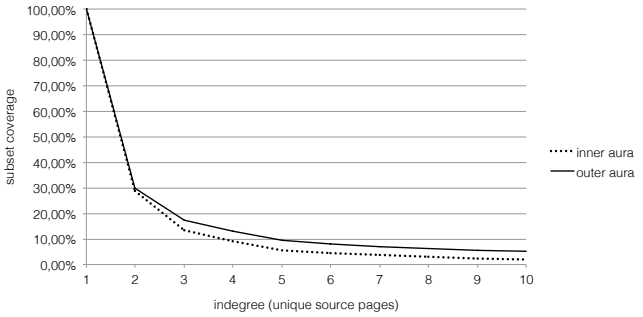

Figure 2: Number of unique source pages (based on MD5 hash) compared to subset coverage

Table 11: Link types

	inner aura	%	outer aura	%
intra-server	5,198,479	94.4	2,065,186	39.8
inter-server	289,412	5.3	3,098,399	59.7
inter & intra-server	18,084	0.4	27,930	0.5

5.1 Indegree

In general, the representation of a target page is richer if it includes anchor text contributed from a wider range of source sites, i.e. has a higher indegree. Therefore, we looked at the number of incoming links for each target URL in our uncovered archive. This is shown in Figure 2, which shows a highly skewed distribution: all target representations in the outer aura have at least 1 source link, 18% of the collection of target URLs has at least 3 incoming links, and 10% has 5 links or more. The pages in the inner aura have a lower number of incoming links than the pages in the outer aura. To check whether this is related to a higher number of intra-server (internal site) links, we also assessed the types of incoming links.

We differentiate between two link types that can be extracted from archived Web content: intra-server links, pointing to the pages in the same domain of a site, and inter-server links, that point to other websites. Table 11 shows the distribution of these types of links of the uncovered aura. It shows that the inner aura has a majority of links from the same source server (i.e. a site on the seedlist), while the outer aura has a much smaller degree of intra-server links. There are very few link targets with both intra-server and inter-server link sources in the inner and outer aura.

5.2 Anchor Text Representations

An influence on the utility of possible representations of sites is also richness of the anchor text. In the aggregated anchor text representations, we counted the number of unique words in the anchor text. Figure 3 shows the number of unique words compared to subset coverage. Like the previous distribution of incoming source links, the distribution of unique anchor text is rather skewed. While 95% of all target URLs in the archive have at least 1 word describing them, 30% have at least 3 words as a combined description, and around 3% have 10 words or more (though still amounting to 322,245 unique pages). The number of unique words per target is similar for both the inner and outer aura.

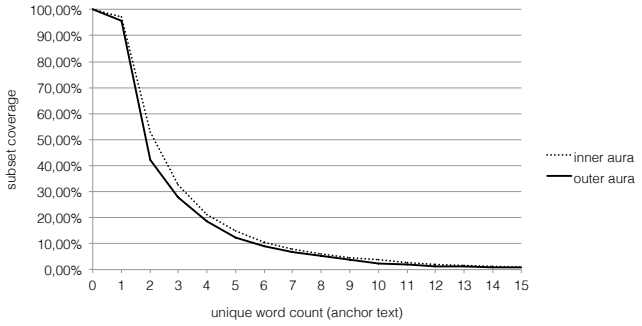

Figure 3: Number of unique words compared to subset coverage

Table 12: Target structure distribution

slashcnt	inner aura	%	slashcnt	outer aura	%
0	3,765	0.1	0	324,782	6.3
1	373,070	6.8	1	921,719	17.8
2	587,416	10.7	**2**	**1,543,129**	**29.7**
3	662,573	12.0	3	535,293	10.3
4	**1,098,947**	**20.0**	4	417,361	8.1
5	535,564	9.7	5	284,237	5.5

5.3 Homepage Representations

As mentioned in section 2.2, anchors have been used for homepage finding, since links often refer to homepages. To verify to what extent our dataset contains homepages, we looked at whether a homepage is available for each captured host in the outer aura. We calculated this number by counting the slashes in the target URLs, keeping the pages with a slashcount of 0, and by creating a set of manual filters for homepages (e.g. URLs that contain 'index.html') for pages with slashcount higher than 0. The results of this analysis indicate that for a total of 481,797 hosts, actually 336,387 homepages are available. In other words, 69.8% of all hosts have their (likely) homepage captured in our dataset. This can be important from a preservation and research perspective, since homepages are essential elements of websites, but also for the representations that we can generate from the link evidence, because homepages often have a higher indegree and more available anchor text.

To obtain a better view of the distribution of pages at different site depths, we also looked at the slashcount of the absolute URLs (see Table 12). From this analysis, we can see that the pages in the outer aura are mainly located at the first levels of the site (i.e. homepage to third level). The links towards the inner aura are pointing to pages that are deeper in the hierarchy, probably because 94% of this subset consists of intra-site link targets (links within a site).

5.4 Qualitative Analysis

Finally, we provide some examples of representations that we can create for target URLs in this dataset. We start with a homepage with a high indegree from our evaluation sample: *vakcentrum.nl*, a Dutch site for independent professionals in the retail sector. It has 142 inlinks from 6 unique hosts (6 different anchor text strings), resulting in 14 unique words. In Table 13 (A) 9 of the unique words (excluding stopwords) are displayed. They provide a basic understanding of what the site is about: a branch organization for independent retailers in the food sector.

Table 13: Sample aggregated anchor text words

(A) vakcentrum [domain]	(B) nesomexico [non-domain]
vakcentrum.nl (6)	mexico (3)
detailhandel (2)	government (1)
zelfstandige (2)	overheid (1)
ondernemers (2)	mexican (1)
levensmiddelen (2)	mexicaanse (1)
brancheorganisatie (1)	beurzen (1)
httpwwwvakcentrumnl (1)	nesomexico (1)
vgl (1)	scholarship (1)
vereniging (1)	programmes (1)

For other non-homepage URLs it is harder to represent their contents based on the anchor text alone. Take for example *knack.be/nieuws/boeken/blogs/benno-barnard*, a page that is not available on the live web anymore. It only has 2 anchor text words: 'Benno' and 'Barnard'. From the URL, however, we can further characterize the page: it is related to news ('nieuws'), books ('boeken') and possibly is a blog. Hence, we have discovered a 'lost' URL, of which we can get an (albeit basic) description by combining evidence. Of course, this varies for each recovered target URL[2], but based on the number of unique words in both anchor text and URL, we can get an estimate of the utility of the representation.

Other pages have a richer description, even if the source links only originate from one unique host. For example *nesomexico.org/dutch-students/study-in-mexico/study-grants-and-loans* is a page that is not available via the live web anymore (3 incomplete captures are located in the Internet Archive). The anchor text, originating from *utwente.nl* (a Dutch University website), has 10 unique words, contributed from 2 unique anchors. In Table 13 the combined anchor and URL words are shown, providing an indication of the page's content.

Summarizing, the inspection of the richness of representations of unarchived URLs indicates that the incoming links and the number of unique anchor text words have a highly skewed distribution: for few pages we have many descriptions which provide a reasonable number of anchors and unique terms, while the opposite holds true for the overwhelming majority of pages. The succinct representations of unarchived Web pages are indeed very different in nature. The answer to our second research question is mixed. Although establishing their existence is an important result in itself, this raises doubts whether the representations are rich enough to characterize the page's content. We decide to investigate this in the next section.

6. FINDING UNARCHIVED PAGES

In this section, we study RQ3: Are the resulting derived representations of unarchived pages useful in practice? Do they capture enough of the unique page content to make them retrievable amongst millions of other pages? We focus on the retrieval of unarchived Web pages based on their derived representations in a known-item search setting. Our

[2]e.g. *facebook.com/filmhuisbussum* has only few URL words and as anchor text 'facebook'

finding is that the succinct representation is generally rich enough to identify pages on the unarchived Web: in a known-item search setting we can retrieve these pages within the first ranks on average.

6.1 Evaluation Setup

To evaluate the utility of uncovered evidence of the unarchived Web, we indexed 5.19M representations that are in the *outer aura* of the unarchived Web archive contents. These representations consist of a unique assigned ID, the unarchived URL and aggregated anchor text of the pages in the outer aura. We indexed these documents using the Terrier 3.5 IR Platform [22], utilizing basic stopword filtering and Porter stemming. Three indexes were created. The first index uses only the aggregated anchor words (*anchT*). We also created a second index (*urlW*), which uses other evidence: the words contained in the URL. Non-alphanumerical characters were removed from the URLs and the remaining words of 20 characters or less were indexed. The third index consists of both aggregated anchor text and URL words (*anchTUrlW*).

To create known-item queries, a stratified sample of the dataset was taken, consisting of 500 random non-homepage URLs, and 500 random homepages. Here, we define a non-homepage URL as having a slashcount of 1 or more, and a homepage URL as having a slashcount of 0. These URLs were checked against the Internet Archive (pages archived in 2012). If no snapshot was available in the Internet Archive (for example because of a *robots.txt* exclusion), the URL was checked against the live Web. If no page evidence could be consulted, the next URL in the list was chosen, until a total of 150 queries per category was reached. The consulted pages were used by two annotators to create known-item queries. Specifically, after looking at the target page, the tab or window is closed and the topic creator writes down the query that he or she would use for refinding the target page with a standard search engine. Hence the query was based on their recollection of the page's content, and the annotators were completely unaware of the anchor text representation (derived from pages linking to the target). As it turned out, the topic creators used 5-7 words queries for both homepages and non-homepages. The set of queries by the first annotator was used for the evaluation (n=300), the set of queries by the second annotator was used to verify the results (n=100). We found that the difference between the annotators was low: the average difference in resulting MRR scores between the annotators for 100 homepage queries in all indexes was 8%, and the average difference in success rate was 3%.

Subsequently, we ran these 300 queries against the *anchT*, *urlW* and *anchTUrlW* indexes created in Terrier using its default InL2 retrieval model based on DFR, and saved the rank of our URL in the results list. To verify the utility of anchor, URL words and combined representations, we use the Mean Reciprocal Rank (MRR) for each set of queries against each respective index.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{1}{rank_i} \quad (1)$$

The MRR (1) is a statistical measure that looks at the probability of retrieving correct results. It is the average over the scores of the first correct result for each query (calculated by $\frac{1}{\text{rank}}$). We also compute the success rate at rank 10, that is, for which fraction of the topics do we actually retrieve the correct URL within the first 10 ranks.

Table 14: Mean Reciprocal Rank (MRR)

MRR	# Queries	anchT	UrlW	anchTUrlW
homepages	150	0.327	0.317	0.489
non-homepages	150	0.254	0.384	0.457
combined	300	0.290	0.351	0.473

Table 15: Success rates (target page in top 10)

Success@10	# Queries	anchT	UrlW	anchTUrlW
homepages	150	46.7%	39.3%	64.0%
non-homepages	150	34.7%	46.0%	55.3%
combined	300	40.7%	42.7%	59.7%

6.2 Availability of Pages

We used unarchived pages uncovered from the Dutch Web archive, that are either available in the Internet Archive, or still available on the live Web, in order to have the ground truth information about the page's content. This potentially introduces bias—there can be some difference between the pages that still are active, or have been archived, and those that are not—but the URLs did not suggest any striking differences. Out of all randomly chosen homepages surveyed, 79.9% were available via either the Internet Archive or the live Web. However, this was not the case for the non-homepages (randomly selected pages with a slash count of 1 or more), as only 49.8% could be retrieved via the Internet Archive or the live Web. The underlying reasons that many URLs could not be archived include restrictive robots.txt policies (e.g. Facebook pages), contents specifically excluded from the archive (e.g. Twitter accounts and tweets), but also links resulting from page scripts (e.g. LinkedIn 'share' buttons). The unavailability of URLs strengthens the potential utility of generated page representations, for example via aggregated anchor text, since no page evidence can be retrieved anymore.

6.3 MRR and Success Rate

MRR scores were calculated for the examined homepages and non-homepages to test to what extent the generated representations suffice to retrieve unarchived URLs. The final results of the evaluation based on MRR are summarized in Table 14. We found that the MRR scores for the homepages and non-homepages are quite similar, though some differences can be seen. Using the anchor text index, the homepages score higher than the non-homepages, possibly because of the richer representations available for these homepages. The scores for the URL words index are naturally higher for the non-homepages: they have longer URLs and therefore more words that could match the words used in the query. Finally, we can see that the combination of anchor and URL words evidence significantly boosts the retrieval effectiveness: the MRR is close to 0.5, meaning that in the average case the correct result is retrieved at the second rank.

We also examined the success rate, that is, for which degree of the topics do we actually retrieve the correct URL

Table 16: Division based on indegree of unique hosts

indegree	pages	word count	MRR anchT	homepage
1	251	2.9	0.29	42.6%
2	28	3.8	0.19	82.1%
3	12	4.5	0.29	100%
4+	9	7.3	0.49	88.9%

within the first 10 ranks? Table 15 shows that again there is some similarity between the homepages and non-homepages. The homepages score better using the anchor text index than the non-homepages: 46.7% can be retrieved. On the other hand, the non-homepages fare better than the homepages using the URL words: 46.0% of the non-homepages is included in the first 10 ranks. Again, we see that combining both representations results in a significant increase of the success rate: we can retrieve 64% of the homepages, and 55.3% of the non-homepages in the first 10 ranks.

The MRR scores indicate that anchor text in combination with tokenized URL words can be discriminative enough to do known-item search: the correct results can usually be retrieved within the first ranks. Secondly, the success rates show that by combining anchor text and URL word evidence, 64% of the homepages, and 55.3% of the deeper pages can be retrieved. This provides positive evidence for the utility of these representations.

The performance on the derived representations is comparable to the performance on regular representations of webpages [10]. Here we used a standard retrieval model, without including various priors tailored to the task at hand [17].

6.4 Impact of Indegree

Another aspect of the evaluation examines the influence of the number of unique inlinks on the richness of anchor text representations. For example, the Centre for European Reform (*cert.org.uk*) receives links from 3 unique hosts: *portill.nl*, *europa-nu.nl* and *media.europa-nu.nl*, together contributing 5 unique anchor words, while the page *actionaid.org/kenya* has 1 intra-server link from *actionaid.org*, contributing only 1 anchor word. For the combined 300 topics (domains and non-domains together), we calculated the mean unique word count, the MRR and the degree of homepages in the subset. Table 16 summarizes these results.

It shows that, depending on the number of inlinks from unique hosts, the mean word count rises, but it also illustrates the skewed distribution of our dataset: the majority of pages (251 out of 300) have links from only one source host, while a much smaller set (49 out of 300) have links from 2 or more unique source hosts. The table also provides evidence of the hypothesis that the homepages have more inlinks from unique hosts than non-homepages: at an indegree of 2 or more, the homepages take up more than 80% of the set of pages. We can also observe from the data that the MRR using the anchor text index in our sample is highest when having links from at least 4 unique hosts.

Summarizing, we investigated whether the derived representations characterize the unique content of unarchived webpages in a meaningful way. We conducted a critical test cast as a known-item finding task, requiring to locate unique pages amongst millions of other pages—a true needle-in-a-haystack task. The outcome is clearly positive: with MRR scores of about 0.5, we find the relevant pages at the second rank on average, and for the majority of pages the relevant page is in the top 10 results. The answer to our third research question is again positive: we can reconstruct representations of unarchived webpages that characterize their content in a meaningful way.

7. DISCUSSION AND CONCLUSIONS

In this study, we proposed a method for deriving representations for unarchived content, by using features extracted from a dataset of archived webpages. We used link evidence to firstly *uncover* target URLs outside the archive, and secondly to *reconstruct* basic representations of target URLs outside the archive. This evidence includes aggregated anchor text, source URLs, assigned classification codes, crawl dates, and other extractable properties. Hence, we derived representations of URLs that are not archived, and which otherwise would have been lost.

We tested our methods on the data of the selection-based Dutch Web archive in 2012. The analysis presented above first characterized the contents of the Dutch Web Archive, from which the representations of unarchived pages were subsequently uncovered, reconstructed and evaluated. The archive contains almost as many mentions of unarchived pages as the number of the actually archived pages. Hence, using data extracted from archived pages, information can be recovered about unarchived pages which once closely interlinked with the pages in the archive.

The recovery of the unarchived pages surrounding the Web archive, which we called the 'aura' of the archive, can be used for assessing the completeness of the archive, and may help to extend the seedlist of the crawlers of selection-based archives. Additionally, representations of pages could also be used to enrich the index and provide additional search functionalities. Including the representations of pages in the outer aura, for example, is of special interest as it contains evidence to the existence of top websites that are excluded from archiving, such as Facebook and Twitter. This is supported by the fact that only two years since the data was crawled, 20.1% of the found unarchived homepages and 45.4% of the non-home pages could no longer be found on the live Web nor the Internet Archive.

The evaluation of the unarchived pages described in this study shows that the extraction is rather robust, since both unarchived homepages and non-homepages received similar satisfactory MRR average scores. However, there are some limitations to the method described in this study. The first concerns the aggregation of links by year, which may overgeneralize timestamps of the unarchived pages and therefore decrease the accuracy of the representation. Second, the recovered representations are rather skewed, hence most of the uncovered pages have relatively sparse representations, while only a small fraction has rich representations. Third, we used data from a selective archive, whose crawler settings privilege select hostnames and are instructed to ignore other encountered sites. This affects the relative distribution of home pages and non-homepages, both in the archive as well as in the unarchived pages. In future work we will examine the impact of the crawling strategy.

Web archives preserve Web content for posterity, assuming that what is not selected for archiving might be lost forever. This study shows that it is still possible to recover representations of pages that were not selected for archiving. We have developed a method for uncovering evidence

of unarchived pages from Web archives, and for reconstructing representations of their past existence based on link and anchors in crawled pages. Our analysis of the Dutch Web archive crawled in 2012 shows that the number of unarchived pages that can be uncovered is as large as the number of the intentionally archived pages. Although the representation of the unarchived pages based on anchor text and link structure is skewed (that is, few uncovered pages have very rich representation while the representation of most pages is relatively poor), our analysis shows that anchor text and link information suffice to retrieve the unarchive pages within the first two ranks on average. Our initial results in this paper are based on straightforward descriptions of pure anchor text and URL components and standard ranking models. In follow up research we will examine the effect of including further contextual information, such as the text surrounding the anchors, and advanced retrieval models that optimally weight all different sources of evidence.

Acknowledgments

Part of this paper is based on an initial report on uncovering and characterizing unarchived pages, published as [25]. We would like to thank the anonymous reviewers for their helpful suggestions. We gratefully acknowledge the collaboration with the Dutch Web Archive of the National Library of the Netherlands. This research was supported by the Netherlands Organization for Scientific Research (WebART project, NWO CATCH # 640.005.001).

REFERENCES

[1] A. Alsum, M. C. Weigle, M. L. Nelson, and H. Van de Sompel, "Profiling web archive coverage for top-level domain and content language," in *TPDL*, ser. LNCS, T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, and C. J. Farrugia, Eds., vol. 8092. Springer, 2013, pp. 60–71.

[2] N. Brügger, "Web history and the web as a historical source," *Zeithistorische Forschungen*, vol. 9, no. 2, pp. 316–325, 2012.

[3] CENTR, "Domain wire stat report," Council of European National Top Level Registrars (CENTR), Tech. Rep., 2013.

[4] N. Craswell, D. Hawking, and S. Robertson, "Effective site finding using link anchor information," in *SIGIR*. ACM, 2001, pp. 250–257.

[5] N. Dai and B. D. Davison, "Mining anchor text trends for retrieval," in *ECIR*, ser. LNCS, C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. M. Rüger, and K. van Rijsbergen, Eds., vol. 5993. Springer, 2010, pp. 127–139.

[6] M. Day, "Preserving the fabric of our lives: A survey of web." in *ECDL*, ser. LNCS, T. Koch and I. Solvberg, Eds., vol. 2769. Springer, 2003, pp. 461–472.

[7] Z. Dou, R. Song, J.-Y. Nie, and J.-R. Wen, "Using anchor texts with their hyperlink structure for web search," in *SIGIR*, J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, Eds. ACM, 2009, pp. 227–234.

[8] A. Fujii, "Modeling anchor text and classifying queries to enhance web document retrieval," in *WWW*, J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, Eds. ACM, 2008, pp. 337–346.

[9] D. Gomes and M. J. Silva, "Characterizing a national community web," *ACM Transactions on Internet Technology (TOIT)*, vol. 5, no. 3, pp. 508–531, 2005.

[10] D. Hawking and N. Craswell, "Very large scale retrieval and web search," in *TREC: Experiment and Evaluation in Information Retrieval*, E. Voorhees and D. Harman, Eds. MIT Press, 2005, ch. 9.

[11] H. Hockx-Yu, "The past issue of the web," in *Web Science*. ACM, 2011, p. 12.

[12] "Web Archiving Why Archive the Web?" http://netpreserve.org/web-archiving/overview, 2014, accessed: 2014-03-23.

[13] J. Kamps, "Web-centric language models," in *CIKM*, O. Herzog, H.-J. Schek, N. Fuhr, A. Chowdhury, and W. Teiken, Eds. ACM, 2005, pp. 307–308.

[14] M. Klein and M. L. Nelson, "Moved but not gone: an evaluation of real-time methods for discovering replacement web pages," *Int. J. on Digital Libraries*, vol. 14, no. 1-2, pp. 17–38, 2014.

[15] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[16] M. Koolen and J. Kamps, "The importance of anchor text for ad hoc search revisited," in *SIGIR*, F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, and J. Savoy, Eds. ACM, 2010, pp. 122–129.

[17] W. Kraaij, T. Westerveld, and D. Hiemstra, "The importance of prior probabilities for entry page search," in *SIGIR*. ACM, 2002, pp. 27–34.

[18] P. Lawrence, B. Sergey, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford University, Technical Report, 1998.

[19] J. Masanès, *Web archiving*. Springer, 2006.

[20] D. Metzler, J. Novak, H. Cui, and S. Reddy, "Building enriched document representations using aggregated anchor text," in *SIGIR*. New York, NY, USA: ACM, 2009, pp. 219–226.

[21] P. Ogilvie and J. P. Callan, "Combining document representations for known-item search," in *SIGIR*, 2003, pp. 143–150.

[22] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma, "Terrier: A high performance and scalable information retrieval platform," in *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[23] M. Ras, "Eerste fase webarchivering," Koninklijke Bibliotheek, Tech. Rep., 2007.

[24] A. Rauber, R. M. Bruckner, A. Aschenbrenner, O. Witvoet, and M. Kaiser, "Uncovering information hidden in web archives: A glimpse at web analysis building on data warehouses," *D-Lib Magazine*, vol. 8, no. 12, 2002.

[25] T. Samar, H. C. Huurdeman, A. Ben-David, J. Kamps, and A. de Vries, "Uncovering the unarchived web," in *SIGIR*, ser. SIGIR '14. ACM, 2014, pp. 1199–1202.

[26] S. Soman, A. Chharjta, A. Bonomo, and A. Paepcke, "Arcspread for analyzing web archives," Stanford InfoLab, Tech. Rep., 2012.

[27] UNESCO, "Charter on the preservation of digital heritage (article 3.4)," 2003.

[28] H. Van de Sompel, M. Nelson, and R. Sanderson, "RFC 7089 - HTTP framework for time-based access to resource states - Memento," Internet Engineering Task Force (IETF), RFC, 2013.