Report on Enhancing Services to Preserve New Forms of Scholarship

Jonathan Greenberg, NYU Libraries; Karen Hanson, Portico, ITHAKA; Deb Verhoff, NYU Libraries

December 2021

DOI: https://doi.org/10.33682/0dvh-dvr2

Executive Summary

Scholars are experimenting with increasingly diverse digital technologies to express their research. Publishers, in turn, are working to evolve their platforms and services to support publications that integrate dynamic features such as data visualizations, multimedia, maps, and more. In this effort to keep up with the creative demands of scholars, publications may evolve in ways that present a serious challenge to preserving or even sustaining them in the long term.

In a project funded by The Andrew W. Mellon Foundation and led by NYU Libraries, a group of digital preservation institutions, libraries, and university presses collaborated to study examples of these dynamic forms of scholarship. There were two goals: Determine whether publications could be preserved in their current form and whether it would be possible to do this at scale, and use these findings to develop guidelines that will help authors and publishers to make these publications easier to preserve, and as a side-effect more sustainable.

Preservation Guidelines

There is a companion document to this report containing recommendations for creating digital publications that are more likely to be preservable. These guidelines are meant to be shared with publishers, authors, editors, digital production staff, software developers and those who design and maintain publishing platforms. An interactive web version of the guidelines is published at https://preservingnewforms.dlib.nyu.edu, and a full static, print-ready version is available at https://doi.org/10.33682/221c-b2xj.

Contents

Changes in Scholarly Publishing 5

Project Description 9

Preservation Objectives 13

Preservation Activities 17

Assessment 35

Works Cited 38

Appendix A: Publications Analyzed 39

Appendix B: Acceptance Criteria Template 42

Appendix C: Enhancing Services to Preserve New Forms of Scholarship Project Participants 46

Changes in Scholarly Publishing

The proliferation of digital tools in scholarly publishing over the past 25 years has created a challenge for the preservation of scholarship. The preservation of print materials has traditionally been the responsibility of libraries, who, often under consortial, national, and regional agreements, steward materials that they have acquired. Digital materials, which are sometimes acquired outright and sometimes licensed to libraries, have occasioned the development of new organizations and structures to support the ongoing preservation of the scholarly record. In addition to national libraries, CLOCKSS, Portico, and HathiTrust have been among the largest providers of digital preservation services for scholarship. In order to operate at scale, these organizations have developed processes for preserving material in common digital formats: PDF, XML, EPUB, HTML, and others. This network appears successful at preserving publications in those forms that hew closely to the conventions of scholarly publishing: linear, text- and image-centric works that could be expressed well in printed formats. These are the vast majority of journal articles and books, many of which are published simultaneously in print and digitally.

The amount of scholarship published using digital technologies beyond these conventions has increased steadily over the past two decades. From hypertext and interactive web-based scholarship of the 1990s, scholars and publishers have expanded the range of forms that scholarship can take. But whereas major preservation service providers now have clear and scalable processes for text-based journal articles and books expressed in XML, PDF or EPUB, they do not currently have reliable, scalable processes to preserve publications that have been further enhanced with multimedia or interactive features, or that encourage web-based, non-linear navigation. Even web-archiving–based workflows, such as that used by LOCKSS, often fail to preserve many features when they diverge from expected patterns.

Because print books published by university presses and other scholarly publishers were so widely preserved by libraries in the twentieth century, scholars and other library patrons have an expectation that books will be preserved, i.e., that future researchers will have access to them. Digital preservation services have emerged to fill this gap. Preserving the scholarly record is now a collaborative activity between publishers, libraries, and these preservation services. Publishers and content aggregators use standards in creating digital objects, and then engage preservation services to safeguard and make them available in the event that the publication becomes unavailable in the future. Libraries preserve digital outputs that they have

stewardship over, such as scholarship in their institutional repository and digital projects and web sites sponsored by the library or its parent institution. This collaboration is currently constrained: first, by shrinking library and publisher budgets, and second because researchers and publishers continue to expand the forms and features of their publications. The expectation of preservation remains, or at least it remains unexamined. As more research outputs make use of complex web technologies, third-party software, and remote resources, the gap between expectations and reality widens.

Enhancing Services to Preserve New Forms of Scholarship, led by New York University Libraries in collaboration with Portico, CLOCKSS, Michigan Publishing, the University of Minnesota Press, Stanford University Press, UBC Press, and NYU Press aimed to narrow the gap between reality and expectations for the preservation of digital scholarship. This goal was met in two ways: first, Portico and CLOCKSS conducted exploratory preservation processes for a wide range of complex digital projects from university presses in order to determine what could be preserved and how scalable the process might be; and second, we developed guidelines that publishers, authors, preservation services, and platform developers can use to improve the preservability of complex digital publications. This report situates the project and describes our methods and findings. The guidelines have been published as a companion document to this report and are available in a sortable website at https://preservingnewforms.dlib.nyu.edu.

NEW QUESTIONS

The ongoing development of digital technologies, and the concurrent development of scholarly research methods, will continue to pose challenges for preservation. To an extent, this is an inherent feature of digital scholarship. As new tools for conveying ideas become available, and as scholars find new ways to build on and incorporate existing digital resources, the tools that libraries and preservation services have built will never fully accommodate all elements of digital scholarship. In order to develop services and workflows that will meaningfully preserve scholarship for the future, several theoretical questions require consideration.

In the print context, research outputs are generally ontologically well-defined and stable. This allows for stable services for preservation; it also makes the process of defining the work itself easier. In contrast, digital research outputs sometimes make definition of the work difficult. If an author has embedded a video from YouTube in

the body of the text, is that video necessarily part of the work, or could it sometimes be considered supplementary? What if the YouTube-hosted video was created for the project? Conversely, when might digital resources marked as supplementary in fact be necessary for some future users'[1] understanding of the author's argument? In short, how do we define the extent of the work that is to be preserved?

This might also be viewed in terms of the scholarly record: When research outputs are well-defined and uniform, scholarly communities can more easily define the scope of outputs that form the core of scholarly communication. As output formats proliferate, and boundaries between disciplines and between audiences blur, libraries and preservation services can no longer take long-standing assumptions about the scholarly record for granted. When do data sets, visualizations, audiovisual clips, and other interactive features become part of the corpus necessary for maintaining academic communities over time? And what about user-generated annotations or contributions?

Even when questions about objects of preservation can be answered, organizations that publish, collect, or preserve scholarly work often cannot provide reliable stewardship over new forms of scholarship. One premise of Enhancing Services to Preserve New Forms of Scholarship was that libraries, publishers, and preservation services must work together more than they have in the past in order to preserve complex digital publications. This premise tacitly acknowledges that in the past, each of these players was able to fulfill its goals with only a modest level of ongoing collaboration. Libraries asked publishers to use materials and production methods that would aid in preservation and conservation, but publishers themselves needed little knowledge about preservation in their usual workflow. As publishers now create publications with digital features that vary in complexity and technology, they need to work more closely with preservation services and libraries to make their products preservable. How can publishers, libraries, and preservation services adapt to work together?

Questions of scale were crucial to this project. While we were interested in improving digital preservation for a range of preservation methods and services, our primary concern was with scalable processes such as those with which Portico and CLOCKSS operate. If individual publications require days of software development in order to be harvested by a web harvester or converted to a sustainable format, preservation will

[1] Because those who interact with enhanced digital scholarship do much more than *read*, we use the term "user" rather than "reader" throughout this report.

only be feasible for well-funded projects. Services such as Portico and CLOCKSS, whose business models require scale, may not be able to process them. Still, the more complex a digital publication, the more likely it will require more labor to preserve; and the project sought to improve preservation services for those publications as well. So, we asked, what compromises must publishers or preservation services make to preserve complex publications at scale? Are there solutions that both satisfy the publisher and the needs of preservation services?

Participants knew from the start that preserving some of the more complex digital publications from university presses was a significant challenge. There are limits to what can be preserved, especially in a world where new applications, protocols, and standards are constantly being introduced. But we didn't know what those limits are. So we set out to understand what is possible under the current business and technological constraints.

Project Description

SCOPE AND MOTIVATION

Enhancing Services to Preserve New Forms of Scholarship aimed to investigate a variety of enhanced digital scholarly publications to identify which of their features can be preserved at scale using tools currently available, and which are likely to be lost over time. All of the publications were book-like, in that they were comparable in scope to scholarly books, although publishers and users may not consider some of the publications to be books.[2]

The project included two main activities. First, publishers transferred digital assets and metadata for scholarly publications representing varying content types to the participating preservation service organizations. The preservation practitioners analyzed the materials sent in order to determine whether their existing processes could be adapted to reliably preserve a publication as a whole using tools currently available. The content selected for this project ranged from formats such as EPUB with audio or video supplements, to bespoke web publications with complex interactive features, annotations, and/or dependencies on third-party platforms. The transfer and analysis of publications happened in three three-month sprints in which publications were grouped by complexity and likely methods for preservation.

Second, these findings were reviewed by an invited team of librarians, archivists, publishers, and technologists in order to gather inputs for a set of guidelines. The guidelines provide advice to publishers and scholars for creating enhanced digital publications that are more likely to be preservable, or at least ensure that the implications of adding certain features are clear so that alternative paths can be taken when possible.

[2] Enhancing Services to Preserve New Forms of Scholarship came on the heels of a wave of projects from university presses and academic libraries that built infrastructure and capacity for digital monographs (Waters, 2016). While the work of these projects was of particular interest to us, and we focused solely on long-form outputs, we hope that our research and the resulting guidelines will apply to digital scholarly publications more generally.

PARTNERS

Project participants represented scholarly publishers, preservation services organizations, and libraries that may provide publishing services, preservation services, or both. Publishers included NYU Press, Michigan Publishing, the University of Minnesota Press, UBC Press and Stanford University Press. Four out of five of the participating publishers also participated as platform developers: NYU Press for Open Square, Michigan Publishing for Fulcrum, the University of Minnesota Press for Manifold, and RavenSpace at UBC Press. Preservation service organizations included CLOCKSS, Portico, and the libraries of the University of Michigan and NYU.

At NYU, a project manager was hired to oversee day-to-day activities for the project. They organized the content transfer sprints as well as the assessment reviews with invited experts. At Portico, a senior research developer analyzed content, assessed publications for preservation readiness within Portico, and contributed to the preservation guidelines for publishers. The CLOCKSS technology development was carried out by the LOCKSS team at Stanford University. They likewise analyzed content and assessed publications for success within the CLOCKSS preservation service.

HOW THE WORK WAS ORGANIZED

Our 18-month–long project was divided into three sprints, with publications assigned to the sprints according to their technical features. The team processed the least complex publications in the first sprint and the most complex publications in the third sprint.

During the first sprint, Portico and CLOCKSS worked with EPUB-based publications from Michigan Publishing's Fulcrum platform (https://www.fulcrum.org) and NYU Press' Open Square (https://opensquare.nyupress.org). These publications have been ingested into web-based reading systems, and include a variety of multimedia and supplementary material either within the EPUB itself or as a platform-level resource

During the second sprint, Portico modeled solutions for preserving web publications with a linear, text-based structure on Manifold (https://manifoldapp.org) and digital publications from Michigan Publishing not on their Fulcrum platform. Like the publications in the first sprint, these publications allow for many web-based interactions, but are limited to a predictable set of interactions. However, Manifold publications allow for a broader range of added digital resources, both alongside and apart from the main text. Many of the publications in both the first and second sprints

support enhanced features such as annotations, embedded multimedia, and data visualizations.

The third sprint covered the most complex, media rich, and nonlinear publications for which an interactive experience is at the forefront. In most of these more dynamic works, third party dependencies are an integral component. In this sprint, Portico and CLOCKSS worked with publications from UBC Press's RavenSpace platform (https://ravenspacepublishing.org), Stanford University Press (https://www.sup.org/digital/), and Michigan Publishing.

WHAT WE DID

The workflow within each of the sprints was designed to capture data from the participants during each phase of submission and evaluation for a publication. (The template used to record this data has been reproduced in Appendix B.)

To begin, publishers submitted publications to be considered for processing. The project manager organized and assigned these to sprint project teams allowing for several publications to be in flight at the same time. During an initial evaluation phase, the assigned publishers and preservation partners collaborated to perform a detailed review of each publication. Together they defined the core intellectual components of the publication — those that must be preserved for future audiences to fully understand the work's substance and arguments. Reviews included detailed instructions for the playback or reading experience of the material submitted. They described what an intended audience should be able to do when the archived content is made available in the future. These core intellectual components served as acceptance criteria for the success of the work done in subsequent phases. In addition, description and documentation of these components gave preservation providers a more complete understanding of the context and dependencies for a work.

Pre-Transfer Activities

Pre-transfer activities included a determination of how the publication would be transferred to the preservation partner followed by a detailed description of the content made available. For submission information packages sent via file transfer, publishers provided a full description of the file types, what each file or group of files represent, and how the files together form the work to be preserved. Included in these files was any available metadata and information about how the metadata is mapped to

corresponding files. For web transfers, publishers described in detail the content made available to a web harvester. The publishers and preservation providers worked together to define content sets and starting points. And for works that were to be emulated, the publishers described the content made available so that the publication could be recreated on a virtual machine that could run in an emulation environment. In all of these scenarios, publishers noted any external dependencies on media or software that were not part of the content package marked for submission.

Preservation Actions

In the preservation action phase, the preservation partners evaluated the submitted materials and either (a) attempted to preserve the work as outlined during the initial evaluation or (b) created a detailed mockup of the proposed preservation process. They documented iterative and final decisions about preservation actions, as well as any concessions made. Any questions, roadblocks, or tasks that warranted further exploration were noted and spun off into issue tickets, to be worked on separately from the primary work they were generated from. The publisher and preservation partner collaborated to ensure that the publication could be successfully recovered from the preservation copy according to the agreed-upon acceptance criteria. Works that progressed through the preservation actions were moved forward for assessment.

Evaluation

At the end of each cycle, the publisher answered questions related to the playback experience of the preservation copy of a work. Their answers captured the degree to which the archived content available matched with the preservation goals and expectations about what would be preserved. The preservation provider responded to prompts that aimed to capture what was preservable using current tools. They recorded any constraints such as technical limitations or limits on what was feasible in the time frame provided.

Together, the project team recorded lessons learned from each work, which form the basis of guidelines for better preservability. We made note of modifications that a publisher could have made during the creation of the original work to improve the preservability of the material while maintaining the essential aspects of the content.

Preservation Objectives

MANAGING EXPECTATIONS

Following established practices led by the preservation service providers, we began our work by asking publishers to articulate their goals for each publication. They identified the content and functional elements that we would attempt to preserve, and these served as the criteria upon which we planned and evaluated subsequent preservation activities. The instructions and context necessary for the playback or reading experience of the material submitted also helped capture their expectations for what the intended audience should be able to experience when the archived content is made available. This process allowed the preservation service provider to focus their work on what was deemed essential and to determine the best method for transferring content from the publisher.

DEFINING A WORK AND THE ELEMENTS TO BE PRESERVED

The publications considered during this project ranged in complexity from enhanced, web-based renditions of EPUBs to complex web publications with interactive, dynamic features. Some of the web based publications make use of platforms that have a consistent, finite, and fairly predictable set of interactions. The more complex web publications embed dynamic features which drive unpredictable interactions with a server such as full text searching, complex data filtering, map navigation, adding highlights or annotations.

During the project, a total of 18 publications were analyzed. Six of the publications are described below to provide examples of the kinds of publications that were included in the project and the criteria for preservation.

By Any Media Necessary: The New Youth Activism is an EPUB-based work published on NYU Press's Open Square platform. It is a version of the paper edition of the same book that has been enhanced with embedded video. The main criterion for successful preservation from the publisher was to preserve the text, images and video elements so that a user can read the book in sequential order. Ideally, users should be able to view the media at the appropriate location within the text; however, a less seamless user experience would be acceptable. The embedded videos are hosted on YouTube, and not in the control of NYU Press or the author. The publisher was able to transfer a submission information package with an EPUB file and accompanying metadata, but

not files of the referenced videos. In some instances, the external video content had already been removed from YouTube by the time our project began.

99 Theses on the Revaluation of Value: A Postcapitalist Manifesto is a University of Minnesota Press title published on Manifold. It consists of five project texts which include annotations and highlights from users who opt to share them publicly. The publisher's expectation was that these features would be captured and embedded into the preservation copy of the work at the same anchor points at which they exist in the original Manifold edition. Future users are also meant to understand the overall structure of the work. Another University of Minnesota Press project, *Cut/Copy/Paste: Fragments of History*, includes a draft chapter from the forthcoming book of the same title, as well as 124 individual project "resources." Among these are spreadsheets, images, and links that resolve to a Twitter query for the project's hashtag.

A Mid-Republican House From Gabii is a University of Michigan Press title which displays EPUB-based text alongside a WebGL visualization that allows users to explore an interactive 3D rendering of the archaeological site described in the text. In this work, a user can click on links in the text that will display the corresponding location in the 3D model, and similarly, clicking on the model brings a user to the corresponding pages in the text. This relationship between the text and the 3D visualization was noted by the publisher as a central element for preservation. Additional objectives were to preserve supplemental images and tables, the 3D model itself and an external database referenced in the work.

As I Remember It: Teachings (ʔəms taʔaw) from the Life of a Sliammon Elder is a media rich publication on RavenSpace, a Scalar-based platform developed by the University of British Columbia Press that "embraces collaboration, respects Indigenous protocols, and uses digital tools in imaginative ways to make knowledge accessible and shareable across communities and generations." The work, which includes interactive maps, audio, and video in a non-linear presentation, represents joint scholarship between Tla'amin elder Elsie Paul, historian Paige Raibmon, and members of Paul's family. The publishers identified the audio and visual content to be of critical value to this work. In addition to preserving media in relation to the text, other essential features included the "Protocol for being a respectful guest" popup message presented to all visitors to the site, interactive pop-up notes for key terms, navigation pathways nested in multi-path structure, and a custom keyboard for Indigenous languages.

Chinese Deathscape: Grave Reform in Modern China is a publication from Stanford

University Press which relies heavily on an interactive map that displays alongside the text. Linkages to sections of text define changes in the map view and focus point. Users can navigate through the map to discover data at different moments in time or in specific date ranges. The map also provides access to source data via references to particular newspapers and articles reporting the information mapped in the publication. These interactive elements were deemed essential to the work. As the publisher stated in the Acceptance Criteria, "To lose the map is to lose what matters about this publication."

EXISTING PRESERVATION-ORIENTED FEATURES OF PUBLICATION PLATFORMS

Most of the publications reviewed during this project are published on open source platforms designed for publishing new forms of scholarship. These platforms allow authors to integrate audiovisual media, digital supplements, and interactive features such as data visualizations with textual content. One of our research objectives was to consider the role platforms might have in scaling the preservation of complex digital publications.

Fulcrum was built on the same Samvera stack as the University of Michigan Library's Deep Blue institutional repository and is governed by the same preservation policies. The platform was built with an express intention that the materials would be durable and accessible over time. During the course of this project, we considered the durability of relationships and interactivity between textual content and related digital resources.

Manifold was designed as a digital publishing platform for iterative as well as conventional publications. Its export feature converts textual content as well as the project materials to a ZIP archive that conforms to the BagIt specification. Other features such as the platform's user-contributed content, supplements embedded from third party platforms, and some of the relationships between the resources and content are not fully expressed in the standard export function. Similar to Fulcrum works, we focused on preserving the relationships between the text and media enhancements and also explored preserving some of the enhanced features that are integrated into the platform.

Scalar, a project of the Alliance for Networking Visual Culture, is an authoring and publishing platform designed to allow for flexibility in structuring media rich works.

Publications can be exported as RDF-JSON or RDF-XML as a backup of text components, metadata, and relationships among text and non-text components. The export feature is designed to facilitate the migration of projects from one Scalar install to another. During this project, we considered the export function as a tool for preservation. As with other platforms, we focused on preserving the relationships between elements in a work as well as some of the interactive and experiential aspects of these dynamic publications.

Preservation Activities

METHODS

For each work, preservation technologists designed scalable preservation pathways that might function for similar works. In each instance, the publisher described the features of the publication that they would like to preserve ("required" through "nice to have"), shared the location of the live publication, and transferred any available export packages or metadata to the preservation institution. The publication was manually reviewed in the live environment, and then compared to the package where provided. After the initial review, the preservation institution experimented with workflows for converting the publication to an archival package containing all of the information required to replay it later if triggered on their platform.

Due to the number and complexity of publications as well as the variation within platforms, it became clear early in the project that attempting to fully process each publication using Portico's and CLOCKSS' existing preservation systems would have made the pace of the project too slow and limited in scope. The preservation institutions instead focused on combining standard workflows with features and functionality not yet in place. For example, Portico tested EPUB rendition, emulation, and web harvesting, each of which would have required significant development time to implement. They configured new required tools, wrote scripts, and developed proof-of-concept implementations in order to demonstrate what the output of the process could look like if the publications were consistent. Portico also created mock-ups of their access website to tie together the workflow outputs and show how it could be presented if triggered.

Described below are three general approaches that were taken in order to preserve the publications: file transfer of information packages; web archiving; and emulation, as we employed them for specific publications and platforms. As the project progressed to more complex examples in which the experience of the platform was an integral feature of the work, methods for preservation became more complex and experimental. In selecting which of these three methods to use, the general approach was to preserve what we had some confidence could be preserved (e.g. text, raw supplements, and metadata), but also reach for the solution that could fulfill all of the publisher requirements. Thus, we used the file transfer method for all publications, often as a secondary mode that would plainly fail to meet the acceptance criteria. This is standard procedure for Portico, ensuring the retention of at least some content using

the least precarious methods. For this reason, many of the publications included both an export package and an emulated or web harvestested version in the archival package.

FILE TRANSFER OF INFORMATION PACKAGES

The file transfer method was deployed by Portico for all works. In some cases, we saw the possibility that this method, the standard one for most of Portico's operations, could suffice for providing access to the work in the future; other works were too complex in their structure or user experience to be reconstructed with discrete files and metadata. Publications that could most easily be exported and reassembled from their component parts were best able to make use of this method. For each of these publications, publishers created a submission information package (SIP) containing the original EPUB file, supplemental media files, and metadata which was transferred to Portico via FTP. Files provided were assigned a file type (based on the output of several file format tools). To the extent possible, each file was then validated and a detailed report for that format was produced and incorporated into the object's technical metadata. While working on this project, Portico added a new EPUB module to the JHOVE application using the World Wide Web Consortium's EPUBCheck tool. JHOVE is an open source tool used to validate and characterize file types and it can be used to identify structural issues that might affect future playback of the file. The new EPUB module is available to the community via the official Open Preservation Foundation's version of JHOVE.

Key to the success for this file transfer method was the inclusion of all resources contained within the publication along with metadata that provided enough information to map embedded resources to their appropriate positions in the EPUB. The availability and structure of metadata varied across publishers. In the case of Fulcrum-hosted publications, transferred SIPs contained an EPUB file, media files, and a Fulcrum manifest in the form of a CSV file exported directly from the publishing repository. Media files and associated software for playback, viewing, or interaction were embedded or linked from within the EPUB via a URL rather than contained within the EPUB, making rendition dependent on the persistence of those URLs.

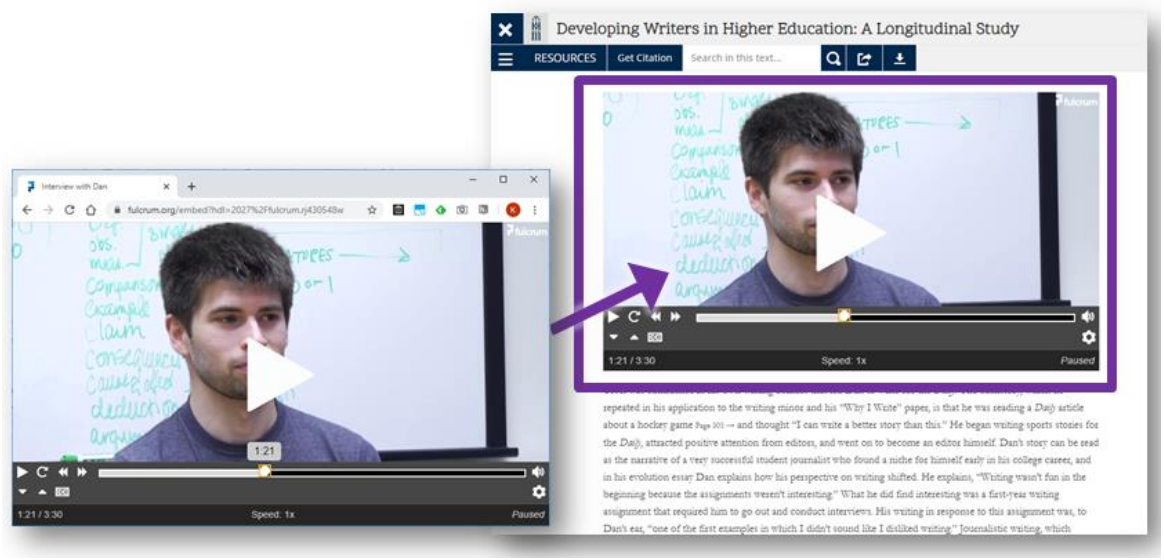

Figure 1: Custom video viewer in a Fulcrum EPUB. Embedded using an <iframe> displaying the viewer on the Fulcrum website. From *Developing Writers in Higher Education.*

At the start of the Portico workflow to create an archival package, the CSV file containing metadata about the media supplements was converted to an XML format and tagged as descriptive metadata. The EPUB included an Open Package Format file that contained additional information, and this was also tagged as descriptive metadata and validated for errors. The two XML files were merged and transformed into header metadata in Book Interchange Tag Suite (BITS) XML, a standard for academic book content and metadata used by Portico to describe books. Where TEI XML was included in the package, full text BITS was generated.

Though it would have been possible to preserve the separate resources and leave it to a future user to reconstruct the work, this approach would not have fulfilled the publisher's aim of preserving the original reading experience with media embedded within its textual context. Nor would it have functioned within Portico's business model, in which publications are "triggered" in the event that they become unavailable or inaccessible through their publisher. For this reason, Portico experimented with ways to make access scalable. The Fulcrum EPUBs use iframes to display enhanced media players that are live on the Fulcrum site. That means if the website is unavailable (as it would be in the event of a Portico trigger), the EPUB will display empty boxes where the media players once were with no visible indication of how to reach those resources. One workflow that was tested required two changes. First, access copies of the media files were moved to a folder within the EPUB. Second, XML transforms were used to replace the Fulcrum media player iframes with simplified EPUB3-compatible media players that would allow the media file and player to be

contained within the EPUB package. The result was a larger new EPUB that could be used for dissemination through the Portico Portal in the event that the publication was triggered. The original unmodified EPUB was also preserved and presented alongside the access version through the portal. While this process of modifying the EPUB resulted in an access experience that was closer to the original, it did result in some loss. The feature-rich Fulcrum media players were not preserved. Instead, transform efforts focused on the vital intellectual components of the experience, for example, the ability to play an audio clip inside the text and simultaneously read the transcript.

With later examples, Portico suggested that descriptive captions be added under each of the embedded multimedia elements in the new access version EPUBs and that they include their unique persistent identifiers. The caption could provide information about the embedded material if it should fail to play back correctly, and the identifiers could resolve to their new Portico locations rather than the original Fulcrum website if triggered. For Portico to add a caption during the transform would require minor editorial decisions about the look and feel of a publication with its transformed content. Portico prefers not to make such decisions without input from the publisher. The transforms and media embedding process for each individual Fulcrum publication were deemed successful and the publishers' intent was realized. It would currently be difficult to execute this process across all Fulcrum publications due to minor variations in how media embedding was handled between the EPUBs. This meant there was a need to check for potential issues resulting from transforms and to repeatedly tweak the process to accommodate for those variations.

Adapting to challenges related to embedded third party resources

Preservation specialists aimed to retrieve resources that were visually embedded in the text but located outside of the EPUB container and maintain their relationship to textual content within the EPUB. This was easier to accomplish when these resources were published via the same platform as the publication, as with many examples on Fulcrum. Where these embedded resources were hosted by a third party service they required special attention.

The NYU Press publication *By Any Media Necessary* contains embedded YouTube videos which are integral to the work. Portico generally requests that the publisher submit a copy of the video files along with text-based files in EPUB or XML format. The publisher of an EPUB-based book like *By Any Media Necessary* could embed the video files directly in the EPUB. However, for various reasons, publishers with platforms for enhanced digital books have often chosen not to embed video in ebook files intended

for distribution: EPUB files may become large and lose the portability for which the EPUB standard was designed; platform developers have also leveraged these platforms to allow for playback features that may not be available even in widely used reading systems such as those built by Amazon, Kobo, and Apple. In this case, the video content was hosted on YouTube at the time of publication. NYU Press does not hold copies of the video files, and could not provide them to Portico. The risk of referencing externally hosted content was evident from the start of the project: one of the videos has already been removed from YouTube, resulting in a grey box within the publication.

Without explicit permission from rights holders, Portico will not capture and store content from YouTube or similar commercial services. Portico looked for options to improve the likelihood that the video content would be preserved and that did not involve storing a local copy of the content. One method that showed promise utilized the Internet Archive's Save Page Now service which allows human- or machine-initiated capture for a single webpage and immediately produces a persistent URL to the archived version of the page. Portico experimented with a proof of concept workflow in which embedded YouTube videos were identified within an EPUB; the corresponding URLs were submitted to the Internet Archive; and a notice was added beneath each video in the publication linking to the archived location. This method seems possible to automate with code and provides a level of content stability. However, because the process is not always successful, adoption of this method may require manual quality checking. Ideally, the archiving process would take place during production—or even earlier, during the research or writing process—to mitigate the risk that videos may be deleted from YouTube, as was the case with *By Any Media Necessary* .

Adapting to challenges at the platform level

A scalable workflow for the Fulcrum publications may be possible with more consistency around how media is embedded since every variation introduces further complexity and fragility. For example, some Fulcrum publications used a IIIF viewer to display images, others embedded the image file directly; some used captions, others didn't. Since Portico prefers not to make editorial decisions on layout, nor risk a brittle preservation workflow, a better solution would be for an alternative EPUB export to be built into the Fulcrum platform itself. This could generate a larger EPUB for offline use that would be more appropriate for preservation and would evolve in line with the platform development.

Export packages were also generated by the Manifold and Scalar platforms. The Manifold export was in development during the project. As with Fulcrum, its export package contains an EPUB and a comprehensive set of supplements and metadata packaged together using the Bag-It specification. After an initial evaluation of the export feature, preservation technologists expressed concern that the metadata was spread across too many files and folders within the package: there was a top-level project metadata file and two files for each supplement serialized as JSON; XML metadata embedded within each EPUB; and then a variety of small text files within each folder in which the file name was the property name and the contents the value. While comprehensive, this would have required a very complex workflow to process the different metadata formats and files. Portico recommended that these smaller files be consolidated into fewer files that all use the same serialization format. While this package could have been used for more conventional publications, for the complex publications selected for this project, this export would be insufficient to preserve significant aspects of the user experience. In particular, the seamless integration of text, multimedia, and user contributed content such as annotations and highlights were not captured. The Manifold export packages were deemed important to preserve as a baseline, but there was interest in exploring whether we could also preserve the rich experience offered on the platform.

Scalar's export was an RDF package that was more suited for migration between Scalar instances than preservation. The networked nature of the content and the specificity of the stylesheet classes and property names to the platform meant that without the platform, the publication would be extremely challenging to render. An approach that could capture the work as a website seemed like a more scalable option.

WEB ARCHIVING

Both of the preservation service providers employed web archiving for a number of publications during the project. The preservation activities for CLOCKSS were performed by the LOCKSS development team at Stanford. The web archiving approach deployed for this project uses Heritrix, the Internet Archive's open-source, extensible, web crawler project. To archive publications on a particular platform, LOCKSS engineers develop a plugin to the LOCKSS software with descriptor rules and code describing the harvesting process specific to that platform.

At the time of the project, Portico did not have a web archiving service in production, but used the publications presented in this project to test several of the new generation

of browser-supported web crawlers-a different approach from CLOCKSS. Rather than attempting to predict and simulate what a browser does, these tools use an existing browser such as Chrome to automatically crawl the website and assemble any loaded resources into a WARC file. Many modern websites use JavaScript that continues to load new resources after the initial page load and as the user interacts with the page. These new crawl tools each have options for simulating user behaviors such as scroll, click, hover, etc. The Portico preservation technologist tested Brozzler, Squidwarc, Memento Tracer, and Browsertrix. Each had its benefits and drawbacks in the context of this project. While Squidwarc offered the best balance of control and simplicity, Internet Archive's Brozzler was the most mature crawling tool and was used for most of the tests. Since completing the project, Portico has moved forward with a web archiving pilot using Webrecorder's Browsertrix Crawler - a lightweight derivative of Browsertrix that uses a Chrome browser and Pywb for capturing web pages, and Puppeteer JavaScript for simulating user behaviors.

The preservation services tested a web archiving approach with 15 publications. In each case the service evaluated what was possible to archive and how well the approach could scale. The definition of acceptable scale and precision may vary between preservation services. Services like CLOCKSS and Portico work with publishers to build solutions that are tailored to their platforms. A custom workflow is considered to be successful if it can, over long periods, automatically and accurately archive all publications on the platform without needing to be modified frequently as a result of changes or inconsistencies. Ideally the benefits of the work to tailor a solution for a publisher's platform can be transferred to other publishers that use the same platform or standards. The guidelines accompanying this report describe ways to design web-based publications that can be successfully captured using web archiving, and also aim to minimize the effort required for customization and maintenance of preservation workflows. If a website is easy to archive, a wider variety of services can preserve the site at a lower cost, improving the likelihood that the publications will be available to future scholars.

With time to customize the web archiving tools, the preservation services had some success in creating reusable workflows that could capture the majority of the publications that were presented using publishing platforms like Manifold, Scalar, or Fulcrum. These customized workflows were relatively complex, however, with some taking weeks to develop as a proof-of-concept. Where there were certain types of enhanced features embedded within the publication, or the publications were custom-designed rather than presented using publishing software, the results were mixed. The

publishers understood that some loss may be unavoidable using these methods, and this was acceptable for features considered "optional" to preserve (e.g. full text search). In some cases though, the very features that qualified the project for inclusion in this research -- features such as interactive map visualizations and IIIF viewers -- could not be captured using web archiving, or required manual work that was out of scope for the preservation services. Specific successes and challenges are described in the sections that follow.

Adapting to challenges at the platform level

Prior to the project start, LOCKSS had been working on a plugin for Fulcrum publications. The development of the Fulcrum plugin continued during this project; feedback from the project team during weekly meetings fed into an iterative development. The method involved harvesting native publication content from its published location on the Fulcrum website. EPUB files were also harvested from the Fulcrum site when available. However, at the time of the project, LOCKSS was not able to ingest both EPUB and other resources separately in a way that would maintain linkages. Therefore, the publications were treated as web sites for the purpose of harvesting.

The LOCKSS team reviewed individual Fulcrum works that had noteworthy or extensive resources embedded in, or presented alongside the text, such as *Animal Acts: Performing Species Today*. The Fulcrum plugin for LOCKSS could harvest the main text of the publication. Use of third-party, remote fonts, such as those from Typekit or Google Fonts, can be an issue in web archiving. For Fulcrum, the publications' font and icon toolkit is hosted on the platform, making it possible to capture. Links to PDFs, video, audio, and transcripts were easily discovered and downloaded by the web crawler. The functionality on the Fulcrum publication landing page that allows users to sort and group the publication's resources was not archived. This feature adds an open-ended combination of parameters to the URL, resulting in an explosion of "pages" that the web crawler identifies as unique content. While crawling a large number of pages is possible it requires an inordinate amount of time and an impractical amount of storage. The plugin approach was therefore successful for capturing many of the core features of the Fulcrum publications and could potentially be applied to other Fulcrum installations. For some Fulcrum publications, specific dynamic features proved challenging with this approach. These will be discussed later.

The Manifold team at the University of Minnesota Press identified six Manifold projects as candidates for analysis, and the project team selected four as

representatives. The Manifold platform includes an export feature that creates detailed packages composed of EPUB versions of any texts in the project, non-textual resources, and metadata for the project and all provided files. The packages contain the core source material, but lose a lot of the richness of the original presentation such as integration of user-contributed content and convenient navigation between the text and resources. Portico explored both whether the packages could be improved to reflect more detail and whether a web archiving approach could preserve the project as originally presented.

Portico initially attempted to archive a Manifold project by providing several web harvesting tools with the starting URL and allowing the crawler to automatically discover the pages of the publication. This proved ineffective due to the design of the platform site — there is no sitemap; projects use multiple URL arrangements making it difficult to control the scope of the crawl; many HTML link tags, which are typically used by crawlers to discover content, do not reference a target URL but instead initiate a JavaScript action on click to load a new page; and, new data is loaded from the server as the user interacts with some page features. All of these factors contributed to incomplete crawls.

In a second Manifold crawl attempt, Portico worked with Memento Tracer and Squidwarc. Initially the crawl scope was controlled by encoding sequences of user behaviors to automatically crawl a single Manifold project. The size and complexity of these publications made this a difficult task with complex behaviors required on many pages to capture all content (e.g. for each annotation click to open, if the annotation panel has a more button click it, close panel, repeat for next annotation). A simple proof of concept was developed that limited the scope to see if these tools were viable. The crawl was programmed to simulate a user navigating the publication text. It started on the hero page, clicked to the first page of a text, clicked “next” until it reached the end of the publication, and visited each linked resource it discovered. A WARC file was generated that held all of the content loaded into the browser during this process. The replay was much better than for the initial fully-automated crawl attempt but was still very buggy - within a few clicks an error would appear. It also would have been a lot of effort to encode *all* of the user behaviors that would fulfill the publisher’s requirements.

The cause of the replay problems were traced to the fact that Manifold runs as a “single page application” in which the entire template website is loaded into the browser when the user visits the first page, and then additional content is loaded using JavaScript

functions that call the Manifold API to retrieve page data. When the user clicks a link to go to a new page within Manifold, the URL in the address bar is artificially changed by JavaScript, which means the page URL that shows in the address bar has not really been visited as a unique location on the network. The result is that the URLs that appear in the address bar, and the resources that would load if they were visited, may not get added to the WARC file. This causes sporadic errors in the replay of the web archive. To add to the complexity, the developer of Manifold expressed concerns about the web crawling mechanism being stable since it depended on CSS Selector or JavaScript DOM paths to simulate user behaviors. Given that the platform is in active development, these paths could change frequently.

Through these experiments, Portico concluded that a full list of resource URLs needs to be identified for crawling. This list should cover both the top level URLs that load in the address bar, and all of the API URLs that are called while the user is interacting with the site. Portico wrote a script to generate this using the Manifold API. The script took approximately two weeks to write. At this point, Portico re-evaluated the crawler tool options in the context of the new method. Memento Tracer did not support feeding in a list of URLs. Squidwarc did allow this, but during testing it was discovered that it lacked functionality to download PDFs or MP3 files, which were vital to the work. Since full control of browser behaviors was no longer needed for this approach, Portico switched to using Internet Archive's Brozzler - another browser-supported crawler. Brozzler does allow basic customization of behaviors, which were useful in other experiments, but they were no longer needed for Manifold. Brozzler was also deemed a more stable option given that it has been running in production as part of Archive-It for some time.

The URLs generated by the custom script were fed into Brozzler, which was configured to only visit the URLs provided and not attempt to discover new links. The resulting website capture was tested by Portico, and again by the publisher and project manager. The same script was used to generate the URLs for three other Manifold publications; for the second, minor improvements to the script were required to incorporate features not expressed in the first iteration. By the third and fourth publications, no further changes to the script were required. Though further testing is required, this indicated that a reusable script to generate resource URLs for crawling might be a scalable solution for Manifold publications.

Adapting to challenges related to social media and user contributed content

Several of the web-based publications in the project had various forms of user

contributed content. While this presented new technical challenges, it also raised a variety of legal and ethical questions around copyright, privacy, and safety.

A standard feature of Manifold projects is that Twitter content related to the publication is presented alongside the text. The project hero page links to the author's Twitter profile and integrates a text-only version of any Tweets that mention the project. The publisher was interested in including this context as part of the web archive. The technical challenges of preserving the full Twitter experience are well documented - many social media platforms are extremely dynamic, with JavaScript driving continuous updates to the page content in response to user interactions, and unexpected platform updates prompting web archivists to adapt their methods. Many archives opt to instead preserve Twitter in a more raw form using the Twitter API. This is much more stable compared to the GUI, though some loss is expected.

In addition to the technical challenges, there was a broader set of questions for Portico related to rights and ethics around harvesting user-generated content and archiving it without permission. Portico decided it would not be appropriate to archive an author's Twitter profile page, even if it was technically possible. The Tweets embedded in the Manifold hero page seemed like more of a grey area — they are text only, which limits issues of privacy and copyright surrounding visual material; they display only a Twitter handle (no full name or profile picture); and they are all related to the project. They were successfully captured as part of the web harvest, but further consideration is required to determine whether this is appropriate on legal and ethical grounds. If not appropriate, it raises a new and difficult technical challenge - to exclude a section of content from the web harvest even though it is seamlessly integrated into the page with data imported from Twitter on the server end. Standard web archiving practice would be to simply exclude the Twitter URL, but this is not possible because of Manifold's architecture.

While this case did not come up in any of the examples, where it is possible to obtain rights to preserve social media content, in general a screenshot of the post with a link to the original location, or even a link to an archived copy would be more stable than embedding the Tweet. As with YouTube videos, services such as Save Page Now from the Internet Archive, may be used to generate an archive link for social media content. The success of Save Page Now in recording social media posts will depend on the evolution of social media platforms and web archivists' ability to keep pace.

With Manifold and several other publications, there was also the broader question of

user contributed content in the form of comments and annotations. Several platforms integrate the web annotation tool Hypothesis. Because Hypothesis' terms of use indicate that annotations are in the public domain, they are legal to preserve. Manifold, however, has a local annotation tool that does not specify rights associated with user-generated annotations, which raises the same questions as for Twitter: is it legal and/or ethical to harvest these for preservation? An adjustment to the Terms of Use for Manifold could forestall legal concerns in this instance.

Lines become blurred once again when it comes to using third party comment plugins. *Rhizcomics*, published on a custom-built site by Michigan Publishing, uses a service called Disqus so that user comments can be added to each page. Some of these comments include uploaded images, and users have full names and profile photos on display within the webpage. Archiving these profiles freezes them in time, with the user losing the ability to control their profile or comment in the preserved copy, which again raises ethical questions about people's right to control their public online identity. While the publisher may wish to preserve this commentary, the most likely approach for the preservation institution is to exclude the Disqus URL from the harvest. A local comment service that is text-only and is covered by an appropriate terms of use statement would allow for archiving by more cautious parties.

Adapting to challenges related to dynamic content

Publication features which require communication with a server where that communication is unpredictable or results in an open-ended number of related URLs cannot be captured well with web archiving. Examples of this include full-text search, embedded Google Maps, or IIIF viewers, all of which depend on user interactions to load additional data. Capturing this dynamic content can be further complicated if the feature relies on third-party services since rights issues may then come into play and continuity becomes dependent on that service being sustained. The preservation technologists analyzed and attempted to address some of the challenges presented by such dynamic content.

Oplontis Villa A ("of Poppaea") at Torre Annunziata, Italy, Volume 1. The Ancient Setting and Modern Rediscovery, published on Fulcrum, proved to be challenging for a web archiving approach. The images in the work are displayed through a IIIF Leaflet widget that allows a user to pan from left to right and zoom. Links to the images are not contained in the HTML page. They are fetched by JavaScript and appended to the page after it is downloaded to the browser. What reads to a viewer as a single image is made up of nine tiles within Leaflet. As the user interacts with the IIIF viewer using zoom

and pan, the JavaScript dynamically fetches new image tiles from the server through an API. In some cases, as was the case with *Oplontis,* the CLOCKSS team was able to determine the image URLs by reverse engineering the process to find that images were hosted by IIIF image servers on the Fulcrum website. In the case of *Lake Erie Fisherman*, which displays images within Fulcrum with multiple zoom levels, it was not possible to discern which resolution the Leaflet tool displays at first. Without this information, the resulting image replay in the web archive appears as a gray box. In order to replicate the original presentation experience, the harvester would need to fetch all possible combinations of image tiles - another limit to preservation. One way to further the likelihood that an ebook could be harvested as a web site is for publishers to display a simplified version of their site when it is accessed by a web crawler such as LOCKSS.

As with Fulcrum, resources embedded in Manifold publications such as images, PDFs and videos—which generate predictable and limited responses from servers—were possible to capture as part of the web archive. However, data visualizations presented a particular challenge. The publication *Cut/Copy/Paste* contains an interactive map created by the author which displays different combinations of image tiles as the user navigates over the map. This dynamic content changes based on user interactions and presents a similar challenge for web archiving as the IIIF viewers encountered by CLOCKSS. The Portico team opted to treat the underlying data and code for the visualizations separately from the web archive file. The data and code were preserved along with the WARC files. This combined approach to archiving was applied to another Manifold work, *Metagaming: Playing, Competing, Spectating, Cheating, Trading, Making, and Breaking Videogames,* which contains downloadable versions of the games referenced in the text: disk image files for Mac and executables for Windows. These were preserved as digital objects in addition to the WARC files. The addition of contextual metadata related to these files would help users to identify what would be needed to run them.

A Mid Republican House in Gabii from University of Michigan Press on Fulcrum presented a different challenge related to preserving a data visualization. A navigational device in the work is driven by WebGL, a JavaScript API used to create interactive 3D graphics within a browser. The publisher prioritized this 3D model and its relationship to specific locations within the text as an important element for preservation. In addition to this, the 3D visualization included DOI links to data records outside of Fulcrum. These records were part of a larger website that supported the ability to browse or search the full dataset. Portico first ran the work through their

Fulcrum EPUB file transfer workflow, which successfully captured all of the component parts including the visualization and metadata. However, the desired interactions between these components and the links to content from the external database were missing from this initial package export approach. A second experiment was made with web archiving, using the contents of the FTP package exported from Fulcrum to identify a list of URLs to crawl. It was possible to visit each link using Brozzler which was configured to crawl the links in the database. The resulting preservation copy was successful on playback, including the relationships between the EPUB and WebGL 3D model. The DOI links to the supplemental dataset, accessed via the WebGL visualization, were not expressed in the metadata and were therefore lost. As an experiment, Portico envisioned a scenario where Fulcrum had included these database DOIs in the supplied metadata file as resources so that they would automatically be added to the web harvester crawl list. In a proof-of-concept test, this approach was successful for retaining the connection between the WebGL visualization and the web-based data, though it did not result in a comprehensive capture of the database. Portico suggested that in addition to including these DOIs in the metadata, adding the raw supplemental dataset to the package would ensure future researchers had a way to access the full data in some form, even if they could not fully navigate it using the archived version. For Portico, this case was an outlier among the Fulcrum works because it required a secondary approach to preservation (web archiving) in addition to the EPUB export. This added a layer of complexity to an already customized workflow. To support outliers like this, either extra effort is required to make the publication work well with a web archiving approach, or some manual effort is required on the part of the publisher and/or preservation service to support unique cases.

The Stanford University Press publication, *Constructing the Sacred: Visibility and Ritual Landscape at the Egyptian Necropolis of Saqqara*, a Scalar publication that contains a complex 3D visualization which is central to the work and therefore, a priority for preservation. The feature allows users to zoom, pan and rotate select objects in the embedded 3D visualization which is hosted on ArcGIS, a third party service. The ArcGIS instance contains the coordinates for the placement of the objects, and bit.ly is used to store detailed bookmarks that rotate and zoom the visualization within ArcGIS. The metadata that is needed to understand the relationship of the publication's components is documented within the Scalar based publication. Portico attempted to archive the visualization automatically using Brozzler, and then manually using Webrecorder. Neither resulted in a comprehensive capture of the visualization. The initial view of the visualization was captured but would break as soon as the user

interacted with it. The instance of ArcGIS—a third-party service—is essential to this publication and cannot be preserved in its current form. The data would need to be able to work independent of these third party services in order to be preservable. One possible solution for improving the preservability would be to make a package for the GIS data that is independent of the Scalar platform which would include the metadata and information about the relationships between the parts. An alternative solution, if the volume of data is small enough, might be to use a visualization tool that does not require ongoing communication with the server - as seen with the WebGL visualization in *A Mid Republican House in Gabii* in which the data supporting the tool is loaded into the browser when a user first opens the page. This would allow for better web archiving and remove a fragile dependency on a third party platform. A final option would be to create a fallback equivalent, perhaps a video demonstrating the visualization, that could be displayed if the ArcGIS service is not available. Each option requires collaborative effort from the publisher and author to ensure the item can be preserved.

EMULATION

The priorities for preservation for the RavenSpace work *As I Remember It* were threefold: preserving customizations for representing Indigenous metadata, languages and orthographies; maintaining the multi-path structure of the work; and an opening pop-up agreement requiring users to agree to respect the expressed cultural protocols before accessing the publication. Web archiving tools allowed Portico to preserve: the general page layout including citations, notes, and embedded media; the table of contents to a second level; and the popup agreement, which required extra configuration for playback. Missing from the web-archived version were: custom search functionality (including the ability to search using the First Nations keyboard), content visualizations, and a dynamic curriculum explorer. Also missing were external resources: Google Maps, iframe content containing two other websites, and a YouTube video. While the web archive was deemed acceptable by the publisher, there were some concerning issues with capture and playback that might cause problems for access in the future. The web archive seemed worth preserving despite these issues, but Portico decided to explore a more advanced approach in response to shortcomings of web archiving for *As I Remember It*.

Portico preserved the work a second time using emulation, which attempts to replicate the server side configuration of a website or application. RavenSpace is built on Scalar, a Linux/Apache Server/MySQL/ PHP (LAMP) application. Portico performed the

emulation work by creating a virtual machine on an instance of EaaSI, the Emulation-as-a-Service Infrastructure developed by Yale University Library. The virtual machine contained a LAMP stack with the GUI enabled, a web browser, Scalar 2.5.12 with the Import/Export plugin installed, and the RavenSpace customizations. The Scalar Import/Export plugin supports the transfer of the text data between platforms, but does not include the non-text resources or update media links. The vital components to recreate *As I Remember It* on the new server were: (a) the Scalar export data from the original platform, which was supplied as a large JSON file,(b) copies of the media files that were embedded into the publication but hosted on Omeka and DSpace (c) the Scalar-generated publication folder from the original web server containing all other resources used in the publication. Fortunately, these remote media files were documented in Scalar as resources, and so their descriptive metadata and original file URLs were included in the data export. This supported the process of matching files to their descriptions and placing them in the context of the text. The JSON was imported into RavenSpace on the virtual machine; the publication folder was copied to the appropriate location on the web server; and all media files were copied to the publication folder under a new sub-folder so that they were local to the publication. Media paths were updated in the Scalar database to point to their new local path. The resulting preservation copy was a fully functioning website that could run offline. It included elements that were missing from the web archived copy: the search functionality, all levels of path navigation, and content visualizations. Google Maps and a YouTube video were not migrated due to copyright and technical challenges, but these had been deemed non-essential by the publisher.

Because of its predictable content layout, the Scalar platform lends itself to both web archiving and emulation. CLOCKSS and Portico were able to create a web archiving workflow by extracting a sitemap-like list of URLs from the Scalar API. Though the workflows were complicated to configure, they could potentially be applied to other Scalar publications. In addition, Portico identified patterns in the installation of Scalar that could make it possible to recreate the platform with the publication loaded into it so that the website could be emulated in the future. In order for either of these methods to be scalable, publishers would need to handle attachments in a consistent way, and communicate any customizations to a preservation service provider. Scalar's import/export tool is useful for the emulation process, but the need to handle the publication folder and media files from DSpace and Omeka separately is burdensome. The process was labor-intensive, requiring manual work and a series of conversations to identify what pieces were required. That said, it may be possible to create a script that would automatically install a publication onto an empty RavenSpace instance for

emulation. This would require a very consistent and structured handoff between the publisher and the preservation institution. We are unsure how scalable such a process would be.

Filming Revolution, which its publisher Stanford University Press describes as a "meta-documentary about independent and documentary filmmaking in Egypt," relies on a custom LAMP (Linux, Apache, MySQL, PHP) application to deliver a user experience based on a visual network structure. The website is a single page application, which means it is heavily dependent on JavaScript to load data into the browser as the user interacts with it. Users navigate content by clicking within a visualized web of relationships between essays, videos, and themes. Both the interactive navigation and the Vimeo hosted video content were identified as primary aspects of the work to be preserved. In a package provided via FTP, the publisher provided the database and code from the server as well as copies of the video files from Vimeo. Portico experimented with both client-side (website harvesting) and server-side archiving (emulation) for this data-driven website.

For the client-side web harvesting approach, the existing sitemap URLs were crawled with Brozzler. Most of the Vimeo videos were not successfully captured, and the playback experience for navigation produced many technical errors. Stanford did have some prior success with a web harvesting approach while working with an expert from Webrecorder, who used different capture and playback tools. Understanding Webrecorder's approach and whether it is possible to automate and scale may inform further exploration of the possibilities for this method.

Since the raw materials for *Filming Revolution* were provided, as with RavenSpace, it was possible to build a functional replica of the publication's web server and attempt to emulate it. The ability to emulate software is dependent on the ability to encapsulate it on a single virtual machine - in other words, remove or modify all external dependencies so that the entire application works offline. Because this publication depended on numerous Vimeo videos and an external font, the site would not function without a connection to the live web and the persistence of all of those connections. The entire application depends on loading a font from the Google API and thus would not load at all once disconnected from the Internet. Because many LAMP website installations are fairly standard, an automated process could be developed if the publisher used a consistent process for building the web server or could provide a script to install all dependendencies onto an empty Linux machine. Portico therefore sought to determine what would need to change within the package provided in order

to develop a scalable emulation process. A new virtual machine was created by installing the LAMP stack; the website files for *Filming Revolution* were copied to Apache Server on the new machine; and the website database was restored to the MySQL server. The publisher supplied an offline copy of all video assets for use in the emulation. The videos were compressed and moved to a file path on the Apache server and renamed to match their Vimeo IDs in order to make the required code change simpler. The code was then modified to use these local copies of the videos instead of the copies on Vimeo. A method was devised to move the web fonts local to the application, a process that could have been simplified if the fonts chosen were local and non-proprietary. Portico also worked with the publisher to remove any data or server information that was not appropriate for access or preservation. The results were demonstrated and shared via an EaaSI instance hosted by Portico for the project.

Code modifications and data cleaning took approximately five days for the Portico technician, an experienced web developer who had never seen the code before. For a scalable approach, Portico could not spend five days on each publication — slightly more time than it took to localize *As I Remember It*. In order to scale this approach, an application would need to be provided to Portico in a form that is ready for a generic LAMP installation. The most efficient way to do this would be for the original developers to design the website with sustainability and encapsulation in mind — ensuring files are local to the application where possible and that there is a simple way to fall back to local functionality for integrations such as Vimeo. When delivering the package to the publisher, the creator would then produce a clean package, which would be provided to the preservation institution. In addition to enhancing preservation, limiting third party dependencies and building in fallback mechanisms would allow the live application to degrade gracefully and reduce future maintenance.

Assessment

The project team found the preservation activities to be a source of significant learning and experience. While we began the project with many hypotheses about what challenges we would find, we discovered many areas that were either unexpectedly resource-intensive or surprisingly easy to address. These challenges and some opportunities to circumvent them are defined in the companion preservation guidelines document. They are both technical and social. Like all preservation challenges, they do not have one solution. Rather, authors, publishers, and preservation services must collaborate to make the best decisions based on the nature of the work at hand, existing resource limitations, and established standards for digital preservation. Decisions often require tradeoffs that affect the user experience, functionality, or level of confidence in long-term preservation. This section will outline those challenges at a high level and connect them to examples or patterns observed during the project.

One area of widespread concern is the embedding of features that depend on third party services. Video from YouTube, Vimeo, or other video services, social media feeds, data visualizations, or even whole web pages that are embedded within a publication present recurring challenges to preservation. We hope that our recommendations—such as hosting resources locally or procuring copies of third party resources, providing clear descriptive and rights metadata, and avoiding HTML inline frames (iframes) that present content on third-party websites —will mitigate some of the risks, but there are two important tensions here in scholarly communication that scholars and publishers should keep in mind when thinking about preservation:

First, there is a tension between different registers of scholarly communication. Many scholars write in blogs and journals where the freedom and convenience of embedding YouTube videos might outweigh the need for long-term preservation. Scholars who study audiovisual media often make liberal use of embedded third-party resources in online publications. Some of these publications fall into conventional conceptions of the scholarly record; some do not. But as Lavoie and others have discussed, the concept of the scholarly record has itself evolved in the age of digital publication:

> The scholarly record, by virtue of its transition to digital formats, is now much more mutable and dynamic than in the past; it is made available through a blend of both formal and informal publication channels; and the scholarly record's boundaries are expanding to include a much wider context surrounding the

publication of a scholarly outcome. (Lavoie, et al, 2014, pp. 8–9)

A number of University of Minnesota Press publications on Manifold are drafts or "in-process" work from their Forerunners series. Conventions around the scholarly record would indicate that drafts are not part of the scholarly record and do not require long-term preservation for the sake of that record; however, insofar as these publications are used, cited, and themselves participate in scholarly discourses, they may still represent important pieces of discourse that would leave lacunae in the scholarly record.

The second relevant tension for scholars and publishers is around innovation. For some digital humanists and other scholars who make use of digital tools, innovative technologies do not simply enhance scholarship but rather constitute an essential element of their work. For these scholars, some technologies that can hinder preservability also allow for new kinds of scholarly communication. Networking between one publication and other online content is sometimes an essential component of digital scholarship, whether that content is data, software, social media content, audiovisual content, or text-based work on the web. The authors and publishers of *Cut/Copy/Paste* and *The Chinese Deathscape* may regard the interactive maps in these works to be essential; in cases like these, more preservation-friendly alternatives are suitable only to the extent that authors and publishers consider them suitable within the vision of work.

With these tensions in mind, there are a few activities in which scholars, publishers, platforms, and communities can engage to mitigate the challenges of preserving new forms of digital scholarship. The first is to make preservation a formal and intentional part of workflows, training, and professional development. When scholars, production staff, and platform developers think about even the most basic tenets of digital preservation, they can make decisions and ask questions that will prevent preservation roadblocks when their work reaches preservation services. If authors and publishers are in the habit of procuring copies of remote resources, for instance, preservation services can generally be confident that they will be preserved, despite many potential uncertainties in web archiving and emulation.

Second, all parties involved should use and continue to develop standards that help preservation services to scale the preservation of digital publications. This includes using non-proprietary, broadly supported, and widely adopted open file formats; building websites according to web standards for development and accessibility; and

following (and making sure vendors follow) the recommended standards for EPUB and PDF. It also means advocating for new standards, such as a standard for archival EPUBs, that would help both scale the preservation of EPUBs and further guide publishers toward preservability even when the full standard cannot be met.

The companion document to this report, "Guidelines for Preserving New Forms of Scholarship," was written in response to the work of this project. The guidelines suggest ways to make new forms of digital scholarship—from ebooks in conventional formats enhanced with interactive resources to digital publications custom-built using web technologies—in ways allow for more reliable and scalable preservation. While they build on decades of experience and scholarship on digital preservation and digital publishing, these guidelines represent the beginning of an intervention with publishers and platform developers, who, we assert, must now take part in a preservation process that in the past did not require their participation. As such, we expect to build on these guidelines as they are put into practice, and as we learn more from publishers, developers, and preservation service providers.

Works Cited

Lebow, A. (n.d.). *About. Filming Revolution*. Retrieved November 2, 2021, from http://www.filmingrevolution.org

About Scalar. (n.d.). Retrieved November 2, 2021, from https://scalar.me/anvc/scalar/

About Us – RavenSpace. (n.d.). Retrieved November 2, 2021, from https://ravenspacepublishing.org/about-us

EPUBCheck. (n.d.). Retrieved November 2, 2021, from https://github.com/w3c/epubcheck

Greenberg, J., Hanson, K., & Verhoff, D. (2021). *Guidelines for Preserving New Forms of Scholarship*. NYU Libraries https://doi.org/10.33682/221c-b2xj

Kunze, J., Littman, J., Madden, E., Scancella, J., & Adams, J. (2018). *The BagIt File Packaging Format (V1.0)*. https://www.ietf.org/rfc/rfc8493.txt

Lavoie, B., Childress, E. R., Erway, R., Faniel, I. M., Malpas, C., Schaffner, J., & van der Werf, T. (2014). *The Evolving Scholarly Record*. OCLC. https://www.oclc.org/research/publications/2014/oclcresearch-evolving-scholarly-record-2014-overview.html

Mulliken, J. (2020, August 12). Archival Success! *SupDigital*. http://blog.supdigital.org/sup-webrecorder-partnership

Waters, D. J. (2016, July 22). *Monograph Publishing in the Digital Age*. Shared Experiences Blog. https://mellon.org/shared-experiences-blog/monograph-publishing-digital-age/

Appendix A: Publications Analyzed

FULCRUM

Chaudhuri, U., & Hughes, H. (2014). *Animal Acts: Performing Species Today*. University of Michigan Press. https://doi.org/10.3998/mpub.5633302

Clarke, J. R., & Muntasser, N. K. (c2014). *Oplontis: Villa A ("of Poppaea") at Torre Annunziata, Italy. Volume 1. The Ancient Setting and Modern Rediscovery*. American Council of Learned Societies. https://hdl.handle.net/2027/heb.90048.0001.001

Gere, A. R. (2019). *Developing Writers in Higher Education: A Longitudinal Study*. University of Michigan Press. https://doi.org/10.3998/mpub.10079890

Gray, J. (2017). *Show Sold Separately: Promos, Spoilers, and Other Media Paratexts*. New York University Press. https://hdl.handle.net/2027/fulcrum.w0892995q

Lloyd, T. C., & Mullen, P. B. (1990). *Lake Erie Fishermen: Work, Identity, and Tradition*. University of Illinois Press. https://hdl.handle.net/2027/heb.33109

Optiz, R., Mogetta, M., & Terrenato, N. (2016). *A Mid-Republican House From Gabii*. University of Michigan Press. https://doi.org/10.3998/mpub.9231782

MANIFOLD

Boluk, S., & LeMieux, P. (2017). *Metagaming: Playing, Competing, Spectating, Cheating, Trading, Making, and Breaking Videogames*. University of Minnesota Press. https://doi.org/10.5749/9781452958354

Massumi, B. (2018). *On the Revaluation of Value: A Postcapitalist Manifesto* (Draft Text). University of Minnesota Press. https://doi.org/10.5749/9781452958484

Trettien, W. (2021). *Cut/Copy/Paste: Fragments of History* (Draft Chapters). University of Minnesota Press. https://doi.org/10.5749/9781452958576

Wythoff, G. (Ed.). (2016). *The Perversity of Things: Hugo Gernsback on Media, Tinkering, and Scientifiction*. University of Minnesota Press. https://doi.org/10.5749/9781452958422

RAVENSPACE

Paul, E. (2019). *As I Remember It: Teachings (ʔəms taʔaw) from the Life of a Sliammon Elder*. RavenSpace Publishing. https://doi.org/10.14288/SNS9-9159

DLXS

Brennan, S. (2018). *Stamping American Memory: Collectors, Citizens, and the Post*. University of Michigan Press. [The version analyzed, on Michigan's DLXS platform, is no longer available from the publisher. The newer Fulcrum version does not contain the Google Trends data visualization.]

SCALAR

Delmont, M. F. (2019). *Black Quotidian: Everyday History in African-American Newspapers*. Stanford University Press. https://doi.org/10.21627/2019bq

Liebhaber, S. (2018). *When Melodies Gather: Oral Art of the Mahra*. Stanford University Press. https://doi.org/10.21627/2018wmg

Sullivan, E. A. (2020). *Constructing the Sacred: Visibility and Ritual Landscape at the Egyptian Necropolis of Saqqara*. Stanford University Press. https://doi.org/10.21627/2020cts

OPEN SQUARE

Gray, J. (2010). *Show Sold Separately: Promos, Spoilers, and Other Media Paratexts*. New York University Press. https://opensquare.nyupress.org/books/9780814733158/

Jenkins, H., Shresthova, S., Gamber-Thompson, L., Kligler-Vilenchik, N., & Zimmerman, A. (2016). *By Any Media Necessary: The New Youth Activism*. New York University Press. https://opensquare.nyupress.org/books/9781479829712/

STANDALONE WEBSITES

Helms, J. (n.d.). *Rhizcomics*. Michigan Publishing. Retrieved November 2, 2021, from https://www.digitalrhetoriccollaborative.org/rhizcomics/

Lebow, A. (2018). *Filming Revolution: A meta-documentary about filmmaking in Egypt since the Revolution*. Stanford University Press. https://doi.org/10.21627/2018fr

Mullaney, T. S., ed. (2019). *The Chinese Deathscape: Grave Reform in Modern China*. Stanford University Press. https://doi.org/10.21627/2019cd

Appendix B: Acceptance Criteria Template

The goal of the Acceptance Criteria document is to create and record an actionable plan for transferring, preserving, and viewing the work as outlined by the publisher and preservation partner. All monitoring and changes to the original plan will be recorded in this document.

Name of Preservation Partner
Name of Publisher
URL of original work
Link to Mock-up

SECTION A: PRE-TRANSFER ACTIVITIES

Preservation objectives

A1. Brief description of the agreed upon goals for this publication - what content and behavioral elements are we attempting to preserve? This is the criteria or basis upon which we will plan and evaluate the preservation activities. [Publisher]

Transfer of content to preservation partner

A2. Describe and document how the package will be collected, e.g. scraped from a website vs publisher delivered by FTP. What is the frequency and volume of transfer? Note any special tools for transfer used. [Preservation Provider]

Describe contents submitted for preservation

Summary

A3. A very general description of content sent to or made available to the preservation provider. e.g. a zip file which contains the contents of an e-book. [Publisher]

Submission information package (For file transfers)

A4-1. Describe in detail the content sent. This is a full description of the file types, what each file or group of files represents and how the files together from the work to be preserved. Include metadata and any information about how the metadata is mapped to the corresponding files. This might be descriptive, technical and/or structural metadata. [Publisher and Preservation Provider together]

Define content sets and starting points (For web transfers)

A4-2. Describe in detail the content made available for web harvester. [Publisher and Preservation Provider together]

Define parameters for full-server disk imaging / VM migration (For emulation)

A4-3. Describe in detail the content made available for imaging and transfer to emulated hardware. [Publisher and Preservation Provider together]

Playback / experience instructions

A5. What should the intended audience be able to do when this archived content is made available? Provide instructions and necessary context for the playback or reading experience of the material submitted. What features are possible with the files included? What is absent or limited? e.g. embedded remote YouTube videos which are not included. Are there software dependencies for playback that are not in the package? e.g. a 3D viewer. [Publisher]

A6. Are there relationships between the components that need to be explained? e.g. the relationship between an EPUB monograph and annotations saved as separate JSON files will need some context about how they were originally connected/presented. [Publisher]

SECTION B: PRESERVATION ACTIVITIES

Assessment of submitted materials

Archive evaluates contents of a submitted package. This section describes the iterative and final decisions about what to preserve and any concessions made. [Preservation Provider]

Preservability concerns

B1. Document any concerns about the overall preservability of the material e.g. say an EPUB3 package references external JavaScript.

Documented decisions

B2. Make notes on whether changes were made to the original submission package, or when there are decisions around acceptable limitations, or if additional tickets were created during this process.

Preparation for Archiving

B3. Use this section to document the process of converting to a package that can be archived. [Preservation Provider]

Workflow

B4. How was the submitted package reorganized for ingest into the archive? Was anything excluded (e.g. thumbs.db), unzipped, validated, or migrated?

Metadata

B5. Was any metadata added or transformed during the process?

Archived package

B6. Description of the archived package

Access to Archived Copy

Describe how to view the archived copy including a description of the user experience. [Preservation Provider]

How to access

B7. Document how a user can view the preserved copy

Dissemination package

B8. Description of what the user will see in the archive.

UI changes

B9. Description of any changes to the access interface to support presentation of the material

SECTION C: ASSESSMENT ACTIVITIES

Evaluation

These questions will be asked at the end of each cycle in order to feed into the guidelines and to draw out the information inferred in other sections.

Playback / experience

C1. How closely did the archived content available match with the preservation goals

and publisher's requirements / expectations about what would be preserved? Will the intended audience be able to experience this archived content as initially proposed by the publisher?

What was preservable using current tools?

C2. What could and could not be preserved. Explain the constraints – were there technical limitations based on what can be managed using existing tools? or fiscal based on what is financially feasible or possible in the time frame provided? The former might be something like the preservation of embedded YouTube videos, the latter might be user data accessible only by logging in.

Guidelines for better preservability

C3. What could the publisher/author have changed during the creation of the original work to improve the preservability of the material while maintaining the message of the content?

Alternative approaches to consider

C4. What were other approaches considered in determining preservation of the work? What resources would have been necessary to execute an alternative methodology?

Appendix C: Enhancing Services to Preserve New Forms of Scholarship Project Participants

Jonathan Greenberg, NYU Press / NYU Libraries
Winter Scarlett, NYU Libraries
Deb Verhoff, NYU Libraries
Thib Guicherd-Callin, LOCKSS, Stanford Libraries
Fei Li, LOCKSS, Stanford Libraries
Craig Van Dyck, CLOCKSS
Karen Hanson, Portico
Patrick Goussy, Michigan Publishing
Jeremy Morse, Michigan Publishing
Susan Doer, University of Minnesota Press
Daniel Ochsner, University of Minnesota Press
Terence Smyre, University of Minnesota Press
Zach Davis, Cast Iron Coding
Jasmine Mulliken, Stanford University Press
Krista Bergstrom, RavenSpace, UBC Press
Crystal Chan, RavenSpace, UBC Press
Darcy Cullen, RavenSpace, UBC Press
Kellen Malek, RavenSpace, UBC Press
Meaghan McAneeley, RavenSpace, UBC Press
Ali Serag, RavenSpace, UBC Press
Erik Loyer, Scalar