Introduction

This report was prepared by a subgroup of the Yale Web Archiving Working Group (WAWG) to examine digital preservation considerations for captured websites, in collaboration with the Yale University Library Digital Preservation unit. The report provides an overview of needs and considerations for long-term preservation of captured web content to inform the Yale WAWG’s further investigations into and development of recommendations pertaining to infrastructure, policy, processes, and staffing to support web archiving activities at Yale. Along with a discussion of technical issues for website preservation, the report also outlines recommended administrative considerations relevant for supporting a broad base of participation in web archiving activities. Particularly as the WAWG’s strategic goals include developing and supporting a community of practice for web archiving at Yale, which could therefore increase collecting activities by community members who are not affiliated with library/museum repositories, we recommend further discussion of these administrative considerations for preservation as an integral component of the WAWG’s development of a community of practice.

Summary of Preservation Challenges, and Recommended Next Steps

- Increasing participation in web archiving from new “collections owners” from non-repository areas of the University could lead to the accrual of collections lacking designated or secure resources (staffing, funding, policies, etc.) to support these materials’ long-term management and preservation
  - Recommendation: For captured web content identified for long-term preservation, preservation plans should be created as a best practice, whether responsibility for materials remains with original contributors or is assigned to established repositories

- Archive-It (AIT) is widely used for capture of web content but explicitly excludes long-term preservation as part of the Archive-It service that Yale uses. Exporting materials from AIT produces Web ARChive (WARC) files that are often incompatible with the WARC viewer rendering in Preservica, Yale’s current Digital Preservation System (DPS), making long-term management more difficult by limiting one-click viewing of WARCs via the DPS interface. The Digital Preservation unit is planning to add support for AIT WARCs to Preservica in the future.
○ **Recommendation:** Add functionality to Preservica to enable it to render AIT WARC s. To aid in ongoing communication, we recommend creating a Digital Preservation Liaison role on WAWG.

- Captured websites stored in compliance with the WARC preservation standard format will contain a wide variety of file formats within them, requiring active monitoring and support strategies for accessibility of stored content
  - **Recommendation:** Monitor discussion within the web archiving community about further development of the WARC format, and strategies that institutions are evaluating for stewardship of web content/approaches for supporting formats within WARC files, such as implementation of software emulation environments, and/or format migration of files stored within WARC files.

- Yale’s website collecting activities include participation in the consortial Ivy Plus Web Resources Collecting Program. Addressing preservation requirements within consortial collections presents challenges in terms of identifying and sustaining management responsibility, but also provides significant opportunities to benefit from collaborative development of shared practices through this partnership.
  - **Recommendation:** Monitor the discussions and decision-making process of the Ivy Plus Web Resources Collecting Program as a source of expertise and leadership regarding recommended practices for preservation of captured websites. Recommend annual update from the Ivy Plus Web Resources Collecting Program regarding their preservation practices.

**Sections in this Report**

This report is organized into five sections:

1. [Captured Web Content at Yale](#) - an overview of already-captured web content and ongoing collection efforts across Yale University
2. [Technical Considerations and Requirements](#) pertaining to preservation of web content
3. [Preservation Capabilities of Current Systems](#) - evaluation of preservation capabilities of currently implemented systems for web archiving at Yale
4. [Administrative Responsibilities and Requirements](#) for Long-term Management and Preservation of Captured Web Content
5. [Appendices](#) - additional details on Yale’s currently implemented web archiving systems
Captured Web Content at Yale - Existing Collections and Ongoing Collecting Activities

Early Initiatives at Yale

The scope of web content collecting activities at Yale, and of the participants within these various activities prior to the commencement of the University’s Archive-It contract, is detailed in a document created in July 2016 by the Web Archiving Working Group, “Web Archiving Expertise at Yale.” As noted in that document, a range of collecting repositories and organizational units across the University have participated in initiatives to capture and preserve web content identified as within-scope for their areas. Some initiatives, such as the YUL Manuscripts and Archives unit’s capture of University sites via HTTrack from 2004-2008, have accumulated data requiring long-term preservation. Not all of these projects accumulated data that will require preservation by Yale, however. For example, the Yale Law Library’s ongoing participation in perma.cc as a program registrar provides support for the capture of web sources but is not responsible for preserving the captured data. Similarly, the Music Library’s participation in the Contemporary Composers Web Archive (2013- ), an Ivy Plus project led by Columbia University Libraries, does not involve preservation of data for the archived sites, as Ivy Plus will be responsible for the preservation of captured data.

Current Collecting Activities: Archive-It, Preservica, and Ivy Plus

Currently, web collecting activity at Yale primarily takes place among collecting repositories participating in the University’s contract with Archive-It (AIT), a hosted web harvesting service developed and provided by the Internet Archive. Yale’s AIT contract has been in place since 2015, when the Yale Center for British Art (YCBA), the Beinecke Rare Book and Manuscript Library (BRBL), and Yale Information Technology Services (ITS) worked together to obtain a contract between Yale University and Archive-It that allows for organizational units/repositories at Yale to subscribe to the web harvesting service separately, thereby allowing each repository to manage its own subscription for control over data allocation, user accounts, and access permissions. Yale ITS joined as an individual subscriber to AIT during the first year and has opted not to renew for subsequent years. Yale University Library’s Manuscripts and Archives (MSSA) joined the contract midway through the first year and has renewed each year. As of March 2018, Yale’s Archive-It subscription will continue for a third year, with three University units participating in the 2018 contract: Beinecke Library, YUL Manuscripts and Archives, and the Yale Center for British Art.
Content captured via the Yale Archive-It subscriptions largely reflects the collecting areas of the subscribing units, as well as content created by and pertaining to University activities. Examples of captured sites include Yale University and YCBA websites and social media sites, as well as websites and social media content selected by Beinecke Library curators and Manuscripts and Archives staff as relevant within their traditional collecting areas. **There is currently no workflow, mechanism, or cost model by which smaller departments or units within YUL (e.g., DASHRS) or Yale offices and departments can use Archive-It to do web archiving separately from the BRBL, YCBA, and MSSA accounts, although some of these units have expressed interest in using the service.**

Along with Archive-It, the YUL’s Digital Preservation System (DPS) (Preservica) provides another means by which collecting repositories can capture web content and a set of services to preserve WARC files captured using either toolset. The YUL DPS is still in relatively early phases of implementation, but there has been interest among DPS users in the system’s website capture capabilities. The YUL Manuscripts and Archives unit, although an Archive-It subscriber, would like to evaluate the DPS’s web ingest workflows as an alternate or additional means of capturing web sites, and is also interested in evaluating ingesting collections created in MSSA’s Archive-It subscription to the DPS for long-term preservation and management. The YCBA has also been actively building collections via Webrecorder as well as in Archive-It, and would like to ingest these materials into the DPS for long-term preservation. Beinecke is also planning to ingest its Archive-It WARC files into Preservica. The Yale Law Library has also begun testing web ingest workflows in the LAW tenancy of the DPS, and the YUL Library IT Department has also expressed interest in exploring the system’s capabilities for capturing LibraryIT departmental sites.

Along with collecting activities within the University, Yale will also continue participating in the Ivy Plus Web Collecting Program formed by a subset of 11 institutions in the Ivy Plus consortium. This group recently contributed funding to create the Ivy Plus Web Resources Collecting Program, directed by the Collections Development Subgroup and the Director of Collection Initiatives and in coordination with the Web Resources Collection Coordinator at Columbia, which will collaboratively build thematic web collections of common interest. As noted above in relation to the Contemporary Composers Web Archive project, Yale’s participation in the Ivy Plus Web Collection program does not include direct responsibility for storing the captured data. Storage and preservation of captured sites will be part of the Ivy Plus program’s overall work, however, and is discussed in the “Preservation Capabilities” section of this report.
Content Capture limitations of current tools

Both Archive-It and the DPS are limited in their scope and capabilities for website capture. The available tools are often unable to adequately capture the data that underpin interactive websites, such as the (often very large) databases that are accessed by web-servers to dynamically change the websites as users interact with them. This limits the content that WARC files contain, in turn limiting the ability of digital preservation practitioners and their toolsets to present a complete and authentic experience of the captured websites to future users. The Webrecorder tool attempts to mitigate some of these limitations by guiding an archivist through experiencing a website in order to capture all of the content-dependencies of that specific experience of the website, and does so to varying degrees of success.

The most comprehensive and therefore minimally-lossy capture method is to snapshot the original web-server computers, thus capturing all the content-dependencies of the target websites. This method is often limited in applicability due to the diverse and distributed nature of the resources that make up modern interactive websites. Any individual Facebook page, for example, is made up of content served to users via many different web servers, and a different set of web servers nearly each time the content is refreshed. Attempting to preserve the web-servers that make up a Facebook homepage then is practically impossible at this point in time. However this method can sometimes be successfully implemented, such as when capturing internal internet or intranet sites (e.g. the Yale.edu website) or other websites to which an archivist has server-level access, so it should not be completely dismissed.

Technical Considerations and Requirements for Preservation of Captured Web Content

To ensure the long-term accessibility of captured web content, preservation efforts need to satisfy the minimum requirements that are common across preservation for all digital objects, as well as to address additional challenges that may result from websites’ structural complexity, and the large variety of data types present within them. These minimum requirements therefore include support for bit-level preservation, as well as addressing risks from software and file format obsolescence.

Bit-level Preservation

At the lowest level, websites are composed of bits, which are the smallest unit of all digital
information. The precise physical arrangement of a vast number of bits in aggregate forms the specific digital content to be preserved, and ongoing stability of this physical arrangement is required to preserve its exact configuration and therefore the content it represents. Data storage, file fixity and data integrity, and information security activities all support bit-level preservation of digital content.

File Format and Software Obsolescence Issues

File formats specify how data are encoded, so that software can process and render the encoded information accurately and enable users to interact with it. While bit-level preservation ensures the integrity of data so that they are uncorrupted during ingest, storage, and delivery for access, it is also necessary to identify how the preserved data are encoded (i.e. what file formats are in use) and how they should be rendered as digital information (i.e. what software should be used to interact with them). Different types of content, such as text, images, or sound, are stored using different file formats, which require specific software to be able to interact with the information the files contain.

Websites are complex digital objects and pose a number of preservation challenges. They are usually composed of multiple files, and are often organized according to a hierarchical structure that is visually presented to the user when browsing a website’s source code, and which determines a user’s interactive path through the website when accessed using a web-browser, by conceptually linking files in relation to each other. Websites frequently include a wide range of content types presented via a variety of files structured using a variety of file formats, thus posing challenges for organizations seeking to implement recommended file format preservation practices such as working with a limited set of known open formats which are carefully inventoried and monitored.

Web Archive (WARC) format

The Web ARChive (WARC) format contains and documents the structure of captured websites by bundling multiple resources within the scope of a crawl as resource records and metadata within an aggregate archival file.¹ The WARC format was standardized by the International Internet Preservation Consortium which provides detailed publicly available documentation for it as an open standard, and community consensus around the use of WARC for packaging captured sites has led to the format’s widespread use. This bodes well for ongoing documentation and improvement of the standard. All of these factors increase the prospects for

¹https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/
sustainability of the WARC format for accessing stored web content in the future.

Yale’s primary web capture systems are in compliance with current recommended best practices for collecting and storing captured web content in the WARC format, but providing future access to the collected content will require support for formats within captured sites. All of the content captured within the University’s Archive-It subscriptions is collected and packaged as WARC files, and the YUL DPS provides configurable web ingest workflows that can either store content packaged as WARC files, or as a copied directory structure replicating the original files. Archive-It’s reporting capability provides data about MIME types of stored collections within an account, but does not currently provide the ability to monitor and report on specific file format types within WARC files. The YUL DPS characterizes file formats within WARC files as a by-product of its within-‘container format’ characterization tools. The DPS does not currently support automated reporting on these formats (ie these formats are not included in the DPS’s standard file format reports) as web content preservation in the DPS is conceptualized as being WARC-based, meaning that the object of preservation is the WARC. The WARC file format is currently not obsolete (it is well supported) and migration workflows have not yet been required in the DPS for use with WARCs.

One consideration pertaining to creating collections from WARC files within the YUL DPS is that playback of content within these files depends upon how the collections of WARCs are organized, which can affect rendering of the content across different systems. For example, collections within Archive-It may include multiple seed URLs, and playback will operate at the collection level. Additionally, Archive-It’s option to export WARC files from a collection produces a list of WARCs that are organized by crawl date, and which can also include multiple seeds. The WARC player currently available within the DPS may require identification of a single seed URL for a deliverable unit containing one or more WARC files to provide playback of the captured content within that deliverable unit. For this reason, playback capability within the DPS does not currently match the playback operation of collections within Archive-It.

Note that these current playback limitations do not prevent the WARC files from being preserved in the DPS, nor from being presented using other access tools and/or portals. It may be that playback within the DPS, while a convenient feature for browsing content contained in ingested WARCs, is not relevant for how the WARCs will be structured as deliverable units for long-term storage. More testing is recommended, however, to explore options for structuring WARCs exported from Archive-It as deliverable units, to determine recommended structures for
web collections within the DPS.

Additionally, even for content captured in the WARC format, there are difficulties posed by the diverse nature of the content within WARC files. Because websites contain a large number of content types, WARC files may contain resources in any format. The ability of WARCs to encapsulate any kind of content is essential for preserving Web content, as it is vital to capture and store files such as the images or audiovisual formats that are frequently embedded in HTML pages that contribute to the richness and diversity of content on the Web. The ability to capture and contain these files, however, does not mean that software will automatically be able to interact with the captured files in the future. The range of file types contained within the WARC files themselves will still be at a high degree of risk if the captured files were originally created as formats which are not as well-supported in the future.

To summarize, adherence to the WARC preservation standard format for capturing and storing web content is recommended, but is not in itself sufficient to ensure future accessibility of web collections. Ensuring accurate identification, representation, and organization of files included in websites is required for their long-term accessibility as content. A complete preservation policy for web content must account for the need to better understand and respond to the multiple layers of challenges within WARC files. Ongoing assessment of existing and new tools as they become available, and development and implementation of preservation workflows relative to these advancements is recommended.

Software obsolescence and Emulation

Software is a key component of the experience of websites. Websites are served to users via web-servers\(^2\) (such as Apache, NGINX, or IIS), software applications that provide the files comprising websites to users over a network (normally the internet) when requested via a web-browser or another software application. Web browsers are the software applications that are normally used to interact with websites stored on web-servers. Examples of web-browsers include Google Chrome, Firefox, Internet Explorer, Microsoft Edge, and Netscape Navigator.

Web server and browser functionality changes over time\(^3\). These changes can cause the

---

\(^2\)The term ‘Web-Server’ is also used to refer to the hardware that supports the web-server software.

\(^3\)Infamously, the web browser ‘Netscape Navigator’ supported the “Layer” element ([https://en.wikipedia.org/wiki/Layer_element](https://en.wikipedia.org/wiki/Layer_element)) for a very limited period of time/through a limited set of software versions, during which the website “The Unreliable Archivist” was rebuilt making use of that element, causing it to quickly become inaccessible when accessed using contemporary web browsers: [http://three.org/unreliablearchivist/1998/#](http://three.org/unreliablearchivist/1998/#)
content a user is presented with for interaction to also be changed or can cause access to the content to be lost altogether. This ‘software obsolescence’ is a huge challenge for the preservation of website content and the preservation of users’ ability to interact with and experience websites without any content loss or distortion.

Fortunately there are ways to mitigate the challenge of software obsolescence, both on the web-server side and the web-browser side. In both cases, the original software used to present the website experience to users can be preserved using standard bit-preservation techniques, then re-executed by running the software on ‘emulated’ versions of the computer hardware on which it originally ran. In the case of web servers it may be difficult to replicate the full web-server configuration without having captured it from the source. However, if the original files that composed a specific website can be captured, and the web-server software version information also captured, then it is possible to configure a ‘fresh’ version of the original web-server software that can serve up the website source files using an almost identical process as the original.

Similarly, if it can be ascertained which web browsers were generally used to interact with a website, then it is possible to preserve those applications and re-execute them on emulated versions of the computer hardware on which they originally ran. Access to web archives using original web-browsers made re-executable using emulation is currently being provided as a service by Rhizome at [http://oldweb.today](http://oldweb.today). Various users of the Emulation as a Service (EaaS) framework ([http://eaas.uni-freiburg.de](http://eaas.uni-freiburg.de)) are also able to provide access to older web browsers to enable interaction with older websites. In addition, the University of Freiburg (developers of the EaaS software suite) has recently simplified the process of connecting both an emulated web-server computer and an emulated computer running an end-user web-browser, enabling them to be accessed as an individual ‘object’ with a single click via a contemporary web browser.

**Preservation Capabilities of Current Systems at Yale**

The Archive-It service is not a preservation service. The terms of Yale’s agreement with the Archive-It team do not include preservation of captured websites.

The DPS can provide bit-preservation of websites and the EaaS service that the Library is currently implementing can ensure the software components of website experiences can be preserved and made accessible in the future.
Preservation Capabilities of the Ivy Plus Web Resources Collecting Program

Addressing the preservation requirements of web archives within consortial collections can pose challenges, given the different approaches to preservation that institutions may employ, and the need to identify ownership of collections and therefore management responsibility for archived content. A subgroup of the Ivy Plus Collection Development Committee will be exploring different possibilities, including an Ivy Plus LOCKSS Network and the Archive-It/DuraCloud integration option, for preserving websites captured as part of the Web Resources Collecting Program. At the moment, Yale’s local preservation decisions for archived websites are not being informed by the process Ivy Plus will be undertaking in the future. The formalized structure of the Ivy Plus Web Resources Collecting Program within the collections development effort of the consortium, and in particular the hiring of the dedicated Web Resources Librarian suggest that Ivy Plus initiatives will continue to provide valuable expertise and leadership in this area.

Administrative Responsibilities and Requirements for Long-term Management and Preservation of Captured Web Content

In addition to the technical considerations detailed above, clear and comprehensive administrative oversight will be required to provide effective long-term management and preservation for web collections. This section outlines considerations for establishing sustainable administrative oversight for the long-term preservation of captured web content at Yale.

As noted above, Web collecting activity at Yale has already been marked by a notable degree of cross-organizational cooperation and decision-making. As collecting activities continue to grow and spread, particularly with the WAWG’s development of a University-wide Community of Practice for web archiving, this work may involve participants from areas of the University that have not yet had an ongoing role as content contributors/depositors to the Library, Museums, or archival repositories⁴. Many of these participants may not yet have had experience as managers of collections requiring long-term administrative oversight and preservation.

Appendix C of this report discusses YUL’s digital preservation system as one model of a

⁴As an indication of some of the possible participants from across the University, see the Use Cases document created by the Web Archiving Working group.
centralized method for working with different content contributors. The management approach implemented in the YUL DPS depends upon and functions in relation to the existence of stable and ongoing collecting units who retain administrative responsibility for their materials, in combination with funding allocated for the permanent retention of stored content. Applying this approach to cooperative stewardship to support existing repositories’ collecting of web materials would only require building upon this established foundation in which the collecting units provide repository expertise via professional librarians/archivists/museum staff who make collections decisions, along with organizational continuity and financial responsibility, and the Digital Preservation unit provides systems support and preservation guidance for these collections.

Expanding support for web collecting activity in the YUL DPS to support the participation and contributions of University areas currently unaffiliated with already-defined YUL repositories/DPS units would require developing additional governance structure for administrative responsibility and decision-making. So far, the YUL DPS has been charged with and funded to support providing a long-term preservation environment for the collections of YUL, the Yale Center for British Art, the Yale Law Library, the Peabody Natural History Museum, and the Yale Art Gallery. If collections managers for web archives from areas of the university outside of these units were to become included users of the YUL DPS system for long-term management of content, this would be an expansion of the system beyond the scope of the current collection development policy and funding scheme, which would need to be clarified. For example, funding sources for storage of non-YUL individual collecting areas would need to be determined and provided over the long-term. Adding a significant number of funding sources in addition to the current YUL and collections-specific funding sources is technically possible, but could potentially increase the overall workload of the Digital Preservation unit. The system supports the configuration of additional funding sources, but in practical terms implementing non-consolidated funding streams from a broad range of University sources could add significant complexity and therefore staffing time required for the management of a complicated array of funding options, constituting a potentially significant cost that would be born by the YUL’s Digital Preservation unit, unless additional staffing or financial support was provided.

Additionally, even in working with established repositories/units, the expansion of YUL DPS activities to include support for web collections within the system would increase the overall workload of the YUL Digital Preservation unit, to include identifying and implementing system requirements, providing staff training, and implementing new workflows relative to web
collections. Also worth noting is that the YUL Digital Preservation unit and the Preservation Department overall are not currently participating in Yale’s Archive-It contract. If collections managers would like assistance from the Digital Preservation unit for long-term storage and management of web content within Archive-It and/or for exporting content from Archive-It into the YUL DPS for long-term management, it would be necessary for Digital Preservation staff to have access to and build expertise in the Archive-It service.

Appendices

The following Appendices provide further details on the preservation capabilities of Yale’s current web archiving systems (Archive-It and the YUL DPS), and a description of how administrative responsibility operates within the YUL Digital Preservation system.

Appendix A. Archive-It Web Harvesting Service (Hosted by the Internet Archive)

The Archive-It Service provides an approachable and low-barrier-to-entry option for organizations to begin collecting web content. The convenience of the service and the dearth of better, lower-cost options has had the effect of consolidating the majority of web collecting activity within this service. From a preservation perspective, the widespread use of the Archive-It service represents a potential strength and a challenge. Organizational subscriptions to the Archive-It service include storage of captured web content, and at present all of Yale’s web content captured via Archive-It is also stored by the service, but the storage options included within the Archive-It service contract do not include enough details about preservation practices to constitute a reliable preservation repository.

Archive-It provides a couple options for better-documented storage of content captured using the service. For subscribers willing to pay an additional fee and to rely on off-site managed preservation storage, Archive-It offers subscribers the option of an integrated backup of captured content to DuraCloud. Another route is for subscribers to download captured content as WARC files from Archive-It’s servers and manage their own preservation storage for the content within their institutions’ existing storage infrastructure. The Harvard environmental scan includes comments from a number of institutions that elect to follow this practice. For long-term storage of web content captured via Archive-It, automated downloading of content could be configured from within Yale’s Archive-It subscriptions; downloaded content should then be managed within a preservation repository. More testing is required, however, to understand how
options for configuring crawls affect the interoperability and rendering of exported WARCs with the YUL DPS.

The need for more transparency about Archive-It’s storage specifications is a concern shared by other subscribers. For example, the 2016 Harvard environmental scan of web archiving\(^5\) includes comments from a number of institutions concerned about reliance on Archive-It storage alone for captured web content, and emphasizing the need for more specifics about Archive-It’s preservation details. As a service with an active community of users, Archive-It may respond to organized/aggregated user requests for increased transparency about the service’s preservation practices. If so, long-term storage of content within Archive-It could be evaluated more fully; with more details about AIT’s preservation conditions, cost and comprehensiveness of storage within the service could be assessed in comparison with the time, effort, and overall cost required for exporting content from the service for long-term storage within Yale infrastructure. Archive-It’s Storage and Preservation Policy is available online at the Archive-It User Guide, and was last updated in March 2018, suggesting that the policy is actively maintained and that Archive-It may be receptive to requests from institutional subscribers to increase overall transparency about preservation practices.

Appendix B. Yale University Library’s Preservica Digital Preservation System

The YUL DPS also provides website capture capabilities via website ingest workflows that users can configure to define a crawl for a single seed URL. Webcrawls created in these ingest workflows use the same Heritrix web crawler software that is used by Archive-It, but the method of creating collections is quite different. Whereas Archive-It provides a dashboard from which users can build collections that are flexible in terms of the number of seed URLs included, website ingest workflows in Preservica execute a crawl on a specified seed URL, save the captured content in the WARC format (or as files if this option is configured in the ingest workflow context), and automatically ingest the crawled site into the system as a deliverable unit for long-term management.

Once the automated ingest process completes as part of a website ingest workflow, WARC files can be managed in the DPS similarly to other content. Preservation conditions such as storage media types, redundancy of copies, storage locations, and integrity checking for saved WARC files can be determined at the point of ingest, and updated at any point afterwards. Disk and

\(^5\)https://dash.harvard.edu/handle/1/25658314
tape storage options in the DPS are provided by Yale’s ITS Storage@Yale service, so additional specifications for evaluating these data storage environments are provided according to ITS policy, which may therefore offer greater transparency than relying on vendor-provided details on data storage (as in the case of Archive-It). As noted above, however, automated file format monitoring and migration within WARC files is not currently possible in the DPS.

In terms of information security, the YUL DPS can provide logs for actions performed on files, including deletions and preservation actions. The DPS also provides the capability to identify and restrict user access permissions to read, write, move, and delete content, at the level of individual files. Access permissions for DPS units are currently documented for each unit in spreadsheets that are saved in Yale’s Box file sharing and storage workspace service and shared with the DPS unit managers. The unit spreadsheets record the users for that unit who have service accounts created in the system, and the unit-specific roles that are assigned to each user, along with permissions assigned to the unit’s content tags.

**Preservica web archiving capabilities – Known limitations, and suggestions for further testing by WAWG Tools subgroup, and/or the YUL Digital Preservation unit**

1. Limitations on format reporting and migrations for WARC files. The YUL DPS scans WARC files upon ingest to determine formats inside the archive file, but these contained files are not included in overall reporting of formats within the system, and system preservation policies are not applied to these files.

1. Current limitations on rendering/playback of WARCs in the DPS. To render ingested WARC files in the DPS, a base URL is provided for the WARC or group of WARCs as a deliverable unit. WARCs exported from AIT can sometimes have multiple base-URLs, however, which may limit their ability to be rendered within the DPS. Further testing of AIT crawl configuration and export of WARCs is recommended, to better understand how Archive-It organizes and renders collections, and how export and packaging of WARCs from AIT to the DPS should be managed.

2. Relative to current playback limitations of WARCs in the DPS, if the DPS is to be used for storing and preserving WARC files exported from another capture system such as ArchiveIT, additional rendering options should be explored. This could include a future access system for the DPS that has a player for rendering WARC files, in which case WARC files could be stored as preservation copies, and the player in the access system could manage rendering issues. Additionally, the Webrecorder project’s WARC player
could be explored as a viewer option for WARC files stored in the DPS.

3. Possibilities for collaboration between WAWG and the YUL Digital Preservation unit. What role will the WAWG play in testing/documenting/advising on these functions of the DPS? Will the Tools subgroup of the WAWG work with Digital Preservation Librarians in the Digital Preservation unit to test these DPS functions?

Appendix C. Administrative Responsibility within the Yale University Library Digital Preservation System

This section describes the assignment and operation of administrative and financial responsibility for digital materials within the YUL Preservica DPS as an example of how decision-making currently functions between repositories across the Library and Museums, and the Yale University Library organizational unit tasked with providing long-term preservation services for these repositories in managing and preserving their content.\(^6\)

The YUL DPS is managed by the YUL Preservation Department’s Digital Preservation unit, with support from YUL’s Library IT Department, and the infrastructure and service the DPS provides are shared by individual units within the Library. The YUL Digital Preservation unit creates policy recommendations to inform repositories’ use of the system and preservation services, and has defined organizational units in the Digital Preservation System based on the established organizational structure of repositories within YUL, the Law Library, and Yale Museums. The organizational structure of DPS units in the system determines which staff members have access to the system, and which workflows and actions they can execute within it, as well as how ingested materials are located within the overall collections hierarchy of the system, and access permissions for these materials.

Within this organizational structure, YUL Digital Preservation works with individual units to provide training and assistance for ingesting content, and for managing content over the long term. Individual units retain ownership and sole decision-making responsibility for ingested content, however; the Digital Preservation unit does not take ownership or primary responsibility for the materials themselves, only for ongoing provision of the DPS in which these collections are managed. The DPS units (i.e., the collections managers identified by the collecting repositories) determine descriptive metadata for their ingested collections and will also be

\(^6\)In addition to the YUL units, YUL Digital Preservation also provides preservation services to Yale’s Museums and the Yale Law Library, by managing the separate Preservica DPS tenancies for these Yale organizations.
responsible for future decisions about managing this content.

In terms of financial responsibility for units’ ingested materials, the Digital Preservation unit identifies a long-term funding source for all materials prior to ingesting content into the system. For the majority of digital content already held by YUL repositories and identified for long-term preservation, YUL has assumed direct responsibility for long-term funding to cover ongoing storage costs. The system also encompasses content with long-term financial support allocated by other funding sources, however. For example, for a few particularly large collections held by individual units within the YUL (such as the Beinecke Library and the Fortunoff Video Archive) funding for long-term storage costs has been allocated from financial sources specific to those collections.

Resources

4. “WARC Specifications” https://iipc.github.io/warc-specifications/ - an IIPC github site that provides extensive documentation of the WARC format, indexing, and playback.
   a. Within this IIPC site, take a look at the Primers section, “Introduction to Web Archive Formats”: https://iipc.github.io/warc-specifications/primers/web-archive-formats/
7. Preservica documentation for web ingest workflows is in the Standard Workflows guide, sections 2.11 and 2.12. This document is available via YUL’s DPS instance, as well as from Preservica’s forums site, “Documentation” section.
7. Preservica Forums thread, Spring 2018 - ingesting WARC files from Archivelt:
   a. note: all Yale community members can sign up for a Preservica Forums account to view all discussion threads. Sign up for a forums account by requesting an account with your @yale.edu email address.

8. Webrecorder project documentation site on github
   https://github.com/webrecorder/webrecorder

9. Webrecorder service - includes a link to download the “Webrecorder Desktop Player App” - for viewing/playback of WARC files https://webrecorder.io/

   a. This issue was linked from “Looking for Feedback on Standardized Directory Layout for Web Archive Collections”, a Google group post to a web-archives list asking for contributions to this issue.
      https://groups.google.com/forum/#!topic/web-archives/lfN_8wztKaU