A Framework for Aggregating Private and Public Web Archives

Extended Abstract for the JCDL 2015 PhD Consortium

Mat Kelly
Old Dominion University
Department of Computer Science
Norfolk, Virginia 23529 USA
mkelly@cs.odu.edu

ABSTRACT

Efforts to preserve content on the public Web have been effective at ensuring our collective digital heritage is not lost. However, these efforts put priority on the collective's judgement of importance, often neglecting to capture individuals' content due to additional scale, appropriateness, and technical restrictions. Individuals that take it upon themselves to preserve what they respectively deem important are ill-equipped with the base knowledge and toolset to perform personal digital preservation to create private web archives. When a user is able to create private web archives, there is little guidance in integrating private web archives with public web archives for a consistent query to replicate the content as if on the medium where it originated. This body of work will provide the ability for individually aspiring personal web archivists, both those with technical knowledge of the medium and those without, a means of preserving content previously not preserved. A framework will be created, utilizing and amending standard archiving technologies and concepts, to allow controlled access of the web archives by the creator as well as account for the integration, aggregation, and migration of private web archives with private public web archives.

1. MOTIVATION

As a modern podium, the Internet provides a means for free expression, often at the loss of the control on produced data. A large portion of the Web is public, but much of what users of the medium consider important is private or provides some restriction on access to content. Because the live Web is ephemeral, private information (and thus personally important information) is not guaranteed to exist in the same form (potentially in no form at all) on subsequent accesses to content. Preserving the private information for future access is thereby important.

However, the primary focus of contemporary web archives has been mostly focused on preserving publicly available content on the Web. Previous efforts to preserve and provide access to content that requires considerations for privacy and access control are ad hoc, non-standardized, and still fairly problematic to execute. Even when content on the private live web is preserved, the integration with archived public content from the live web is haphazardly handled or the content is not integrated at all, despite both originally existing in the same medium. Relying on institutions to preserve private web content that individuals deem important while maintaining control of access by the requesting users is unfeasible. The responsibility to preserve this content belongs to those interested.

Standard formats exist for digital preservation of live web content. These formats show little consideration for content that requires additional access control beyond simple embargoing of potentially sensitive content and a single level of authentication. Even if a user were to preserve content on the private web, little technical guidance exists on how a user should proceed to ensure good practice is used in terms of format, setup, and access of private web archives. By leveraging standard formats and practices to preserve private content, many of the benefits translate to effective preservation of the target web content.

Consider the case where Linda, a small town librarian, wishes to preserve web pages about her small town. Though not as popular as the CNNs of the live web, the small town newspaper documents the history of the town and posts many more stories on their web site compared to their print edition. Fairly new to the digital medium, the newspaper also exhibits reckless preservation with their web site on the assumption that the Internet Archive will capture the contents in case of data loss. Linda wants to be proactive and keep a historical record of locally relevant content for the library. She would like to replicate the original experience of how the web pages looked and felt and to allow her library's offerings to be shareable with those outside of their community through integration with other public web archives. Secondarily, Linda also wishes to preserve her own private web content using the methods she will learn in performing this web archiving for the library.

She begins with browsing to each web page from the newspaper and selecting her browser's "Save Webpage as" feature. The results produced from the browser are a large collection of files stored on her hard drive for each web page. When the newspaper's home page changes, she performs the same procedure but her naming scheme between "Homepage January 20th" and "Homepage January 21st" seem unwieldy and ad hoc. As a librarian, Linda wishes for a more systematic, tried-and-true approach at "web archiving". She looks to institutional tools but is deterred by the technical knowledge required to use them. She considers that, even if she were able to preserve the contents, she is unaware of any methods to execute her plan of replicating the web for the town's paper as it once was. Further, her secondary goal of using what she learns from the newspaper web archiving experience is not given consideration for archiving her private web content from the live web.

This research aims to explore the dynamics of integrating

private web archives with public web archives as a proactive approach for considerations both technical and accessible. In many instances, the web archiving approaches from humanities scholars and those that simply wish to preserve a portion of the web (like Linda) are either the easiest method to minimally accomplish the goal of personal web archiving (frequently ad hoc) or done so with institutional-level tools that do not take into account the private nature of the content. Further, content that is of a private nature likely contains Web content that is difficult to archive (e.g., dynamic JavaScript and Ajax), even with institutional quality software. Were the content accessible to these tools (they are able to satisfy the authentication requirement), capturing the content and making it suitable for replay would be problematic.

Replaying content on the archived Web acts to simulate the live Web as it was at the time of capture. Segregating public and private archives limits this correlative concept. Frameworks and archiving concepts like the Web Archive (WARC) format, Memento (for temporal navigation of archives), and Wayback (for re-experiencing or "replaying" web archives) exist to facilitate the capture and accessible replay of content from the live Web but do not account for private Web archives. Content on the private web can be captured into WARCs, replayed with Wayback, and queried with Memento. However, the comprehensiveness of the WARCs in terms of the content on the live web will likely be incomplete (for private contents), Wayback has limited access control for content (an issue if there is sensitive information on the captured page), and Memento provides no indicators for content whose representation is indicative of private web content.

In this research, I am proposing a framework to remedy these issues. By creating a framework that provides access control for content in private web archives, users that preserve private live web content can interface with their web archives in a manner that better simulates the whole Web as it was. A hierarchy of additional entities (Section 3) in the web archiving access patterns allow for this regulation and integration as well as remedy other behaviors in the state of web archiving that limit the accessibility of current technologies.

To serve as an example: Linda could download the Heritrix archival web crawler and through configuration files, specify her Facebook news feed URI at `http://facebook.com` to be archived. This initial step alone requires Linda to have some degree of technical expertise. I previously remedied this setup process with software [12] that simplifies archiving public web contents, but this is moot as the example (as follows) stands. Heritrix, by default, will not archive web pages that contain a robots.txt directive that limits crawlers. Facebook provides this directive. Disregarding this initial limitation, what the user sees at `facebook.com` while logged in via a web browser is not what an archival browser observes. The former is behind authentication and requires even further configuration, which is complicated for the user and problematic for the crawler.

Continuing on this example, if what Linda expects as the representation shown in the browser were crawled, the content would be inconsistent (Figure 1) with what she observes when viewing other `facebook.com` captures at Internet Archive's Wayback Machine. This inconsistency is the crux of the problem: should the authentication representation of `facebook.com` and the public non-authenticated

(a) (b)

Figure 1: When a user tries to use archiving tools to preserve content on the live private Web (1a) based on providing the URI, the results from an archival crawler (1b) are inconsistent with what the user observes in the browser. This further illustrates that replaying content behind authentication requires more than a URI for replay.

Facebook homepage be considered the same archived page? The method of accessing a web page using a URI as a key for archival replay seems to imply this, but the combination of private and public archives does not account for potentially sensitive content (e.g., the user's private archived contains an embarrassing Facebook post). Said content would be exposed were the two archives naively aggregated and accessible to the public.

To further exemplify this, if Linda were to capture `http://cnn.com` when realizing that a breaking story was occurring (Figure 2), should her capture be considered as valid as one from an "official" crawl by an institution like Internet Archive? From her perspective, web archive users may want to see how the story progressed, especially if the capture from her perspective was unique and not captured by the Internet Archive. That a non-vetted archivist (Linda) performed the capture may not play a role from the perspective of personal web archivists that wish to combine what they saw with the captures from vetted web archives. If Linda were to share the captures with other users and use Memento to show the progression of the story over time, no means exists to indicate that the user's capture is from a private web archive aside from what is implicit in the URI of the private archives. Per above, the URI alone is insufficient to fully account for the nature of private web archives in many cases.

Private web archiving of live web content is not limited to be executed by individuals. Consider the case where a web archive is only accessible on-site, e.g., the restricted access archives at Library of Congress or those that require one to be physically located in a British Library (BL) reading room to view[1]. These scenarios of isolating access cause a disjoint viewing experience that is not integrated with other captures (e.g., Internet Archive's captures of the same BL sites) and thus less temporally comprehensive. A more detailed picture of a site in time would be represented if the captures from the British Library were aggregated with the results of other web archives yet the original access restriction should only allow this aggregation from either the perspective of the BL (e.g., the user is on-site) or if a method existed to regulate

[1] `http://www.bl.uk/aboutus/stratpolprog/digi/webarch/webarchives.html`

access to the private BL web archives for aggregation with other sources (e.g., both IA's and potentially an individual's captures).

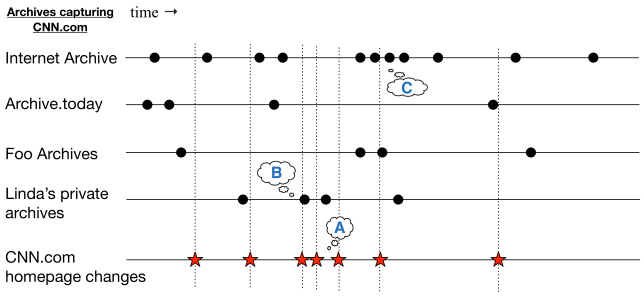

Figure 2: Web archives crawl a web site at different rates of frequency. Upon noticing an event occurring (bubble A), Linda captures the CNN.com homepage (bubble B) and continues to perform the focused crawl while the homepage rapidly changes. When Internet Archive's crawler recognized the significant event (bubble C), it does increase the crawl rate yet multiple versions of the page have already been missed.

In this research I will address the following preservation and access issues for the integration of public and private web archives.

1.1 Preservation Issues

1. Casual web users and amateur archivists use sub-optimal means for personal and private web archiving or must defer to institutions for preservation.
2. Content behind authentication is difficult to capture without compromising privacy (e.g., handing over credentials).
3. Preserving content on the live web requires delegation to tools designed for archiving rather than the perspective of the tool that originally viewed the content (the user's web browser).

1.2 Access Issues

1. Content behind authentication is difficult to replay due to the public interface (e.g., login page) and private interface (e.g., my social media news feed), residing at the same URI.
2. Private web archives and public web archives cannot be aggregated in the same way that public archives are aggregated together (i.e., in a Memento aggregator)
3. Access control to private web archives is boolean without refinement relative to the specific content contained within.
4. Web archive aggregators require manual maintenance with the set of Memento-compatible archives being static.

My prior work (Section 2) has mostly been on the preservation facet of this research while remaining research works to improve preservation and take into account the issues involving access.

2. BACKGROUND & RELATED WORK

Web archives digitally preserve our heritage in a medium where an ever-increasing portion of free expression and culture exhibitions is accumulating. The Internet Archive (IA)[2] and a number of other web archives preserve content from the live web for access at a later date. This content is publicly available and constitutes an example of a "public web archive" with little to no content restrictions.

Memento [25] is a framework for adding the dimension of time to the web - a critical characteristic for web archive access. Memento terminology is used throughout this description of research. A large portion of public web archives (including IA) support Memento. A Memento aggregator (MA) is an entity that acts as a hub for querying and combining the contents of multiple web archives. An MA provides access to the chronological results of the timestamped captures (accessible with a URI-M) of content that resides in web archives (mementos) that once existed on the live web (were accessible with a URI-R). The listing of the mementos returned from a web archive or from an MA is provided as a TimeMap (TM).

AlSum studied the selection of archives represented in the queries by currently existing MAs as a percentage of all archived content [2]. Even with a listing of URI-Ms from a TM, the URI-Ms whose content is accessible varies with the accessibility of the target archive, which varies with time as archives come on and offline [22]. The current management of adding and removing Memento-compatible archives to the Memento aggregator software is a manual process with no subscription-like model nor an API for manipulating the set of archives included in-place. Brunelle studied the caching policies of MAs, a process that aggregators use to optimize the temporally expensive operation of querying and aggregating the URI-Ms from multiple archives [5]. I took these considerations into account when building software atop currently existing aggregators. Rosenthal highlighted further issues with the then-current state of Memento aggregators [21]. Memento provides no structure to represent and differentiate mementos originating from private web archives with those from public web archives.

Rauber discussed privacy issues in archiving private web content and provided a way to programmatically identify when web content contains information that requires special handling when archived [20]. His discussion on the ethical implications of preserving this content and the current practice of access control exhibited by institutional web archives further justifies the need for a proactive means of access control instead of after-the-fact identification of private content content in web archives.

OAuth [8] is an open standard for providing authorization for resources on the web through a means of secure delegation of access without loss of access control. I use OAuth in my framework to establish authorization and regulate access to private archives using OAuth's tokenization model [9]. Regulating access beyond a simple "accept or deny" scheme requires an extensible system to accommodate private web archives' need to tailor access to the resources. Wang suggested a role-based access control system stemmed on proximity in a social networking context for automated inclusion for access [26]. This approach is borrowed in my framework to regulate group access; e.g., when access is limited to those

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

in a proximity like within an IP address range, the authorization procedure need not be repeated but rather, an access token can be reused with a two-factor authentication-like scheme. This scheme can also be utilized to prevent access using this token beyond the IP address range.

PANDAS is a system developed by the National Library of Australia that provided tagging to web archives including restrictions by date (embargoes), authentication, and by IP address and is implemented via Apache's .htaccess file [18]. This system provides no fine-grain access control and suffers from other scale issues but was used as a basis for consideration in OpenWayback's implementation of access control[3]. Niu examined the Australian PANDORA archive among ten other web archives to compare the functionality and personalized-based features offered to users for personal web archiving [17]. These features included comparing web archive access methods such as lookup-by-URI as one one method offered to users. Access in terms of means of lookup will be investigated in the context of private web archives in this research, for which the URI clashing issue remains. Rao's iProxy provided users a means of archiving and replay with access parameters that extended URLs with commands for retrieval [19]. Because of the URI clashing issue, a similar extension of URIs will be needed for lookup in private web archives whose content was behind authentication on the live Web.

OpenWayback, the IIPC-sponsored open source version of the archive replay software that powers IA's Wayback Machine, provides limited access control for web archives that use their software as a basis (whether public or private). Through a collection-based scheme, OpenWayback's "AccessPoints" allow regulation of users who should have access and interact with a collection in the archive[4]. Each AccessPoint specifies an access URL and interfaces for querying and replacing archived pages. The documentation on how to accomplish this is sparse though instances exist (e.g., at University of North Texas[5]) where restrictions such as limiting access by IP address range have been deployed. The "Access Point Adapter" described in an older version of the OpenWayback codebase serves as a foundational model for the Private Web Archive Adapter (Section 3.4).

Abrams described a bookmarking system he labeled as personal/private "Archiving" but described a preservation by-reference approach where contemporary archiving is preservation by-value in addition to maintaining a reference key for lookup and replay [1]. He reiterated this point with the admittance that "bookmarks aren't great describers of the actual content [of the Web page]" reinforcing the link rot that occurs when a representation for a URI has gone stale.

Thelwall explored the bias of the collection of web sites preserved by Internet Archive as a selection of the "whole Web" [24]. This evaluation did not extend to the private live Web for which an even larger bias exists, as the overwhelming majority of content preserved by IA is from the public live Web. Gomes evaluated biases in web archive corpora that occur when the process of choosing which sites to archive in focused crawls is automated with a criteria basis [7]. His consideration of the user in developing access models is relevant in the user-based access models being developed for aggregating private and public web archives.

Marshall enumerated examples of personal digital archiving extending beyond web archiving [15, 16]. The usage patterns give real-world scenarios of how individuals preserve and access their digital content including the distribution of collections, what sort of content is preserved, and the role of the storage medium in ensuring future access. With the audience of this framework ultimately being these same amateur archivists, Marshall's patterns help to understand the technical needs of the users in developing the framework.

In my previous work I highlighted and evaluated the digital preservation capabilities of tools used to preserve content on the live web, particularly in respect to JavaScript [13, 10, 11, 4]. These works accounted for archiving content on the public live web though much of the private live web is dynamic and JavaScript-driven, proving the likelihood of a higher degree of damage in mementos [6, 3]. I have preliminarily used browser-based tools [14] for a subset of the web archives I created from the private live web to generate private web archives.

Strodl described a user-driven framework for digital preservation that facilitates individuals' preservation of private digital content using best practices [23]. Their software prototype predates and shares similarities with my prototype [12] to encourage users to archive their private web content by removing technical barriers in the preservation software. Strodl's work abstracts the access issues that will need to be addressed when the implementation of the framework creates data akin to the sort he describes.

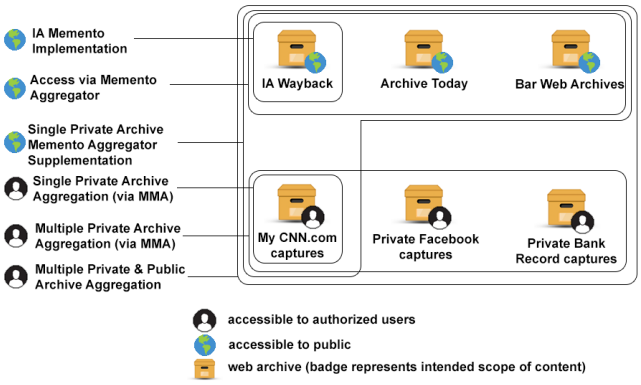

Figure 3: Various currently existing entities in the web archiving spectrum are limited to a small part of what can be preserved. Illustrated here are the additional entities of the framework and their scope relative to the currently existing entities.

3. FRAMEWORK DYNAMICS

The goal of this framework will be to provide a means of controlling access for aggregation of the contents of public and private web archives. The example for aggregation I use is based on a returned Memento TimeMap where the usual exhibition of aggregation is simply between multiple public web archives. My framework supplements these results into a TimeMap potentially consisting of the results from

[3] https://web.archive.org/web/20090209140507/http://webteam.archive.org/confluence/display/wayback/Exclusions+API

[4] https://github.com/iipc/openwayback/wiki/Release-History

[5] http://sourceforge.net/p/archive-access/mailman/message/32026372/

an aggregator amended with results from private content from web archives (e.g., captures of my private Facebook news feed accessible on the live web through authentication), public content from private archives (e.g., a user's CNN.com captures), and Memento compliant public web archives not included in an aggregator, as configured. The role of each entity required to achieve this functional hierarchy first requires that each entity's contribution to be defined (see Figure 3).

3.1 Private Web Archive (PWA)

A private web archive (PWA) in this proposed framework constitutes a collection of web pages captured at a certain time for which some consideration of access control should be applied. The rationale for needing access control is not a factor so as to scale for scenarios such as containing sensitive information, limitation of access based on IP address range, or any number of reasons for segregation from other web archives as a default functionality. Current public facing Memento aggregators do not read from these archives. These private web archives may also contain private captures of public live web content (e.g., a user's CNN.com captures) that can be exposed if queried while simultaneously restricting access to other content (e.g., a user's web-based bank statements).

3.2 User

Identifying the "user" in the flow of the framework allows the authentication and access procedures to be more intuitive and descriptive. Memento is primarily accessed through a user-agent, sometimes a web interface to a Memento aggregator but more frequently via a web browser extension built to interface with both Memento aggregators and archives directly using the Memento communication patterns. In this framework, the user-agent and user are considered synonymous. A user may query:

- A Memento aggregator with a URI-R
- A Memento meta aggregator (Section 3.3) directly with or without credentials to be relayed to archives or in the case of a very concentrated query without aggregation
- A Private Web Archive Adapter (Section 3.4) directly with credentials.

3.3 Memento Meta Aggregator (MMA)

A Memento meta aggregator (MMA) serves as a superset of functionality of a conventional Memento aggregator. Beyond providing access to TimeGates and TimeMaps for a set collection of web archives, an MMA also provides the ability to supplement the results of an MA with additional web archives on request and reference. These other web archives may be public non-aggregated Memento-compliant web archives or private web archives as relayed through a private web archive adapter. Further, an MA is not required to access an MMA at all but can return aggregated results based solely on a set of archives in the set of web archives for which it has been configured or provided upon query. This abstraction provides a level of extensibility to current Memento aggregators for which the additional functionality may not be appropriate, scalable, or interoperable.

As an endpoint, MMAs can also relay credentials to the authorization layer for private web archives (PWAs) and subsequently route the appropriate token to corresponding web archives (private or public) on queries after authentication has been established. Further, MMAs can query other MMAs with the expectation that the results returned will be consistent with those from an MA with additional indicators for content beyond the scope of an MA (e.g., a flag for content from a non-aggregated or public archive).

3.4 Private Web Archive Adapter (PWAA)

A private web archive adapter (PWAA) serves as the entity that regulates access to the PWA. Different access patterns (Section 4) can be used in the implementation of a PWAA but the primary use case consists of setting up persistent access using tokenization to remove the need for reauthorization on each query. PWAAs can also regulate access to a collection of private web archives via ad hoc subsetting (e.g., tagging specified URIs from a set of web archives) producing a "key" for the subset to be used on re-query so the potentially expensive subsetting does not have to again be established. A PWAA's primary interface is via requests from MMAs relaying requests from users.

4. USAGE PATTERNS

Sample usage patterns help to verify the validity of the flow of access from a variety of hierarchical schemes, represented as a composite in Figure 4 and described piecemeal here.

4.1 Hierarchical Entity Interaction

Basic usage (Pattern 1) consists of Linda simply accessing a web archive (e.g., Internet Archive) directly, as shown with the sole user accessing Archive 8. This requires no aggregator but serves as a base case.

The next abstraction (Pattern 2), which is possible in the Memento usage pattern today, is for Linda to access an MA. In the diagram, a user accesses a public aggregator, which aggregates only public web archives (Archives 4, 5, 6, and 7). Beyond these high level patterns resides my contribution.

The following pattern involves access to an MMA instead of simply an MA. In the base case, MMAα (Pattern 3) relays a request for mementos for a URI-R from Linda to the aforementioned MA. The MA treats the request as a request by a user (by design), performs the query, and returns the results to the MMAα , which in turn relays the results to Linda.

The diagram shows the case (Pattern 4) where MMAα is aware of a Memento compatible public web archive of which the MA is either not aware or does not aggregate by default. Along with relaying the request from Linda for mementos for a URI-R to the MA, the same request is sent to Archive 8 from MMAα. Upon response from both the MA and Archive 8 for mementos for a URI-R, MMAα aggregates these results and returns them to the user.

MMAs can execute this pattern recursively, querying other MMAs in the same way that an MA would be queried, further emphasizing the expected polymorphic behavior of MMAs in the hierarchy (Pattern 5). This might occur if Linda were to setup an MMA of her friends' captures, which accesses an MMA of her town's captures, etc. with each MMA supplying both scope of which archives are aggregated as well as potentially ultimately aggregating with mementos from an MA.

Adding in the additional query to PWAAs from MMAs shows how this relay of queries from one MMA to another

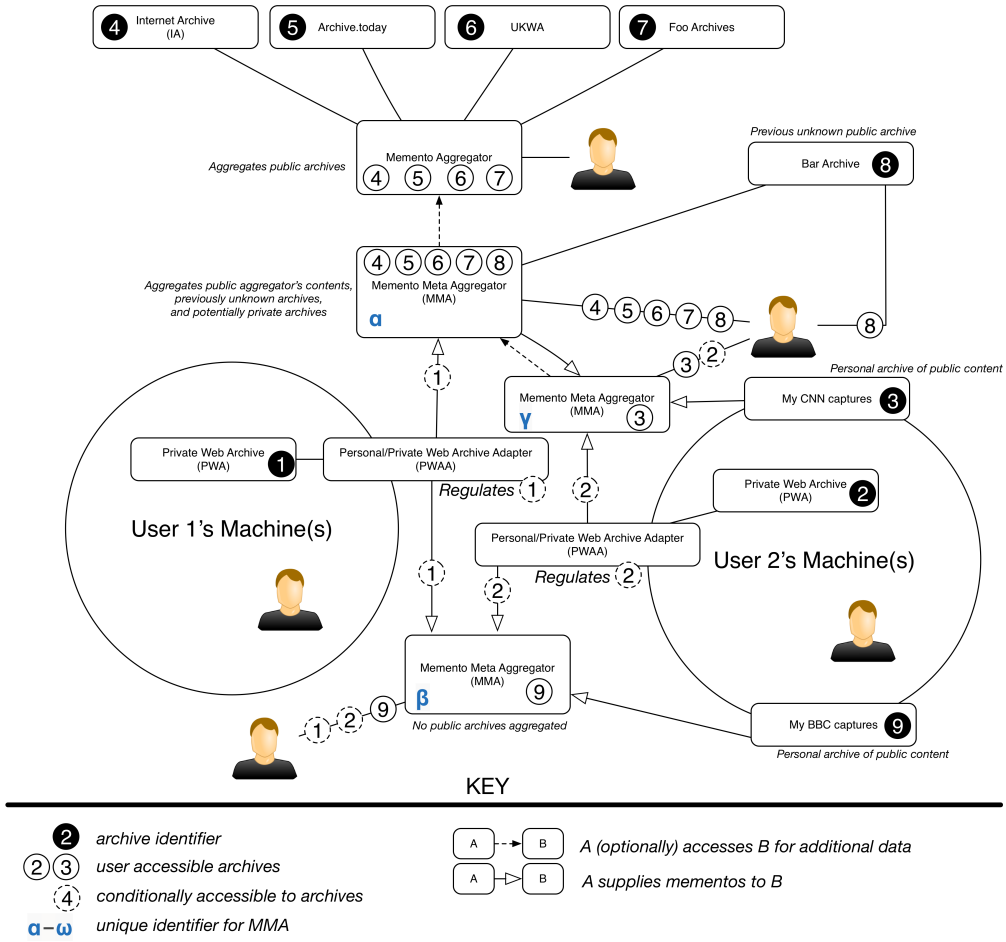

Figure 4: The composite hierarchy diagram displays various entities from Section 3 and usage patterns from Section 4 to effectively regulate access to private web archives for aggregation with public web archives without changing the functionality of the infrastructure in-place (e.g., Wayback deployments, Memento aggregators, etc).

Table 1: A user can establish access to a set of private web archives by communicating and providing credentials to an MMA. This is an example of Pattern 7 in Section 4.

FROM	TO	COMMUNICATION
user	MMAβ	"I want persistent access & what do I need to do?"
MMAβ	user	Supply me with your credentials
user	MMAβ	Here are my credentials! Give me persistent access!
MMAβ	PWAA1	"A user wants persistent access & here are the credentials"
MMAβ	PWAA2	"A user wants persistent access & here are the credentials"
MMAβ	PWAA3	"A user wants persistent access & here are the credentials"
PWAA1	MMAβ	"Access granted & token: abc"
PWAA2	MMAβ	"Access granted & token: def"
PWAA3	MMAβ	Access denied
MMAβ	user	"{PWAA1: abc & PWAA2: def & PWAA3: null}"
user	MMAβ	"Give me mementos for foo.com {PWAA1: abc & PWAA2: def}"
MMAβ	PWAA1	PWAA1: Give me mementos for foo.com (token: abc)
MMAβ	PWAA2	PWAA2: Give me mementos for foo.com (token: def)
PWAA1	MMAβ	Here are the results (50 mementos)
PWAA2	MMAβ	Here are the results (20 mementos)
MMAβ	user	"{PWAA1: {50 mementos} & PWAA2: {20 mementos}}"

and eventually to an MA can be useful for "decorating" results with the additional memento from archives. In Figure 4, Linda queries MMAγ, which gets its results from three sources: MMAα, Archive 3, and potentially Archive 2 (Pattern 6). Whether the query is sent from MMAγ to the PWAA is based on authorization parameters passed from

the user at the discretion of MMAγ. Some of Linda's friends may be willing to share their Facebook captures while others may not. Depending on which credentials Linda sends to her MMA, which are relayed to each PWAA for her friends' PWA, will allow a subset of archives' mementos to be aggregated.

The aggregation of Archive 3 to MMAγ (Pattern 7) demonstrates private web archives for which Linda has configured to be publicly exposed. Were she to not pass any credentials to MMAγ, the lack of specification could either be relayed to the PWAA that regulates access to Archive 2 or not relayed from MMAγ to the PWAA at all based on the design of MMAγ. If empty or insufficient credentials are passed from MMAγ to the PWAA, PWAA would respond with either an HTTP 403 (Forbidden) or a message indicating "0 mementos" to discourage further requests without exposing the state of the authentication (for security). A TimeMap for a query to an MMA, which aggregates the results from a Memento aggregator, an un-aggregated public web archive, and a private web archive is illustrated in Figure 5.

Being a functional superset of MAs, MMAs do not rely on obtaining results from an MA but rather can potentially query results from a disjoint set of archives wholly consisting of PWAs. When a user queries MMAβ with no or insufficient credentials for any of the PWAAs of which MMAβ is aware, only results from Archive 9 are returned, much like Pattern 7. Were credentials passed on a per-archive basis from the user to MMAβ for the PWAAs regulating Archives 1 and 2 (or any number of other PWAAs), whether results were returned and subsequently aggregated by MMAβ with URI-Ms from Archive 9 are based on the authentication result from each PWAA. If the credentials for accessing the content in Archive 1 are sufficiently met, as determined by the PWAA, and those for Archive 2 incorrect, as determined by that archive's respective PWAA, the results returned to a user querying MMAβ could consist of either results from solely Archive 9; Archives 1 and 9; Archives 2 and 9; or all three archives for which MMAβ is aware: Archives 1, 2, and 9.

The set of access patterns can be summarized as follows. More patterns likely exist and will be explored during the course of the research.

Pattern 1: user accesses web archive directly.

Pattern 2: user accesses memento aggregator.

Pattern 3: user accesses memento meta aggregator, which accesses a memento aggregator.

Pattern 4: user accesses memento meta aggregator, which accesses a memento aggregator and an additional previously non-aggregated archive.

Pattern 5: memento meta aggregator accesses a memento meta aggregator for results.

Pattern 6: user sends credentials to a memento meta aggregator, which queries then potentially aggregates results from private and public web archives.

Pattern 7: user queries memento meta aggregator with no or insufficient credentials but still retain access to publicly exposed content in a private web archives.

4.2 Establishing Access

In Section 4.1 I described a user simply submitting credentials to an MMA and these being relayed to the PWAA as the authority on whether access to the target PWA should be granted. Passing credentials to a service repeatedly has security implications, which I mitigate by using OAuth tokenization. Upon initially relaying credentials, an MMA also may send (in the same request) a request for persistent access. After a token has been established from a user to a PWAA via an MMA, a user can use this token to directly query the archive for getting non-aggregated results.

Table 1 illustrates a query from a user to a single MMA (Pattern 7), which serves as the aggregator to three private web archives. Upon rejection of the credentials, the user opts to not re-query and instead sends indicators implicitly selecting a subset of available archives based on the sets of tokens passed in.

5. WORK PLAN

In previous research, I have identified content that is problematic to preserve from the live web, built tools to capture a portion of content that previously was not preserved, and formulated a per-resource metric to evaluate the important of content that is difficult to preserve. Preserving this content is only a first step in replicating the archived web experience in a manner that includes privates web archives. Further work is needed to investigate modular and extensible approaches for access control so as to not couple with standards like OAuth when the nature of private web archiving demands customization. The entities built on top of the Memento framework (Section 3) will need to account for scalability issues and additional caveats that will arise with the additional utilization of the infrastructure in-place to aggregate the archived public web. Because the previous evaluation of resource importance only took into account public web archives, the importance of resources captured from the private live web will likely vary, so repeated experiments to evaluate content that is much more difficult to capture will need to be performed.

Integrating public and private web archives have an inherent problem of URI clash. Because URI alone is an insufficient parameter (Section 1) with accessing content on the private live web, replaying this content on the archived web (containing both public and private live web data) will require a deeper abstraction of access to reliably query content to replicate the experience from the live private web. Additional usage patterns (Section 4.1) very likely exist when the additional entities for controlled access and aggregation are applied to real world scenarios. A user study will assist in accounting for more of these situations and make the hierarchy more robust and useful for wide-scale application.

Based on the issues previously enumerated, I wish to address the following research questions in the course of my thesis.

- What sort of content is difficult to capture and replay for preservation from the perspective of a web browser?
- How do web browser extension APIs compare in potential functionality to the capabilities of archival crawlers?
- What issues exist for capturing and replaying content behind authentication?
- How can content that was captured behind authentication signal to web archive replay mediums that it possesses this characteristic for special handling?
- How can Memento aggregators indicate that private web archive content requires special handling to be

```
...
, <http://web.archive.org/web/20150228155703/https://facebook.com/>;rel="memento";
       datetime="Sat, 28 Feb 2015 15:57:03 GMT"
, <http://web.archive.org/web/20150228163939/http://www.facebook.com/>;rel="memento";
       datetime="Sat, 28 Feb 2015 16:39:39 GMT"
, <http://web.archive.org/web/20150303162841/https://www.facebook.com/>;rel="memento";
       datetime="Tue, 03 Mar 2015 16:28:41 GMT"
, <http://users2machine.local/web/20150305000101/https://www.facebook.com/>;rel="memento";
       datetime="Thu, 05 Mar 2015 00:01:00 GMT"; key="e395935019ee467c797034ee410cc91e"
, <//wayback.archive-it.org/all/20150305215922/https://facebook.com/>;rel="memento";
       datetime="Tue, 05 Mar 2015 21:59:22 GMT"
, <http://previouslyUnaggregated.org/web/20150306123457/https://www.facebook.com/>;rel="memento";
       datetime="Wed, 06 Mar 2015 12:34:57 GMT"
, <http://web.archive.org/web/20150310140721/https://www.facebook.com/>;rel="memento";
       datetime="Tue, 10 Mar 2015 14:07:21 GMT"
...
```

Figure 5: An example partial Memento TimeMap from a Memento meta aggregator (Section 3.3) contains contents from Internet Archive and Archive-It (which resemble Figure 1), a capture from the user's private web archive as well as a capture from an unaggregated yet public web archive. The unaggregated archive supplements the results from a Memento aggregator while the additional key for the private web archive can be utilized for access. This TimeMap would be returned after access has been established via a Private Web Archive Adapter (Section 3.4).

replayed, despite being aggregated with publicly available web archive content?

- What kinds of access control do users that create private web archives need to regulate access to their archives?

Progress for this research can be evaluated based on the preservation and access goals of this research. A timeline of prior, current, and upcoming progress will help steer the research to completion.

Prior publications in my doctoral research have dealt with the aspects of preservation [14, 12, 4] and evaluation [3, 11, 13] aspects of the framework. Future work will be focused on the access portion of the framework.

Following the JCDL Doctoral Consortium in June 2015, I expect to implement the basic MMA and PWAA functionality from May to July 2015. From June to August 2015 I will begin data collection for creating corpora resembling archives 1, 2, 3, 8, and 9 in Figure 4. In August 2015 I will perform my PhD candidacy proposal after refining this work from the feedback obtained at the doctoral consortium and executed over the Summer 2015. In Fall 2015 I will develop a publication title "Evaluation of User Access Patterns for Private Web Archives", which will be a user study to account for more real world private web archiving scenarios beyond those in Section 4.1. This work will additionally explore access control via tagging, role-based control, and other methods in practice for other realms beyond archiving that use similar control. The target submission conference is JCDL 2016, to be submit in mid-January for late June 2016 presentation.

In Fall 2015, I will program support for my initial framework definition into software tools I have previously created for personal web archiving (WARCreate, WAIL, and Mink).

For TPDL 2016 I will submit research titled, "Methods in adding JIT Inclusion of Private Web Archives in Memento" that explore the pros and cons of adding key-like features in TimeMaps (Figure 5). The submission of this work will be in late February for September presentation.

For the ACM Symposium on Access Control Models and Technologies (submitted early Feb 2016, presented in June), I will submit a publication exploring the OAuth-like tokenization and similar methods for access establishment (Section 4.2).

For iPres 2016 (submitted in April 2016, presented in October), I will submit research investigating URI clash and other needed identifiers for distinguishing archived content from the "deep web" with archived content from the public live web.

In late 2016, I plan to defend my dissertation and submit an abbreviated version detailing the work to the International Journal on Digital Libraries.

6. EVALUATION

Evaluation of this research will largely consist of determining the effectiveness of the hierarchy (graphically represented in Figure 4) in addressing the issues and research questions enumerated in Sections 1 and 5 (respectively). Further research is required in the sorts of access control needed in currently deployed private web archives that serve as barriers in protecting the content at the expense of integration with private web archives. The scalability of adding a layer of abstraction on top of currently deployed Memento aggregators will require concrete performance evaluation to determine how to effectively supplement the results aggregated from public web archives with those from private web archives. Quantitative success of the hierarchy can be tested when the scenarios described in Section 4 can be executed with the expected results returned. Correctness of the expected results will need to be determined to establish a baseline to differentiate unexpected results and to account for variations in the fluctuating availability of various public and private web archives.

7. CONCLUSION

In this work I plan to extend the Memento framework to allow for the controlled access and aggregation of content in private web archives with content in private archives. I defined the Memento meta aggregator and Private Web Archive Adapter, two entities necessary to provide this aggregation as well as extend the current functionality of public Memento aggregators, which are tailored to work solely

with public web archives. I defined various usage scenarios for content that would reside in public web archives and how a user might use the framework for aggregation. I regulated access using an OAuth-based tokenization scheme to allow for extensibility and the ability to share access to private web archives to a set of users.

8. ACKNOWLEDGEMENTS

Mat Kelly's advisor is Michele C. Weigle. Michael L. Nelson has provided additional guidance, direction, and expertise in the realm of web archiving. This work is supported in part by the National Endowment for the Humanities (NEH) DHIG (HK-50181-14).

9. REFERENCES

[1] D. Abrams, R. Baecker, and M. Chignell. Information Archiving with Bookmarks: Personal Web Space Construction and Archiving. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 41–48, 1998.

[2] A. AlSum, M. Weigle, M. Nelson, and H. Van de Sompel. Profiling Web Archive Coverage for Top-Level Domain and Content Language. *International Journal on Digital Libraries*, 14(3-4):149–166, 2014.

[3] J. F. Brunelle, M. Kelly, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. Not All Mementos Are Created Equal: Measuring The Impact Of Missing Resources. In *Proceedings of JCDL*, pages 321–330, London, England, 2014.

[4] J. F. Brunelle, M. Kelly, M. C. Weigle, and M. L. Nelson. The Impact of JavaScript on Archivability. *International Journal on Digital Libraries*, pages 1–23, 2015.

[5] J. F. Brunelle and M. L. Nelson. An Evaluation of Caching Policies for Memento TimeMaps. In *Proceedings of JCDL*, pages 267–276, 2013.

[6] J. F. Brunelle, M. C. Weigle, and M. L. Nelson. Archiving Deferred Representations Using a Two-Tiered Crawling Approach. Submitted for publication.

[7] D. Gomes, S. Freitas, and M. J. Silva. Design and Selection Criteria for a National Web Archive. In *Research and Advanced Technology for Digital Libraries*, pages 196–207. Springer, 2006.

[8] D. Hardt. The OAuth 2.0 Authorization Framework. IETF RFC 6749, October 2012.

[9] M. Jones and D. Hardt. The OAuth 2.0 Authorization Framework: Bearer Token Usage. IETF RFC 6750, October 2012.

[10] M. Kelly, J. F. Brunelle, M. C. Weigle, and M. L. Nelson. A Method for Identifying Personalized Representations in the Archives. *D-Lib Magazine*, 19(11/12), Nov/Dec 2013.

[11] M. Kelly, J. F. Brunelle, M. C. Weigle, and M. L. Nelson. On the Change in Archivability of Websites Over Time. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL)*, pages 35–47, Valletta, Malta, 2013.

[12] M. Kelly, M. L. Nelson, and M. C. Weigle. Making Enterprise-Level Archive Tools Accessible for Personal Web Archiving Using XAMPP. Poster and demo presented at Personal Digital Archiving, February 2013.

[13] M. Kelly, M. L. Nelson, and M. C. Weigle. The Archival Acid Test: Evaluating Archive Performance on Advanced HTML and JavaScript. In *Proceedings of JCDL*, pages 25–28, London, England, September 2014.

[14] M. Kelly and M. C. Weigle. WARCreate - Create Wayback-Consumable WARC Files from Any Webpage. In *Proceedings of JCDL*, pages 437–438, Washington, DC, June 2012.

[15] C. C. Marshall. Rethinking Personal Digital Archiving, Part 1. *D-Lib Magazine*, 14(3/4), Mar/Apr 2008.

[16] C. C. Marshall. Rethinking Personal Digital Archiving, Part 2. *D-Lib Magazine*, 14(3/4), Mar/Apr 2008.

[17] J. Niu. Functionalities of Web Archives. *D-Lib Magazine*, 18(3/4), Mar/Apr 2012.

[18] M. Phillips. PANDORA, Australia's Web Archive, and the Digital Archiving System that Supports It. `http://pandora.nla.gov.au/pandas.html`, 2003.

[19] H. C.-H. Rao, Y.-F. Chen, and M.-F. Chen. A Proxy-based Personal Web Archiving Service. *SIGOPS Oper. Syst. Rev.*, 35(1):61–72, Jan. 2001.

[20] A. Rauber, M. Kaiser, and B. Wachter. Ethical Issues in Web Archive Creation and Usage-Towards a Research Agenda. In *8th International Web Archiving Workshop (IWAW08)*, 2008.

[21] D. Rosenthal. Re-thinking Memento Aggregation. `http://blog.dshr.org/2013/03/re-thinking-memento-aggregation.html`, 2013.

[22] T. Schwarz, M. Baker, S. Bassi, B. Baumgart, W. Flagg, C. van Ingen, K. Joste, M. Manasse, and M. Shah. Disk Failure Investigations at the Internet Archive. In *Work-in-Progess session, NASA/IEEE Conference on Mass Storage Systems and Technologies (MSST2006)*, 2006.

[23] S. Strodl, F. Motlik, K. Stadler, and A. Rauber. Personal & Soho Archiving. In *Proceedings of JCDL*, pages 115–123, 2008.

[24] M. Thelwall and L. Vaughan. A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research*, 26(2):162–176, 2004.

[25] H. Van de Sompel, M. Nelson, and R. Sanderson. HTTP Framework for Time-Based Access to Resource States – Memento. IETF RFC 7089, December 2013.

[26] T. Wang, M. Srivatsa, and L. Liu. Fine-Grained Access Control of Personal Data. In *Proceedings of the 17th ACM Symposium on Access Control Models and Technologies*, pages 145–156, 2012.