Accessing Web Archives from Different Perspectives with Potential Synergies*

Helge Holzmann and Thomas Risse

L3S Research Center
Appelstr. 9a, 30167 Hannover, Germany
{holzmann,risse}@L3S.de

Abstract. Web archives constitute a valuable source for researchers from many disciplines. However, their sheer size, the typically broad scope and their temporal dimension make them difficult to work with. We have identified three approaches to access and explore Web archives from different perspectives: *user-*, *data-* and *graph-centric*. In this paper we present related works on these three views as well as discuss their relations and potential synergies. Finally, we propose a generic analysis schema that outlines a systematic way to study Web archives by approaching them from different zoom levels corresponding to the three presented views.

Keywords: Web Archives; Big Data Processing; Temporal Information Retrieval

1 Introduction

Web archives constitute a valuable source for research in many disciplines, including *Digital Humanities*, historical sciences and journalism, by offering a unique possibility to look into past events and their representation on the Web. They are typically big in size and have a very broad scope, often targeting at national subsets of the Web, or, in case of the *Internet Archive*[1], the entire Web.

The natural way to look at the information in a Web archive is through a Web browser, just like users do on the live Web. This is what we consider the **user-centric view**. The most common way to access a Web archive from a user's perspective is the *Wayback Machine*[2], the Internet Archive's replay tool to render archived webpages. Archived pages are identified by their URL and a timestamp, referring to a particular version of a page. To facilitate the discovery of an archived resource if the URL is unknown, different approaches to search Web archives by keywords have been proposed [1, 2, 3, 4]. Another way for users to find and access archived pages is by linking past information on the current Web to the corresponding evidence in a Web archive [5].

* This work is partly funded by the European Research Council under ALEXANDRIA (ERC 339233)

[1] `http://archive.org`

[2] `http://web.archive.org`

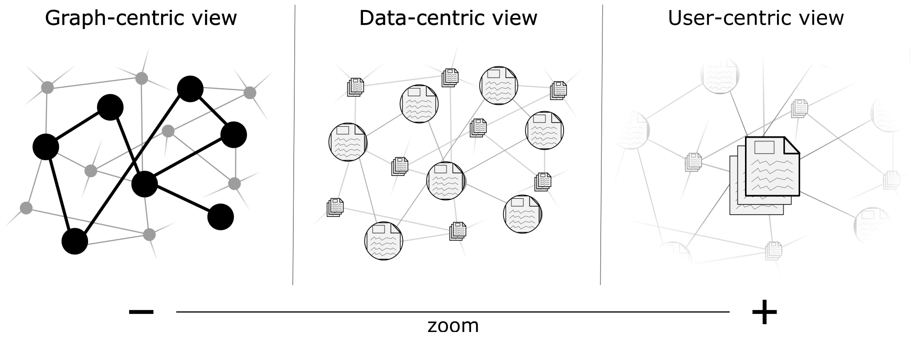

Fig. 1: Three views on Web archives, representing different zoom-levels to look at the archived data.

In contrast to accessing Web archives by closely reading pages, like users do, archived contents can also be processed at scale, enabling evolution studies and big data analysis. In this **data-centric view**, webpages are not necessarily considered self-contained units with a layout and embeds, but single resources are treated as raw data, such as text or images. A question like *"What persons appear together most frequently in a specific period of time?"* is only one example of what can be analyzed from the archived Web. Typically this is not done on a whole archive, but only on pages from a specific time period as well as on specific data types or other facets that need to be filtered first. With ArchiveSpark we have developed a tool for building research corpora from Web archives that operates on standard formats and facilitates the process of filtering as well as data extraction and derivation at scale in a very efficient manner [6].

The third perspective, besides the *user-centric* and *data-centric views*, is what we call the **graph-centric view**. Here, single pages or websites, consisting of multiple pages, are considered nodes in a graph, without taking their contents into account. Links among pages are represented by edges between the nodes in such a graph. This structural perspective enables completely different kinds of analysis, like centrality computations, with algorithms such as *PageRank*.

We present the latest achievements from all three views as well as synergies among them. For instance, important websites that can be identified from the *graph-centric* perspective may be of particular interest to the users of a Web archive. Furthermore, the *user-centric view* is often just a starting point for a much more comprehensive data study. Hence, those views can be considered as different zoom levels to look at the same Web archive data from different perspectives as illustrated in Figure 1. In the remainder of this paper we will begin to look at Web archives from the user's perspective and zoom out to the data- and graph-centric views, before we finally discuss synergies among all three.

2 User-centric View

By *user-centric view* on Web archives we refer to access from the perspective of a user without requiring additional infrastructure or knowledge about the underlying data structures. This includes the normal user who wants to look up an archived webpage as well as scholarly users who *closely* read individual webpages to understand their content and the context rather than or prior to analyzing collections in a *data analysis* or *distant reading* fashion [7]. Hence, the *user-centric view* involves the lookup and display of pages in a way that is suitable for the user. Similar to the use of the live Web, where users either directly enter the URL of a webpage in a browser or utilize search engines to find the desired page, access to Web archives from a user's perspective can be distinguished into *direct access* and *search* as well.

2.1 User Access to Web archives

Direct access to an archived webpage is commonly done through the Internet Archive's Wayback Machine, which is also used by most other archives[3] in order to provide access to their Web collections. This way of access involves entering a URL first and selecting the desired version of the corresponding webpage from a calendar view that presents all available snapshots to the user. As URLs can be cumbersome, Web users often prefer search engines over remembering and typing URLs manually. An alternative to that is to follow hyperlinks from known pages. With Web archives being temporal collections, such a link needs to carry a timestamp in addition to the URL. Within the Wayback Machine this timestamp is as close as possible to the active capture of the originating webpage. However, links can also point from the outside of a Web archive, i.e., the live Web, into the archive, in which case the timestamp needs to be set explicitly.

One approach to form such hyperlinks is to incorporate temporal information mentioned together with the link target. We recently investigated this for the case of (mathematical) software that is cited or mentioned in scientific publications. The websites corresponding to that software often nicely describe and document the tool or application. Furthermore, we found that for many purposes the archived versions can be considered surrogates of the corresponding software version that was referred to in an article [5]. In this case, the publication date is a good indicator, or at least a close estimate, of the target time for linking the publication and mentioned software. Together with an information service for mathematical software[4], we eventually established these links in their system[5]. While this example is very domain-specific to software, the same idea can be applied to other scenarios as well, such as preserving the evolution of people by archiving their blogs and social network profiles [8, 9, 10]. Another example is

[3] `https://github.com/iipc/openwayback`

[4] `http://swMATH.org`

[5] `https://blogs.tib.eu/wp/tib/2017/05/19/what-does-the-internet-know-about-the-development-of-software`

the preservation of Web citations, like on Wikipedia, to provide access to cited page at the time when it was cited[6].

Without those direct links, search capabilities are particularly important in Web archives since the target webpage may not exist anymore or its URL may have changed. To visit such a page in the past the user needs to know the URL from that time, which is challenging without a supporting search engine. However, search in Web archives is quite challenging as well, as not only the textual relevance of a webpage may be important to the user, but also its temporal relevance. For example, a removed word, possibly resulting in a negation of a fact, may be highly relevant without changing the textual relevance to a given query. By contrast, the correction of a typo or a dynamically changing ad is probably less important, just like a new layout, which can cause relatively big changes on many pages. Identifying this tiny change of a removed word to be crucial for search is difficult, especially since the textual relevant to a matching query might not be very different from the captures of the same page just before and after this change. Furthermore, at different periods in time, different websites might be relevant to the same query, even if the textual relevance to the keywords of the query is not changing at all.

2.2 Web Archive Search

Web archive search can be considered a special case of temporal information retrieval (temporal IR). This important subfield of IR has the goal to improve search effectiveness by exploiting temporal information in documents and queries [11, 12]. The temporal dimension leads to new challenges in query understanding [13], retrieval models [14, 15] as well as temporal indexing [16, 17]. However, most temporal indexing approaches treat documents as static texts with a certain validity, which does not account for the dynamics in Web archives as described above. Furthermore, while information needs in IR are traditionally classified according to the taxonomy introduced by Broder [18], user intents are different for Web archives as studied by Costa and Silva [19]. In contrast to the majority of queries being informational, where users search for information, in Web archives queries are predominantly navigational, because users often look for specific resources in a Web archive under a temporal aspect. Costa et al. [20] presented a survey of existing Web archive search architectures and Hockx-Yu [21] identified 15 Web archives that feature full-text search capabilities. With the incorporation of live Web search engines, Kanhabua et al. [3] demonstrate how to search in a Web archive without indexing it.

The Internet Archive's Wayback Machine recently got its own *site search* feature to support users looking for websites about specific keywords without knowing the URLs [22]. Their system is based on anchor texts that are used in hyperlinks, extracted from archived pages across time. While this is a great improvement over the original URL lookup approach, it has some limitations. Due to the enormous amount of URLs in the archive, the feature is restricted to

[6] `https://en.wikipedia.org/wiki/Help:Using_the_Wayback_Machine`

homepages and does not yield deep links into a website. Also, the *Wayback Site Search* has no explicit temporal search support, i.e., users cannot specify a time interval for their queries.

With Tempas we have built a system with the goal to provide *temporal archive search* for *authority pages* given a keyword query together with a time interval [1, 4], e.g., *"what were the most central webpages of Barack Obama before he became president in 2005?"*. This would bring up Obama's senator website rather than his today's website and social media accounts. As we discussed before, such temporal semantics can often not be derived from the webpages under consideration and require external indicators. In our first version[7] we incorporated tags attached to URLs on the social bookmarking platform *Delicious*. Without evaluating the precision of the ranking, which was based on the frequency of a tag used with a URL, we showed that this approach results in a good (temporal) recall with respect to query logs from AOL and MSN [2]. However, since Delicious is a closed system, available data is limited and our dataset only ranges from 2003 to 2011. Also, we found that it shows a strong bias towards certain topics, like technology. For these reasons, we also switched to hyperlinks in the second version of Tempas. Using a graph-based query model, Tempas v2[8] exploits the number of websites and corresponding anchor texts linking to a URL in a given time interval. Its temporally sensitive search for *authority pages* of entities in Web archives has shown to be very effective in multiple scenarios [4].

This outline of Tempas as well as the *Wayback Site Search* system shows the close relationship between search and the *graph-centric view* on Web archives, which is detailed in Section 4. Also, by zooming out of search results, the line between the *user-* and *data-centric view* (Sec. 3) can be blurry as well, as demonstrated with *Shine*[9], an information retrieval system by the *UK Web Archive*[10] that supports trend analysis of Web archive content.

3 Data-centric View

Web archives are commonly organized in two data formats: *WARC files* (Web Archive files) store the actual archived contents, while *CDX files* (Capture Index) are comprised of lightweight metadata records. The *data-centric view* approaches Web archives from these files, which is how data scientists would typically look at it. This perspective provides a higher, superior point of view, looking at whole collections rather than individual records nicely rendered for a user. However, we also have to deal with much lower data access and processing techniques at this level.

In the following, we distinguish between a *data-centric view* with the focus on Web archives as representations of the actual Web, in which the object of study is the Web's evolution and its *dynamics*, and the perspective focusing on

[7] http://tempas.L3S.de/v1

[8] http://tempas.L3S.de/v2

[9] https://github.com/ukwa/shine/wiki

[10] http://www.webarchive.org.uk/

the contents of webpages to derive insights into the real world. The latter is referred to as the discipline of *Web Science* [23].

3.1 Web Dynamics Analysis

Web archives spanning multiple years constitute a valuable resource to study the evolution of the Web as well as its dynamics. Already in a very early work, Cho and Garcia-Molina [24] analyzed webpages to obtain implications for an incremental crawler and found that 40% of them change within one week. However, they studied a rather small collection of only 720,000 pages over 4 months. On a larger scale, Fetterly et al. [25] analyzed 150 million webpages over a period of 11 weeks and report that 67% of the pages never change, 20% are only minor text changes and 10% of the webpages have changes in the non-textual part. Smaller, with 3-5 million pages, but over one year, Ntoulas et al. [26] observed 8% of the pages are replaced by newly created ones every week. About 50% did not change at all during the year under consideration. Koehler [27] was one of the first who studied snapshots over multiple years from 1996 to 2001, but only for a small sample of 360 pages. He showed that navigation pages have a better survival rate than content pages. A more fine-grained study by Adar et al. [28] reported that 66% of their studied pages changed on average every 123 hours. Although their collection was again small with 55,000 pages over 5 weeks, they analyzed hourly and sub-hourly changes, which is a crawl rate that cannot be provided by most Web archives.

With access to existing Web archives, more recent studies of the Web were conducted retrospectively on available data [29, 30, 31]. However, instead of analyzing the whole archive at once, all of them focus on a particular subset, such as national domains. Thanks to the Internet Archive we were provided with their entire subset of German pages over 18 years, i.e., the top-level domain .de from 1996 to 2013, which enabled us to carry out an analysis of the dawn of today's most popular German domains [32]. In this study, we explored how the age, volume and sizes of popular pages have evolved over the last decade. We found that most of the popular educational domains like universities have already existed for more than a decade, while domains relating to shopping and games have emerged steadily. Further, we see that the Web is getting older, not in all its parts, but with many domains having a constant fraction of webpages that are more than five years old and aging further. Finally, we see that popular websites have been growing exponentially after their inception, doubling in volume every two years, and also newborn pages have gotten bigger over time. These insights allow for predictions of the Web's future and lead to new questions, like: *"How does the Web of other countries compare to this analysis of the German Web?"*, *"How do webpages evolve content-wise compared to size and age, and why is the average size of the newborn webpages today larger than the ones in the yesteryear?"*, *"Is it because of an actual increase in content or is it because of the markup due to constantly increasing web authoring technologies?"*.

3.2 Web Archive Data Processing

Due to the sheer size of Web archives, in the order of multiple terabytes or even petabytes, distributed computing facilities are needed to process archived Web data efficiently. Common operations, like selection, filtering, transformation and aggregation, can be performed using the generic *MapReduce* programming model [33], as supported by *Apache Hadoop*[11] or *Apache Spark*[12] [34]. AlSum [35] presents with *ArcContent* a tool specifically for Web archives using the distributed database *Cassandra* [36]. In this approach, the records of interest are selected by means of the CDX records and inserted into the database to be queried through a web service. *Warcbase* by Lin et al. [37] follows a similar approach based on *HBase*, a Hadoop-based distributed database system, which is an open-source implementation of Google's *Bigtable* [38]. While being very efficient for lookups, major drawbacks of these database solutions are the limited flexibility as well as the additional effort to insert the records, which is expensive in terms of time and resources. In a later version, *Warcbase* allows to load and process (WARC) files directly using *Apache Spark* in order to avoid the *HBase* overhead, for which they provide convenience functions to work with Web archives.

With ArchiveSpark[13] we presented a novel data processing approach for Web archives that exploits the CDX metadata records for gains in efficiency while not having to rely on an external index [6]. This tool for general Web archive access is based on *Spark* as well and supports arbitrary filtering and data derivation operations on archived data in an easy and efficient way. Starting from the small and lightweight metadata records it can run basic operations, such as filtering, grouping and sorting very efficiently, without touching the actual data payloads. In a step-wise approach the records are enriched with additional information by applying external modules that can be customized and shared among researches and tasks. These modules can integrate any third-party tools to extract or generate new information from webpage contents, like words of different kinds, out-going links or *Named Entities*, such as persons, organizations or locations. Only at this step, ArchiveSpark seamlessly integrates the actual records of interest stored in WARC files. Internally, ArchiveSpark documents the lineage of all derived and extracted information, which can serve as source for additional filtering and processing steps or stored in a convenient output format to be used as research corpus in further studies. Benchmarks show that ArchiveSpark is faster than competitors, like *Warcbase* and pure *Spark* in typical use case scenarios when working with Web archive data.

4 Graph-centric View

The *graph-centric view* on Web archives enables their exploration from a structural perspective. In contrast to *user-centric* and *data-centric* views, here the focus is not on content or single records in a Web archives, but the relations among

[11] https://hadoop.apache.org

[12] https://spark.apache.org

[13] https://github.com/helgeho/ArchiveSpark

them. In the context of the Web, the most obvious relations are hyperlinks that connect webpages by pointing from one page to another. However, there is more that is less obvious. Looking at hyperlinks from a more coarse-grained perspective, multiple links can be combined to connections among hosts, domains or even top-level domains, revealing the connections among services, organizations or the different national regions of the Web, respectively. Furthermore, by zooming out to the graph perspective after processing the archived data from a *data-centric view* (s. Sec. 3), even relationships among persons or objects mentioned on the analyzed pages can be derived [39].

Works specifically on graphs in Web archives are very limited, but scientists have looked into graph properties of the Web in general both on static [40, 41, 42, 43, 44] and evolving graph [45, 46, 47]. However, it is worth looking at graphs in Web archives as a special case of Web graphs, because archives are never complete and not crawled with a particular attention to preserving the original Web graph structures. At the same time though, they are our only source to analyze the Web of the past and its evolution retrospectively. As a consequence, questions on the completeness and consistency of graphs extracted from Web archives arise, such as: *"How well do structures and properties of graphs extracted from a Web archive resemble the graph of the actual Web and what is the impact of missing pages on the behavior or results of graph algorithm, like PageRank?"* [48]. These aspects have often been neglected in the past and need to be studied in more detail in the future.

Besides these structural questions, the *graph-centric view* on Web archives is crucial to get an overview of available records in an archive and to find the right resources. Hyperlinks among the archived pages can point a user or an algorithm in search or data analysis tasks to the desired entry points within the big and often chaotic Web archive collections. As we discussed in Section 2, we make use of this in Tempas, our temporal search engine for Web archives [4]. The effectiveness of hyperlinks and the descriptive anchor texts of such links for the task of *site finding* was already shown by Craswell et al. [49]. They are reported to be twice as effective as searching the content of pages, which can be consider a rather *data-centric* approach to search. The authors in Kraaij et al. [50] combined anchor texts with content features for *entry page search* and also found that search just based on anchor texts outperforms basic content features. In a similar experiment, Ogilvie and Callan [51] showed that anchor texts are among the most effective features for the task of finding homepages. Koolen and Kamps [52] re-evaluated the effectiveness of anchor texts in ad-hoc retrieval and showed that propagated anchor text outperforms full-text retrieval in terms of early precision.

5 Discussion and Synergies

In this paper, we have presented three views on accessing Web archives: from a *user's perspective*, from a more comprehensive *data-centric* point of view and from a structural perspective, focusing on the *graphs* spanning a Web archive.

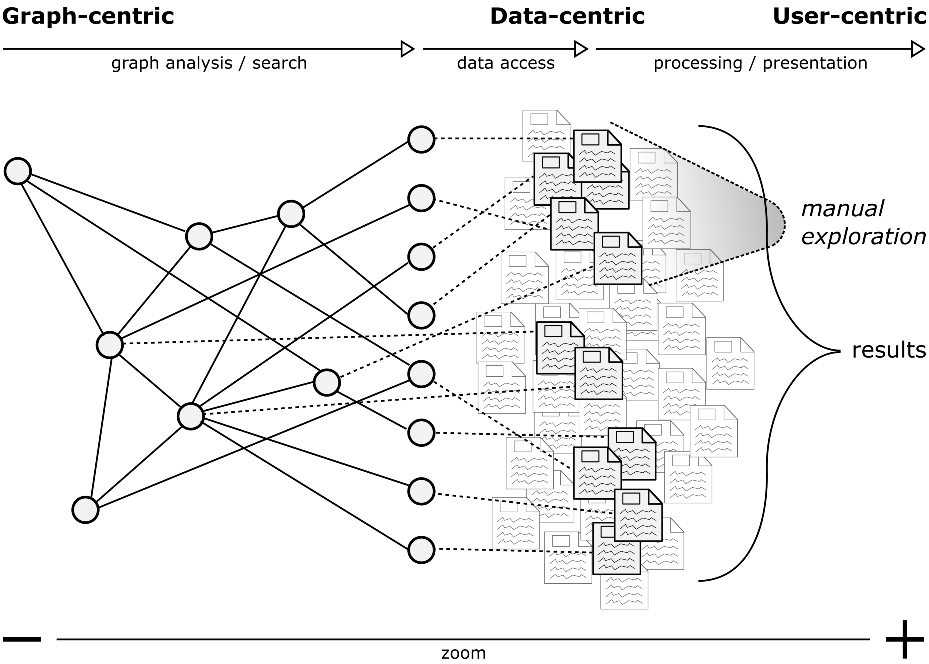

Fig. 2: Combining different views on Web archives for systematic data analysis.

In the practical work with Web archives, these views are often combined and complement each other. They can be considered different zoom levels to look at archived data, as illustrated in Figure 1: the *user-centric view* focuses on single records in a rendered or nicely presented form, while the *data-centric* and *graph-centric* views are zoomed out to whole collections in a more raw form, looking at datasets of archived resources or their relations without even taking the contents into account.

Synergies among these views have already been pointed out in the previous sections. In Section 2 we discussed the usefulness of the broader *graph-centric* perspective to guide a user in search to the webpages of interest. Furthermore, to extract graphs as described in Section 4, a data processing step is required, approaching archives from a *data-centric* perspective as we outlined in Section 3. An example of such an interplay is presented in Fafalios et al. [53] to build a *semantic layer* for Web archives, i.e., triples of facts extracted from versioned documents. The triples in the experiments described in that paper are generated by a data processing pipeline using an extension of ArchiveSpark (s. Sec. 3), called ArchiveSpark2Triples[14], which extracts named entities on a small topical Web archive collection. This *semantic layer* can be queried on a structural level, from a *graph-centric* perspective, by connecting records that share certain facts, like those mentioning the same persons. Once a set of documents that match the query has been identified, a data-scientist may zoom in to look at the contents

[14] https://github.com/helgeho/ArchiveSpark2Triples

with a *data-centric* view, for instance to analyze the context in which the persons appear. Quite commonly, such workflows also involve manual exploration and inspections of the records under consideration from a *user-centric* perspective. This is helpful to get an understanding of the data under consideration.

Figure 2 shows a generic analysis schema that outlines a systematic way to study web archives. This schema can be adopted and implemented for different scenarios, like the one described above. Here, the *graph-centric view* is utilized to get an overview and find suitable entry points into the archive. This may be first done manually by the user to get an feeling for the available data using some search engine, e.g., Tempas (s. Sec. 2), but can also be integrated as the first step in a data processing pipeline to (semi-)automatically select the corpus for further steps. Next, the selected records can be accessed from a *data-centric* view at scale, for instance using ArchiveSpark (s. Sec. 3), to extract the desired information, compute metrics or aggregate statistics. Finally, the results are presented back to the user. A concrete implementation of this pipeline is outlined in Holzmann et al. [4], where we describe the example of analyzing restaurant menus and compare prices before and after the introduction of the Euro as Europe's new currency in Germany in 2001/2002.

This type of analysis can be seen as a generalization of the *distant reading* idea [7], which refers to the analysis of big (text) corpora from a structural point of view by modeling the collection as a network of documents or contained objects, as opposed to reading every single document, referred to as *close reading.* A good overview of the related work in *distant reading* from the visual analytics perspective for *Digital Humanities* has been published by Jänicke et al. [54]. Jackson et al. [55] discuss the integration of *distant reading* features into web archive search by providing visualizations at a different zoom level. This would enable a more integrated perspective by providing access to all three views to the user. At the same time, however, it may limit the flexibility of dealing with each view individually to get a detailed look at Web archives from different perspectives.

References

[1] Helge Holzmann and Avishek Anand. Tempas: Temporal Archive Search Based on Tags. In *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016.

[2] Helge Holzmann, Wolfgang Nejdl, and Avishek Anand. On the Applicability of Delicious for Temporal Search on Web Archives. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016. doi: 10.1145/2911451.2914724.

[3] Nattiya Kanhabua, Philipp Kemkes, Wolfgang Nejdl, Tu Ngoc Nguyen, Felipe Reis, and Nam Khanh Tran. How to Search the Internet Archive Without Indexing It. In *Proceedings of the 20th International Conference on Theory and Practice of Digital Libraries (TPDL)*, 2016. doi: 10.1007/978-3-319-43997-6_12.

[4] Helge Holzmann, Wolfgang Nejdl, and Avishek Anand. Exploring Web Archives Through Temporal Anchor Texts. In *Proceedings of the 9th International ACM Web Science Conference 2017*, 2017. doi: 10.1145/3091478.3091500. (to appear).

[5] Helge Holzmann, Wolfram Sperber, and Mila Runnwerth. Archiving Software Surrogates on the Web for Future Reference. In *Proceedings of the 20th International Conference on Theory and Practice of Digital Libraries (TPDL)*, 2016. doi: 10.1007/978-3-319-43997-6_17.

[6] Helge Holzmann, Vinay Goel, and Avishek Anand. ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL)*, 2016. doi: 10.1145/2910896.2910902.

[7] Franco Moretti. *Graphs, Maps, Trees: Abstract Models for a Literary History.* Verso, 2005.

[8] Nikos Kasioumis, Vangelis Banos, and Hendrik Kalb. Towards building a blog preservation platform. *World Wide Web Journal*, 17, 2014.

[9] Catherine C Marshall and Frank M Shipman. An argument for archiving facebook as a heterogeneous personal store. In *JCDL 2014 (DL 2014)*.

[10] Hany M. SalahEldeen and Michael L. Nelson. Losing my revolution: How many resources shared on social media have been lost? In *TPDL 2012*.

[11] Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. On the Value of Temporal Information in Information Retrieval. *SIGIR Forum*, 41(2):35–41, 2007. doi: 10.1145/1328964.1328968.

[12] Ricardo Campos, Gal Dias, Alpio M. Jorge, and Adam Jatowt. Survey of Temporal Information Retrieval and Related Applications. *ACM Comput. Surv.*, 47(2), 2014. doi: 10.1145/2619088.

[13] Rosie Jones and Fernando Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3), 2007. doi: 10.1145/1247715.1247720.

[14] Klaus Berberich, Srikanta Bedathur, Omar Alonso, and Gerhard Weikum. A Language Modeling Approach for Temporal Information Needs. In *Proceedings of the 32Nd European Conference on Advances in Information Retrieval (ECIR)*, 2010. doi: 10.1007/978-3-642-12275-0_5.

[15] Jaspreet Singh, Wolfgang Nejdl, and Avishek Anand. History by Diversity: Helping Historians Search News Archives. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR)*, 2016. doi: 10.1145/2854946.2854959.

[16] Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. Temporal Index Sharding for Space-Time Efficiency in Archive Search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011.

[17] Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. Index Maintenance for Time-Travel Text Search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '12, 2012.

[18] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[19] Miguel Costa and Mário J Silva. Understanding the Information Needs of Web Archive Users . In *Proceedings of the 10th International Web Archiving Workshop*, 2010.

[20] Miguel Costa, Daniel Gomes, Francisco Couto, and Mário Silva. A Survey of Web Archive Search Architectures. In *Proceedings of the 22nd International Conference on World Wide Web (Companion)*, 2013.

[21] Helen Hockx-Yu. Access and scholarly use of web archives. *Alexandria*, 25(1-2): 113–127, 2014.

[22] Vinay Goel. Beta Wayback Machine - Now with Site Search!, October 2016. URL `https://blog.archive.org/2016/10/24/beta-wayback-machine-now-with-site-search`. [Accessed: 16/03/2017].

[23] Wendy Hall, Jim Hendler, and Steffen Staab. A manifesto for web science @10. *arXiv:1702.08291*, 2017.

[24] J. Cho and H. Garcia-Molina. The Evolution of the Web and Implications for an Incremental Crawler. In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00.

[25] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03.

[26] Alexandros Ntoulas, Junghoo Cho, and Christopher Olston. What's new on the web?: The evolution of the web from a search engine perspective. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04.

[27] Wallace Koehler. Web Page Change and Persistence-A Four-Year Longitudinal Study. *Journal of the American Society for Information Science and Technology*, 53(2):162–171, 2002.

[28] Eytan Adar, Jaime Teevan, Susan T. Dumais, and Jonathan L. Elsas. The web changes everything: Understanding the dynamics of web content. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09.

[29] Scott A. Hale, Taha Yasseri, Josh Cowls, Eric T. Meyer, Ralph Schroeder, and Helen Margetts. Mapping the UK webspace: Fifteen years of british universities on the web. In *Proceedings of the 2014 ACM Conference on Web Science*.

[30] Teru Agata, Yosuke Miyata, Emi Ishita, Atsushi Ikeuchi, and Shuichi Ueda. Life span of web pages: A survey of 10 million pages collected in 2001. *Digital Libraries*, 2014.

[31] Lulwah Alkwai, Michael L Nelson, and Michele C Weigle. How well are arabic websites archived? In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2015.

[32] Helge Holzmann, Wolfgang Nejdl, and Avishek Anand. The Dawn of Today's Popular Domains - A Study of the Archived German Web over 18 Years. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL)*, 2016. doi: 10.1145/2910896.2910901.

[33] Jeffrey Dean and Sanjay Ghemawat. MapReduce: a flexible data processing tool. *Communications of the ACM*, 53(1):72–77, 2010.

[34] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: cluster computing with working sets. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, volume 10, page 10, 2010.

[35] Ahmed AlSum. *Web archive services framework for tighter integration between the past and present web*. PhD thesis, Old Dominion University, 2014.

[36] Avinash Lakshman and Prashant Malik. Cassandra: a decentralized structured storage system. *ACM SIGOPS Operating Systems Review*, 44(2):35–40, 2010.

[37] Jimmy Lin, Milad Gholami, and Jinfeng Rao. Infrastructure for supporting exploration and discovery in web archives. In *WWW'14 Companion*, 2014.

[38] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C Hsieh, Deborah A Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E Gruber. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2):4, 2008.

[39] Miroslav Shaltev, Jan-Hendrik Zab, Philipp Kemkes, Stefan Siersdorfer, and Sergej Zerr. Cobwebs from the past and present: Extracting large social networks using internet archive data. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016. doi: 10.1145/2911451.2911467.

[40] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *nature*, 401(6749):130–131, 1999.

[41] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

[42] Lada A Adamic and Bernardo A Huberman. Power-law distribution of the world wide web. *science*, 287(5461):2115–2115, 2000.

[43] Torsten Suel and Jun Yuan. Compressing the Graph Structure of the Web. In *Data Compression Conference*, 2001.

[44] Paolo Boldi and Sebastiano Vigna. The webgraph framework i: compression techniques. In *Proceedings of the 13th international conference on World Wide Web*, pages 595–602. ACM, 2004.

[45] Bernardo A Huberman and Lada A Adamic. Internet: growth dynamics of the world-wide web. *Nature*, 401(6749):131–131, 1999.

[46] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

[47] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

[48] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. 1999.

[49] Nick Craswell, David Hawking, and Stephen Robertson. Effective Site Finding using Link Anchor Information. In *Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.

[50] Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. The Importance of Prior Probabilities for Entry Page Search. In *Proceedings of the 25th annual international ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 2002.

[51] Paul Ogilvie and Jamie Callan. Combining Document Representations for Known-Item Search. In *Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Informaion Retrieval.* ACM, 2003.

[52] Marijn Koolen and Jaap Kamps. The importance of anchor text for ad hoc search revisited. In *Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–129. ACM, 2010.

[53] Pavlos Fafalios, Helge Holzmann, Vaibhav Kasturia, and Wolfgang Nejdl. Building and Querying Semantic Layers for Web Archives. In *Proceedings of the 17th ACM/IEEE-CS on Joint Conference on Digital Libraries 2017*, 2017. (to appear).

[54] Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. On close and distant reading in digital humanities: A survey and future challenges. *Proc. of EuroVisSTARs*, pages 83–103, 2015.

[55] Andrew Jackson, Jimmy Lin, Ian Milligan, and Nick Ruest. Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL)*, 2016.