

Browsing and Recomposition Policies to Minimize Temporal Error When Utilizing Web Archives

Scott G. Ainsworth
Old Dominion University
Norfolk, VA, USA
sainswor@cs.odu.edu

ABSTRACT

Public web archiving on a large scale began in the late 1990s with archives such as Australia’s Pandora and the Internet Archive. These archives were (and still are) much less rigorous than records management systems; indeed, they do not have the resources needed to approach records management rigor. Thus the archives are incomplete, which leads to temporal discrepancies when browsing the archives and recomposing web pages. When browsing, the user-selected target datetime drifts without notice. When viewing a composite resource, the embedded resources may have a temporal spread of many years, which is invisible to the user. Preliminary investigation has confirmed and partially measured these temporal issues. This proposed body of work will measure drift and spread, determining their impact on web archive usability, and develop policies and heuristics to balance completeness, temporal coherence, and usability based on user goals.

Categories and Subject Descriptors

H.3.7 [Information Storage and Retrieval]: Digital Libraries

Keywords

Digital Preservation, HTTP, Resource Versioning, Temporal Applications, Web Architecture, Web Archiving

1. INTRODUCTION

1.1 Background

To browse archived pages from a web archive such as the Internet Archive [20], the user begins with the selection of a URI followed by selection of a Memento-Datetime (the datetime the resource was archived). The selected memento (archive copy) is then presented to the user [ex. ODU Computer Science home page, Figure 1(a)]. The user then is able to browse the archive’s collection of mementos by clicking links on the displayed pages in a process similar to browsing the live Web. For example, the user might click to navigate to the ODU College of Sciences home page [Figure 1(b)]. However, with each click, the target datetime (the datetime requested by the user) is changed to the Memento-Datetime of the displayed page. In this example, when the user returns to the Computer Science home page, a different version is shown [Figure 1(c)] even though the user has not consciously changed the target datetime.

Simultaneously, another temporal incoherence is subtly occurring. The embedded mementos (e.g., images) can have significantly different Memento-Datetimes than the root memento containing them. Figure 2 graphs an example of the temporal spread of 11 composite mementos (pages and their embedded resources) for a single URI. The red diamonds are the page itself, yellow and gray diamonds are embedded mementos. The maximum temporal spread shown is unexpected (over 2 years) and not reflected by the interface presented to the user, which displays only the root’s Memento-Datetime.

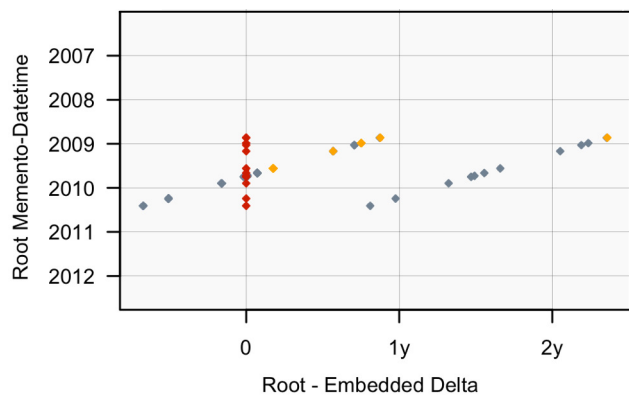


Figure 2: Example of Temporal Spread

These simple examples raise many questions. How much temporal drift do users experience when browsing archives using services such as the Wayback Machine? How much temporal spread exists in composite mementos? How can drift and spread be minimized? What are the best ways to measure drift and spread? Given the somewhat ad-hoc nature of early web archiving, what results can be expected and what is the best that can be achieved? What factors contribute, positively or negatively, to the amount of drift and spread? Does combining multiple archives produce a better result than using a single archive? Would users with differing goals (e.g., casual browsing, historical research) benefit from different drift and spread minimization policies? What are the best policies to meet user goals?

2. BACKGROUND AND RELATED WORK

Although the need for web archiving has been understood since nearly the dawn of the Web [8], these efforts have been for the most part independent in motivation, requirements,



Figure 1: Impact of Drift on Archive Browsing

and scope. The Internet Archive, the first archive to attempt global scope, came into existence in 1995 [14]. Since then, many other archives have come into existence. Some of these use software developed by the Internet Archive and have similar capture behavior and user interfaces; however, other archives such as WebCite [12] have significantly different capture behaviors.

Large-scale web archiving requires resolution of issues and approaches on several axes; although a little out of date, Masanès [17] is an excellent introduction and covers a broad range of web archiving topics. Of significance to this research are the technical aspects of acquisition, organization and storage, and quality and completeness. A major area not addressed by Masanès is archive access, in particular the lack of standards or conventions for accessing archived resources. Van de Sompel et al. [28] addressed this lack with Memento.

2.1 Acquisition

Acquisition is the technical means of bringing content into an archive. Client-side archiving essentially emulates web users following links, obtaining content using the HTTP protocol. The Heritrix [19] crawler and the mirroring capability of *wget*¹ are examples of client-side archiving. A significant issue with client-side archiving is that only those parts of the Web exposed as linked resources are captured. Transactional archiving is specifically designed to overcome this limitation. Transactional archiving inserts the capture process between the user and the data source, for example an Apache web server filter, which requires the cooperation of the server operator. Unique request-response pairs are archived, including requests for resources that are not linked. Server-side archiving makes a direct copy of the content from the server, bypassing HTTP altogether. Although conceptually simple, access to the resulting server-side archive can be difficult, requiring different URIs and navigational structures than the original. Many systems, e.g. content management systems and wikis, perform server-side archiving by design.

2.2 Organization and Storage

Once acquired, content must be stored. Masanès [17] describes three organization and storage methods that are commonly used. Local archives store content in the local file system, transforming the content just enough to allow off-line browsing. Links must be modified to refer-

ence either locally-stored archived resources or the live web. Strict adherence to the original content is generally impractical and size is limited by local storage capacity and speed. Thus, local archives are most suitable for small-scale archiving. Common methods of creating local archives are *wget* mirroring and the Save functions available in common web browsers. Web-served archives, like the IA, commonly store content in WARC (Web ARChive) container files, which allows the original representation and URIs to be stored unmodified, which overcomes many limitations imposed by file systems. Web-served archiving is highly scalable and suitable for large-scale archiving. Non-web archives generally transform web content into other forms. For example, Adobe Acrobat has the ability to download web content and produce a corresponding PDF. This type of archiving is generally best suited for resources, such as digitized books, originally created independently from the Web. Of the three types of organization and storage methods, only web-served archives are relevant to this study.

2.3 Access

An area of web archives that remained unresolved until recently was lack of methods or a standard API for time-based access to archived resources. Each archive provides a user interface (UI) to access the archive's resources. (Many archive's use the Internet Archive's Wayback Machine [26] and therefore share similar UIs.) In general, UI access to archives starts with a user-selected URI and datetime, after which the archive allows the user to simply click links to browse the collection.

Van de Sompel et al. addressed the lack of a standard API with Memento [27, 28], which is an HTTP-based framework that bridges web archives with current resources. It provides a standard API for identifying and dereferencing archived resources through datetime negotiation. In Memento, each original resource, URI-R, has zero or more archived representations, URI-M_i, that encapsulate the URI-R's state at times t_i . A list of URI-Ms for a URI-R is called a timemap (URI-T). Using the Memento API, clients are able to request URI-M_i for a specified URI-R by datetime. Memento is now an IETF Internet Draft [27].

2.4 Quality and Completeness

In general, quality is functionally defined as fitting a particular use and objectively defined as meeting measurable characteristics. This examination of web archive content is concerned with the latter. For web archives, most quality

¹<http://www.gnu.org/software/wget/>

issues stem from the difficulties inherent in obtaining content using HTTP [17]. Content is not always available when crawled, leaving gaps in the coverage. Web sites change faster than crawls can acquire their content, which leads to temporal incoherence. Ben Saad et al. [7] note that quality and completeness require different methods and measures *a priori* or *a posteriori*, that is during acquisition or during post-archival access respectively.

2.4.1 Completeness (Coverage)

When crawling to acquire content, the trade-offs required and conditions encountered lead to incomplete content, coverage, or both. A web archive may lack the resources to acquire and store all content discovered. Associated compromises include acquiring only high-priority content and crawling less often. The content to be acquired may not be available at crawl time due to server downtime or network disruption. The combination of compromises and resource unavailability create undesired, undocumented gaps in the archive.

Although much has been written on the technical, social, legal, and political issues of web archiving; little detailed research has been conducted on the archive coverage provided by the existing archives. Day [10] surveyed a large number of archives as part of investigating the methods and issues associated with archiving. Day however does not address coverage. Thelwall touches on coverage when he addresses international bias in the Internet Archive [25], but does not directly address how much of the Web is covered. McCown and Nelson address coverage [18], but their research is limited to search engine caches. Ben Saad et al. [6, 5] address qualitative completeness through change detection to identify and archive important changes (rather than simply archiving every change). This research primarily addresses *a priori* completeness. *A posteriori* web archive coverage is addressed by Ainsworth et al. [1]. Leveraging the Memento API and pilot infrastructure, Ainsworth et al. [1] obtained results showing that 35–90% of publicly-accessible URIs have at least one publicly-accessible archived copy, 17–49% have two to five copies, 1–8% have six to ten copies, and 8–63% at least ten copies. The number of URI copies varies as a function of time, but only 14.6–31.3% of URIs are archived more than once per month. The research also shows that coverage is dependent on social popularity.

2.4.2 Temporal Coherence

When crawling and acquiring content, web archives must make tradeoffs. Crawling consumes server resources, thus crawls must be polite, e.g. paced to avoid adversely impacting the server. The web archive may not have the bandwidth needed to crawl quickly. These and other constraints increase crawl duration, which in turn increases the likelihood of temporal incoherence.

Spaniol et al. [23] note that crawls may span hours or days, increasing the risk of temporal incoherence, especially for large sites, and introduces a model for identifying coherent sections of archives, which provides a measure of quality. Spaniol et al. also present a crawling strategy which helps minimize incoherence in web site captures. In another paper, Spaniol et al. [24] also develop crawl and site coherence visualizations. Spaniol’s work, while presenting an *a posteriori* measure, concerns the quality of entire crawls.

Denev et al. present the SHARC framework [11], which introduces a stochastic notion of *sharpness*. Sites changes are modeled as Poisson processes with page-specific change rates. Change rates can differ by MIME type and depths within the site. This model allows reasoning on the expected sharpness of an acquisition crawl. From this they propose four algorithms for site crawling. Denev’s work focuses on *a priori* quality of entire crawls and does not address the quality of existing archives and crawls.

Ben Saad et al. [7] address both *a priori* and *a posteriori* quality. Like Denev et al. [11], the *a priori* solution is designed to optimize the crawling process for archival quality. The *a posteriori* solution uses information collected by the *a priori* solution to direct the user to the most coherent archived versions.

Nearly all web archive research shares a common thread: evaluation and control of completeness and temporal coherence during the crawl with the goal of improving the archiving process. In contrast, this research takes a detailed look at the quality and use of existing web archive holdings with the goal of determining how they may be best utilized.

2.5 Use Patterns

To date, little research has been conducted on web archive use patterns. AlNoamany et al. [4] have looked at real-world access patterns through analysis of the Internet Archive’s web server logs. They categorize use into four groups: Dip, Slide, Dive, and Skim. Dip is a request for a single memento. Slide is requests for different multiple mementos for the same original URI. Dive follows hyperlinks while remaining near the same target datetime. Skim requests only timemaps (lists of mementos). AlNoamany et al. also found that human archive users most frequently exhibit the Dip and Dive patterns and seldom access timemaps. Robots split their accesses between the Dip and Skim patterns and request timemaps much more frequently than human users.

2.6 Identifying Duplicates

Our observations indicate that web archives frequently capture, or at least return, identical representations for multiple Memento-Datetimes. For some files type such as images, duplicates are easily identified using message digest algorithms (e.g. MD5) because the mementos are not modified by the archives. Text file types, and HTML in particular, are modified by every archive this researcher has studied. The changes are minor and invisible to most archive users; and, generally comprise simple HTML comments that identify the original URI, capture datetime, and other metadata. However, the placement, content, and formatting of this information varies, even within the same archive, which means determining duplication cannot be successful using message digests. Lexical Signatures [21, 15] and the *SimHash* [9, 16] algorithm may be able to help determine if HTML files are unchanged, or are at least substantially similar.

3. PRELIMINARY WORK

3.1 How Much of the Web Is Archived?

This researcher was the lead author for our “How Much of the Web Is Archived?” paper [1, 2], which was published and presented at JCDL 2011. The study used samples of 1,000 URIs randomly-selected URIs from four sources: DMOZ,

Delicious, Bitly, and a search engine index. We found that each sample set provides its own bias, but that popularity has a significant impact on archival rate. The results from our sample sets indicate that from 35%-90% of the Web has at least one archived copy, 17%-49% has between 2-5 copies, 1%-8% has 6-10 copies, and 8%-63% has more than 10 copies in public web archives. The number of copies varies as a function of time, but no more than 31.3% of URIs are archived more than once per month.

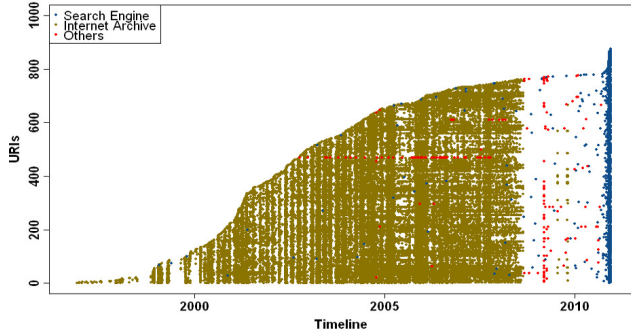


Figure 3: DMOZ Mementos by Date

Figure 3 shows the memento distribution for the DMOZ sample. The vertical axis represents sample URIs, sorted by earliest Memento-Datetime; each row represents a single URI-R. The horizontal axis is the datetime timeline. Each memento is plotted with the Internet Archive in brown, search engines in blue, and other archives in red. The Internet Archive is the major contributor; however, most of its holdings have an embargo period, which is evident at about 2009 in the figure. This work indicates that searching for mementos in web archives should have a good success rate; and, is supported by the temporal spread (Section 3.3) and the temporal drift (Section 3.2) work.

3.2 Temporal Drift

The “Evaluating Sliding and Sticky Target Policies by Measuring Temporal Drift in Acyclic Walks Through a Web Archive” [3] paper will be published at JCDL 2013. This research examines the datetime drift that occurs browsing web archives using user interfaces such as the Internet Archive’s Wayback Machine. A comparison is made between the Wayback Machine and the MementoFox (Section 3.5) extension for the Firefox web browser; these two approaches are concrete implementations of two policies. The Sliding policy allows the target datetime to follow the Memento-Datetime of archived web pages as they are viewed by the user; the sliding policy is the *de facto* policy of the Wayback Machine. The Sticky policy fixes the target datetime and uses it for every memento request. It is the policy employed by MementoFox.

The study conducted 200,000 randomly-selected acyclic walks using the Wayback Machine and Internet Archive Memento API. Figure 4 illustrates the distribution of Wayback Machine mementos by drift. The horizontal axis is the walk step number. The vertical axis is the drift in years from the walk step’s target datetime. Color indicates memento density on an exponential axis. As expected, density is highest for early steps and tapers off with as walk length increases.

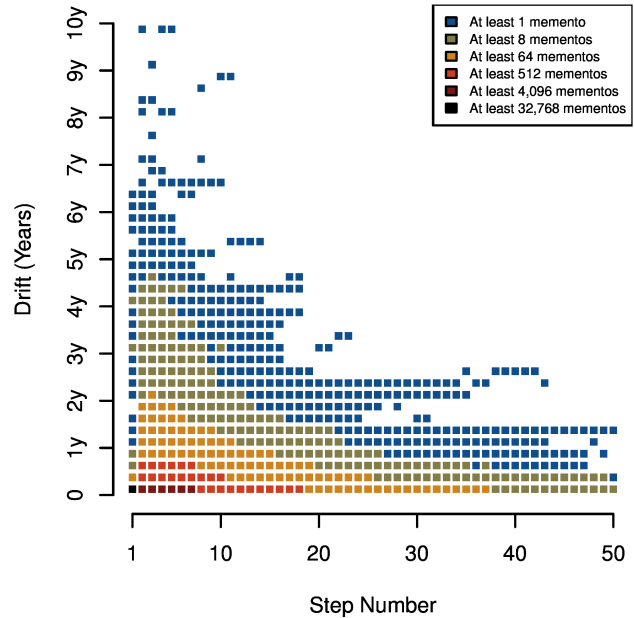


Figure 4: UI Drift by Step

We found that the Sliding Target policy drift increases with the number of walk steps, number of domains visited, and choice (number of links available). However, the Sticky Target policy controls temporal drift, holding it below 30 days on average regardless of walk length or number of domains visited. Sticky Target policy drift shows some increase as choice increases, but this could be due to other factors. In general, the Sticky Target policy generally produces at least 30 days less drift than the Sliding Target policy as shown in Figure 5.

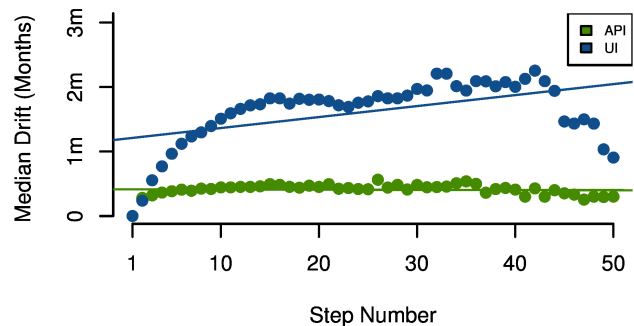


Figure 5: Median Drift by Step

3.3 Temporal Spread

Nearly all web pages are composite resources. As the tree shown in Figure 6 represents, web pages are a composition of many resources, some of which are themselves compositions (i.e., frames). Ideally, all the resources required to re-compose a composite resource are captured simultaneously in the state a user would observe using a web browser. This is seldom the case. The crawlers used by web archives by necessity retrieve HTML pages before discovering embedded resources, which are then queued for later processing. Web archives also receive content from other sources, which may

or may not include all required embedded resources. Spread is a measure of the time delta between the mementos used to recompose a composite resource.

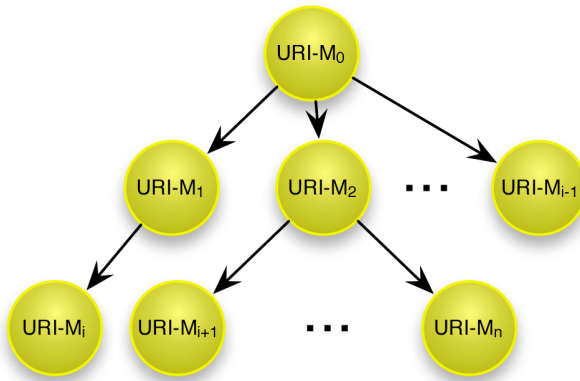


Figure 6: Composite Memento

Evaluation of Temporal spread is in progress. Preliminary results indicate that nearly all composite mementos have significant spread, frequently years, occasionally more than 10 years. Figure 7 is the ODU Computer Science home page memento for 2005-05-14 at 01:36:08. Five of the embedded mementos have their Memento-Datetime's annotated. As can be seen, all were captured at least 1 week after the root was captured and at least one embedded memento was captured over 2.1 years after the root.



Figure 7: Temporal Spread

3.4 Temporal Coherence

In addition to spread, composite mementos also frequently suffer from temporal coherence problems. We classify the temporal coherence of mementos with respect to the root memento as follows:

- **Temporally coherent.** Temporally coherent embedded mementos can be shown to have existed in their archived state at the time the root was captured. There are two ways to show temporal coherence. First, embedded mementos with the same Memento-Datetime

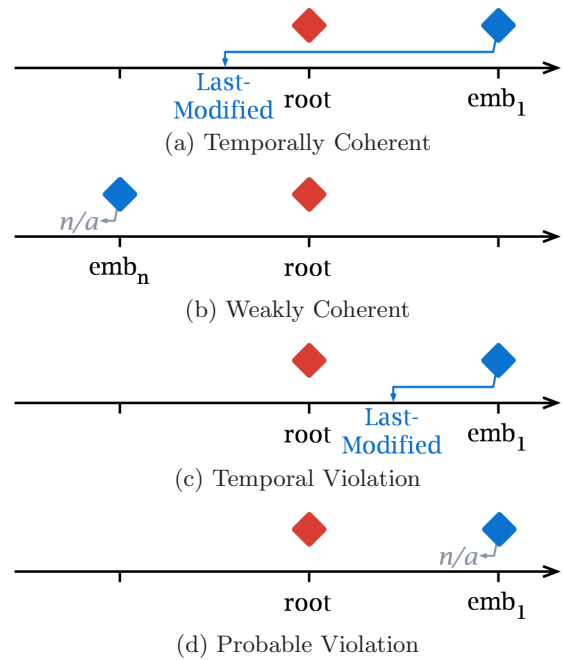


Figure 8: Temporal Coherence States

as the root are temporally coherent. Second, if an embedded memento's lifetime overlaps the root memento's Memento-Datetime, it is temporally coherent [Figure 8(a)]. A memento's lifetime is from its Last-Modified datetime through its Memento-Datetime.

- **Weakly coherent.** Weakly coherent embedded mementos were captured before the root, but lack evidence of existence at the time the root was captured [Figure 8(b)]. However, most are images and other resources that change infrequently.
- **Temporal violation.** Embedded mementos that were captured after the root and have Last-Modified datetimes after the root Memento-Datetime are temporal violations [Figure 8(c)]. The evidence indicates that they either did not exist or were changed after the root was captured.
- **Probable violation.** Embedded mementos captured after the root but lacking Last-Modified datetimes are considered probable temporal violations. [Figure 8(d)]

Figure 9 charts the spread and coherence for <http://rpgdreamers.com>. The horizontal axis is the delta of embedded mementos from the root. The deltas marks are in years. Each root memento is plotted at zero. The vertical axis is the Memento-Datetime of the root memento. Each line of mementos is a single composite memento. The black diamonds represent the root mementos; all roots are temporally coherent. Green diamonds are temporally coherent embedded mementos. Blue diamonds are weakly coherent. Red diamonds are temporal violations and yellow diamonds are probable violations. A fascinating aspect of this chart is that many embedded mementos have Memento-Datetimes captured much later than root memento. We suspect this may be due to the archive storing only one copy of resources that change infrequently, such as site logos.

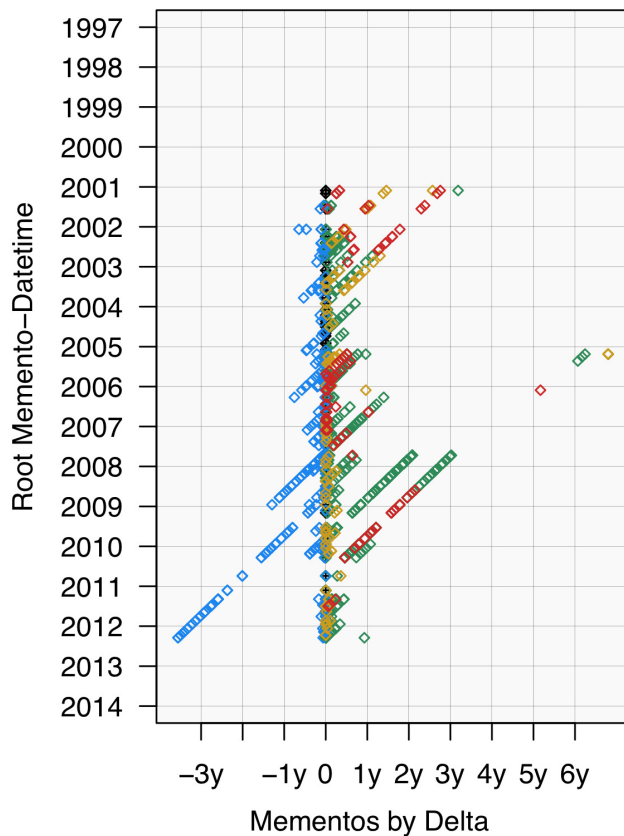


Figure 9: rpgdreamers.com Spread

3.5 MementoFox

The MementoFox extension [22] implements the Memento HTTP extensions in the Firefox web browser. The user is able to specify a target date-time and browse the archived Web using the familiar interface of the web browser. This is enabled by the combination of MementoFox, a Memento aggregator, Memento-enabled web archives, and Memento proxies. Figure 10 illustrates the MementoFox user interface.



Figure 10: Memento API and MementoFox

As of this writing, only four archives have the Memento API publicly available: the Internet Archive², the British Library's UK Web Archive³, the UK National Archives⁴, and Archive.is⁵. For other web archives, a proxy is required. The Memento aggregator allows the use of multiple archives simultaneously. The aggregator accepts Memento requests,

²<http://www.archive.org>

³<http://www.webarchive.org.uk>

⁴<http://www.nationalarchives.gov.uk/webarchive/>

⁵<http://archive.is/>

queries multiple web archives and Memento proxies, and returns a combined result to the Memento client. This researcher developed the first version of MementoFox. MementoFox provided deep insight into the Memento HTTP extensions and provided valuable feedback on practical considerations during the initial design of Memento. Memento is foundational to this research; it enables the development of tools that simplify data collection.

4. FUTURE WORK

There are many avenues of future work for this line of research. Research to date has focused primarily on characterizing temporal drift and temporal spread. The drift research also began exploring policies and heuristics, but much is still required in this area.

4.1 Browsing Patterns, Clusters, and Drift

The drift experiments conducted so far have focused on randomly-generated walks through a single archive. Al-Noamany et al. [4] have looked at real-world walk patterns through analysis of the Internet Archive's web server logs. Using these patterns to guide walks will provide more realistic link streams and result in temporal drift data more in line with actual user experience. There are also domains that users tend to avoid, such as link farms, search engine optimization (SEO) sites, and spam sites. Detecting and avoiding them, as would a user, will also move the data toward real-world user experience. We also suspect that long walk drift is heavily influenced by clusters of closely-related domains and domains that primarily self-reference. Applying an appropriate measure of clustering or similarity may shed some light on this topic.

4.2 Timemaps, Redirection and Missing Mementos

The research conducted so far has accepted the lists of mementos in timemaps as ground truth; however, our experience indicates that timemaps tell only part of the story. For example, a memento listed in a timemap may HTTP redirect to another memento with a different Memento-Datetime. These redirections are reflected in the drift research but not yet incorporated into the spread research. Since the spread research uses the timemap Memento-Datetimes, redirection will change the deltas; whether for the better or the worse needs to be determined. It is also possible that redirection is an indication of duplication; and, the timemap Memento-Datetime should be used. Further investigation is warranted. Timemaps also list mementos that do not exist or are not accessible. The research to date simply lists these as missing. (If you have used the Wayback Machine and discovered missing images, you have probably encountered this.) However, would it not be better for casual users to substitute another version? This needs to be investigated and a policy developed.

4.3 Similarity and Duplication

The spread research conducted so far has calculated deltas as the simple difference between the root and embedded memento Memento-Datetimes. Consider a root resource with a Memento-Datetime of July 5 and an embedded resource with Memento-Datetimes of July 1 and July 10. It appears that the delta for the nearest embedded memento is four days (July 5 – July 1). However, if the embedded memen-

tos are identical (e.g., a logo image that seldom changes), then the resource probably did not change and the effective delta is probably zero. But this is not certain and requires testing to determine its probability.

4.4 Communicating Status

One significant issue with existing web archive user interfaces is that temporal consistency is not communicated to the user. Early in this research, there was discussion of an icon or symbol that could quickly reflect the temporal consistency of composite mementos to users. This is on-going.

4.5 Policies and Heuristics

Web archive users will have different goals. Many are casual browsers or simply curious. Expending significant resources or time computing the best or most accurate composite memento is probably counter-productive for these users. Likewise, there are users for which the most accurate recomposition is a necessity. For example, in the February 2006 edition of the *Journal of Internet Law*, Howell [13] addresses the strict requirements to make evidence captured from the Internet Archive admissible in court. The development of policies and corresponding heuristics is a primary goal of this research and dissertation.

5. CONCLUSIONS

Researchers have proposed methods for improving temporal coherence in future web archive capture processes. However, the temporal characteristics of current web archive holdings are not well understood. The holdings need to be characterized and corresponding policies developed so that the appropriate resources are used to meet user performance, completeness, and accuracy goals. This research is well begun; the future research outline in this paper will advance our knowledge and improve usability of existing web archive holdings.

6. ACKNOWLEDGMENTS

I thank my advisor, Dr. Michael Nelson, for his continued guidance, support, and encouragement. This work supported in part by the NSF (IIS 1009392) and the Library of Congress. We are grateful to the Internet Archive for their continued support of Memento access to their archive. Memento is a joint project between the Los Alamos National Laboratory Research Library and Old Dominion University.

7. REFERENCES

- [1] S. G. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the Web is archived? In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, pages 133–136, June 2011.
- [2] S. G. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the Web is archived? Technical Report arXiv:1212.6177, Old Dominion University, December 2012.
- [3] S. G. Ainsworth and M. L. Nelson. Evaluating sliding and sticky target policies by measuring temporal drift in acyclic walks through a web archive. In *Proceedings of the 13th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL'13, July 2013.
- [4] Y. AlNoamany, M. C. Weigle, and M. L. Nelson. Access patterns for robots and humans in web archives. In *Proceedings of the 13th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL'13, July 2013.
- [5] M. Ben Saad and S. Gançarski. Archiving the Web using page changes patterns: a case study. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 113–122, 2011.
- [6] M. Ben Saad and S. Gançarski. Improving the quality of web archives through the importance of changes. In *Proceedings of the 22nd international conference on Database and expert systems applications - Volume Part I*, DEXA'11, pages 394–409, 2011.
- [7] M. Ben Saad, Z. Pehlivan, and S. Gançarski. Coherence-oriented crawling and navigation using patterns for web archives. In *Proceedings of the 15th international conference on Theory and practice of digital libraries: research and advanced technology for digital libraries*, TPD L'11, pages 421–433, 2011.
- [8] C. Casey. The Cyberarchive: a look at the storage and preservation of web sites. *College & Research Libraries*, 59, 1998.
- [9] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, STOC '02, pages 380–388, New York, NY, USA, 2002. ACM.
- [10] M. Day. Preserving the fabric of our lives: A survey of web preservation initiatives. In *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, pages 461–472, 2003.
- [11] D. Denev, A. Mazeika, M. Spaniol, and G. Weikum. SHARC: Framework for quality-conscious web archiving. In *Proceedings of the VLDB Endowment*, volume 2, pages 586–597, August 2009.
- [12] G. Eysenbach and M. Trudel. Going, going, still there: Using the WebCite service to permanently archive cited web pages. *Journal of Medical Internet Research*, 7(5), 2005.
- [13] B. A. Howell. Proving web history: How to use the internet archive. *Journal of Internet Law*, 9(8):3–9, 2006.
- [14] M. Kimpton and J. Ubois. Year-by-year: from an archive of the Internet to an archive on the Internet. In J. Masanès, editor, *Web Archiving*, chapter 9, pages 201–212. 2006.
- [15] M. Klein and M. Nelson. Revisiting lexical signatures to (re-)discover web pages. In B. Christensen-Dalsgaard, D. Castelli, B. Ammitzbøll Jurik, and J. Lippincott, editors, *Research and Advanced Technology for Digital Libraries*, volume 5173 of *Lecture Notes in Computer Science*, pages 371–382. Springer Berlin Heidelberg, 2008.
- [16] G. S. Manku, A. Jain, and A. Das Sarma. Detecting near-duplicates for web crawling. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 141–150, New York, NY, USA, 2007. ACM.

- [17] J. Masanès. Web archiving: issues and methods. In J. Masanès, editor, *Web Archiving*, chapter 1, pages 1–53. 2006.
- [18] F. McCown and M. L. Nelson. Characterization of search engine caches. In *Proceedings of IS&T Archiving 2007*, pages 48–52, May 2007.
- [19] G. Mohr, M. Stack, I. Rnitovic, D. Avery, and M. Kimpton. Introduction to Heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop, Bath, UK*, September 2004.
- [20] K. C. Negulescu. Web archiving @ the Internet Archive. http://www.digitalpreservation.gov/news/events/ndiipp_meetings/ndiipp10/docs/July21/session09/NDIIPP072110FinalIA.ppt, 2010.
- [21] S.-T. Park, D. M. Pennock, C. L. Giles, and R. Krovetz. Analysis of lexical signatures for finding lost or related documents. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '02*, pages 11–18, New York, NY, USA, 2002. ACM.
- [22] R. Sanderson, H. Shankar, S. Ainsworth, F. McCown, and S. Adams. Implementing time travel for the Web. *Code{4}Lib Journal*, (13), 2011.
- [23] M. Spaniol, D. Denev, A. Mazeika, G. Weikum, and P. Senellart. Data quality in web archiving. In *Proceedings of the 3rd Workshop on Information Credibility on the Web, WICOW '09*, pages 19–26, 2009.
- [24] M. Spaniol, A. Mazeika, D. Denev, and G. Weikum. “Catch me if you can”: Visual analysis of coherence defects in web archiving. In *The 9th International Web Archiving Workshop (IWA 2009) Corfu, Greece, September/October, 2009 Workshop Proceedings*, pages 27–37, 2009.
- [25] M. Thelwall and L. Vaughan. A fair history of the Web? examining country balance in the Internet Archive. *Library & Information Science Research*, 26(2), 2004.
- [26] B. Tofel. ‘Wayback’ for accessing web archives. In *Proceedings of the 7th International Web Archiving Workshop (IWA 07)*, 2007.
- [27] H. Van de Sompel, M. Nelson, and R. Sanderson. HTTP framework for time-based access to resource states — Memento, November 2010. <http://datatracker.ietf.org/doc/draft-vandesompel-memento/>.
- [28] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H. Shankar. Memento: Time travel for the Web. Technical Report arXiv:0911.1112, 2009.