RESCUING LEGACY DIGITAL COLLECTIONS

Lessons Learned from Migrating Historical Bespoke Digital Collections

Kayla Maloney
The University of Sydney
Australia
kayla.maloney@sydney.edu.au
0000-0001-6247-3944

Katrina McAlpine
The University of Sydney
Australia
katrina.mcalpine@sydney.edu.au
0000-0002-2305-3661

Jennifer Stanton
The University of Sydney
Australia
jennifer.stanton@sydney.edu.au
0000-0002-6285-1340

Abstract – The University of Sydney Library hosts many historically significant digital collections. In 2021 and 2022, the Library undertook a project to ensure the accessibility of these collections, migrating them from ageing web servers to our current repository systems. This paper outlines the challenges involved in managing bespoke legacy collections at an institution in the early stages of building digital preservation capacity. We discuss the approaches taken to make use of existing systems, capabilities, and resourcing to rescue collections and prepare for future preservation actions.

Keywords – Digital humanities, Legacy digital content, Data curation, Sustainability, Digital preservation

Conference Topics – We're All in this Together; Sustainability: Real and Imagined.

I. INTRODUCTION

The University of Sydney Library was an early adopter in creating and supporting digital cultural collections. The Library has been hosting online digital collections since 1996, and by 2021 had on the order of 85 different collections being hosted on 15 servers.

Content across the collections varied widely. The collections included historical photographs, digitized manuscripts and images from Rare Books and Special Collections, transcriptions of handwritten content, an archive of archeological grey literature, artworks produced by staff and students from the University's Sydney College of the Arts, and an archive of audio files of Australian adolescents' speech from the 1960s, to name a few. The collections comprise historically relevant content, particularly in an Australian context, and document early digital humanities projects and experiments in using technology and online display in novel ways.

This paper discusses a project to migrate these collections to more modern systems, keeping this historic content accessible and usable for the future, without having a mature digital preservation program in place. We discuss some of the challenges encountered working with legacy collections and infrastructure. We hope that our project can provide insights for people working with non-standardized, bespoke content where there may not be an obvious "right" way forward.

II. BACKGROUND

Despite the experimental nature of several of the collections, little intervention from Library staff was required to keep them online and available over the decades. Consequently, a lot of the institutional knowledge around the collections was gone by the time the Library started this migration project

Library staff have been exploring issues around these legacy collections and how they should be managed since 2017. However, getting a comprehensive picture of the entirety of our content

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/121121

was not straightforward. To save the cost of setting up additional servers, new collections were often added to existing servers, resulting in a complex web of links and sometimes orphaned pages. Some of the servers were originally physical, virtualized years later, and finally, years later again, were moved to the cloud. They were beyond their end-of-life and no longer fit for purpose.

To properly tackle this situation, we needed someone with the appropriate technical skills to dedicate a large amount of time to investigate the collections and determine appropriate solutions for different cases. However, Library IT staff were in high-demand and there were few staff with the skillset needed to navigate the ageing servers. Over the years, at least four different people started to investigate and audit the content on separate occasions, only to be pulled away when urgent tasks elsewhere required attention.

During this time, the Library began to invest in a digital preservation program. Staff undertook training and development activities, including iPres conference attendance, the Digital Preservation Coalition's 'Novice to Know How' course, completing digital preservation maturity modelling and implementing some digital preservation workflows for digital collections. Overall, however, digital preservation at the Library was still in its infancy and a digital preservation framework or system had not been implemented.

In April of 2021, rising institutional cyber security concerns led to a deadline for upgrading or shutting down the legacy collection servers. This was no longer a task that could be put on the backburner until we had the time to do it "properly".

Staff from the Digital Collections team, Library IT and the Sydney University Press compiled a comprehensive list of collections from the legacy servers, based on the earlier audit. The team looked to projects at other institutions on managing and preserving bespoke digital humanities collections to develop approaches for rescuing and migrating the content in our collections [1]. Each collection was assessed for whether it should be kept, and where and how it should be migrated. Tasks were assigned to the appropriate team, and everyone got to work.

III. Challenges

A. Have we found everything?

Gaining a comprehensive understanding of all the collections on our servers had been a major roadblock to getting started on this project for years, and the worry that we might be missing something was with us throughout the entirety of the project.

To ensure that we had a copy of all content, the final step in decommissioning each server was to archive all content and configuration files and put the archive on the University's Amazon Glacier storage. Concerns emerged at one point that two of our more unstable servers could fail before being properly decommissioned. Due to staff availability, we were unable to undertake priority archiving of these servers according to our established process. As a stop-gap measure, team members attempted to use the MacOS application SiteSucker to get a local emergency backup copy of these servers [2]. This was successful for one of the servers, but SiteSucker struggled to capture the entirety of our most complicated server, and we were left with an incomplete emergency backup. Fortunately, both servers remained functional until they were able to be properly archived and decommissioned.

These backups mitigated the risk of data loss, however, they did not solve the problem of knowing what content we needed to migrate, and understanding how that content displayed and functioned in its original context.

Where SiteSucker worked, it provided us with the additional benefit of easily accessed working copies of our content and insights into where we had content that we had not yet identified. We also manually combed through the sites and tried web searches to turn up orphaned pages still hosted and accessible, but no longer linked to from the main pages of the sites. Some orphaned pages were only discovered through serendipity, for instance, a team member finding a reference to a collection in historical documentation, or an inquiry from a member of the public. These finds helped us move forward, but also highlighted the likelihood that we were missing content from our migration plan.

For websites that hosted large numbers of files available for download, such as PDFs, we used the browser extension Simple Mass Downloader to obtain local working copies of the files for migration [3]. This was also helpful for cross-checking with existing and newly created collections metadata, to highlight gaps where we might be missing files or where we needed to create metadata.

Our intention was to migrate content with no downtime, so that the new location would be available prior to removing the old. We eventually reached a point in our checks where we felt the risk of downtime due to missing a collection was acceptably low, and our backups gave us confidence that we would be able to reinstate any content that we missed.

B. Understanding our content

Documentation was uneven across the legacy collections. For some, it was difficult to determine important details such as the copyright owner, agreements that had been made around the collection, who had been involved, or sometimes even why we had it in the first place. This information can be critical in making decisions about what preservation actions can or should be taken for a collection. Statistics around usage and engagement with the different collections would also have been valuable for this decision-making, however, issues with the setup of the servers and the influence of bots meant that we were unable to get trustworthy information.

Interestingly, the fact that many of these collections had continued to remain accessible with minimal intervention over long time periods was a contributing factor to the loss of institutional knowledge. Most of the bespoke collections were built using HTML and we did not need to grapple with the complex issue of preserving custom software. Without problems occurring, no one needed to check in on the collections and staff who had been involved in collection creation left the institution without passing on historical knowledge. For most of the collections, particularly those where the Library was involved in their creation, we were able to turn up the information needed. This took the form of finding historic documentation, relying on institutional memory from some long-term staff, or tracking down contact details from involved parties. In a few cases, the information that we found allowed us to determine that we no longer would make a collection available, for instance, where an agreement had lapsed, or if the purpose that it was made available for was no longer relevant. In some cases where documentation was lacking, we had to decide whether the Library was the best organization to make content available. Other institutions have subsequently digitized some of the same materials at a higher quality. When better versions were openly available elsewhere, we generally opted not to migrate our version.

In all cases, we tried to ensure that the information we turned up and any decisions we made were well documented. Project decisions were recorded in project documentation. Where investigations were required, outcomes from the investigation were detailed in Word documents and stored alongside collection files in our dark archive location. Agreements regarding collections were saved to the University's recordkeeping system and the record numbers were added to administrator metadata for the collection in our repository systems to ensure connection between the information across the systems. A brief statement about the migration was added to items' provenance metadata fields in their new location, visible only to system administrators. We also considered how best to include information for others to use and understand the collections. "About" pages were created detailing the projects that many of the collections belonged to, outlining the history of the projects, funding, references to agreements around collection content and an acknowledgement of the people involved. We also included a link to versions of the sites archived in the Internet Archive's Wayback Machine to allow people to see the original context of the collections. These pages are hosted on our current Digital Collections site.

C. Non-standard structures and scale

The Library no longer hosts servers for individual digital collections to have their own bespoke pages. Instead, we have moved towards having more standardized systems and processes, including the Library's Digital Collections repository [4] (Recollect [5]) and the Sydney eScholarship repository [6] (DSpace [7]) for University research outputs. As repositories, these systems have different affordances to websites. It was not always straightforward to determine how the bespoke, and frequently unusually formatted, website-based content should best be migrated and displayed in a repository system.

The John Anderson Archive provides an example of one of our approaches to unusually formatted content. The Archive presents significant works and papers of John Anderson (Challis Professor of Philosophy at the University from 1927 until 1958) [8]. Among these works are handwritten lecture notes. The original form of the Archive presented transcriptions of the notes as HTML text on the website. Each transcribed page included a link to an image of the original handwritten text. The handwritten notes also included asides, often indicated by text in square brackets. The asides were included in the transcriptions as hyperlinked notes that opened in a separate pop-up window. Significant reformatting was needed to be able to include this content in our Digital Collections repository. The transcribed text was copied and pasted into a Word document, preserving the page numbering of the original text. The hyperlinked notes were included as footnotes. Each document was saved as a PDF and uploaded to the repository. The JPEG images of the original handwritten notes were combined and saved as a PDF and uploaded to the repository as a separate item to the transcriptions. In this way, we were able to preserve the content of the original archive, although not the rather experimental functionality of the linked notes and images. Due to the manual nature of this work, significant resourcing was required. We benefitted from the availability of additional staff, who normally work in client-facing roles, during lockdowns and periods of reduced services during the COVID-19 pandemic.

Other collections were large enough that a manual approach was not feasible. A collection of archeological reports [9] and another of photographs of artworks produced at the University's Sydney College of the Arts [10] each contained well over 1,000 items. No reformatting of the content was needed, however, collection item metadata needed to be combined, mapped, and transformed for ingest to our Digital Collections repository. The artwork metadata was originally stored in a relational database, where many images, each with their own metadata, could belong to a single artwork. We needed to transform this to a flat tabular structure. To do this, we used the pandas Python library for the data wrangling and Jupyter notebooks to allow us to document our code in a more readable fashion for future reuse. We also took the opportunity to involve team members with no coding experience to enable knowledge-sharing and the development of new skills across our team.

D. Digital preservation maturity

The Library was, and at the time of writing still is, in the early stages of implementing a digital preservation program. Ideally, we would have undertaken this migration project with a more mature digital preservation program and an appropriate digital preservation system in place, however this was not an option. Throughout this project we were able to apply some digital preservation practices such as using tools like TeraCopy to transfer files, ensuring there were back up files created and stored and that the project was well documented. However, we were, and are, aware that there were many processes we could not complete due to lack of time and an established preservation framework. This was challenging, as we knew throughout the project that there were digital preservation good practices we were not following, and that there would be extensive future work to undertake to enable us to move our content into a digital preservation system.

IV. Lessons Learned

Our main lesson from this work is that we cannot let the desire for a perfect solution prevent us from getting started. We do not want to go in and start doing work without considering issues and having a plan, but if getting that plan completely "right" means important work never gets started, we need a different approach. Not all issues can or should be solved upfront, and we can work through problems as they come. This may lead to stress when something unexpected crops up, or we realize that we have overlooked something; not everything will be done in the ideal way. Even with these bumps along the way, it is a far better outcome than never getting started and losing everything.

Documentation is critical for being able to appropriately manage and preserve content, but historical practices have not always given us the information that we need. This includes information about copyright holders, agreements and reuse conditions, project stakeholders, and collection outcomes and impact. Tracking this information down can take a lot of resourcing. Where needed, taking a risk-management approach can help us to make acceptable decisions. Whatever happens, it is essential to set ourselves up better for the future by documenting this important information and what we have done using the tools available to us, including recordkeeping systems, collection metadata, project histories and project documentation.

We also learned that we should consider whether content, functionality or both need to be retained when migrating to new systems. Our systems did not always allow us to preserve the functionality of the content we migrated, however, this web-based content had been archived by the Wayback Machine, allowing us to link to earlier versions to provide users with the initial context for the collection.

Collections may be hosted in one place, but over time, they will be harvested and linked to elsewhere. Any time collections move, issues will appear in the network of places they now exist in. Permalinks can help to mitigate this issue, but they will not entirely solve it. Issues can be chased down over time as they are noticed, and this should be seen as something to be aware of, but not something that we can fully plan for from the beginning of a project.

Finally, a project like this will require a large range of skills to complete. Wherever possible, we tried to prioritize and make the time to share knowledge and skills. This will mean that some tasks take longer than if the staff member with the most knowledge completes them fully. Particularly in areas where only one staff member has a skill, the growth in team capacity is well worth this extra time.

V. Future Considerations

This paper has outlined a project to rescue legacy collections from being lost entirely. The current systems that they have been migrated to are repository systems that enable access but are not preservation systems. The University of Sydney is increasingly interested in digital preservation, and there is likely to be future institutional support for growing our digital preservation capacity. The actions taken to standardize collections in this project will assist us in future preservation activities and working with future systems.

Digital humanities projects and bespoke digital collections similar to those addressed by this project are still being created. Migrating the collections has given the Library and the University further insights into what needs to be considered for managing these projects and outputs in the future. Do we need to be creating service level agreements for ongoing support of collections? What information, agreements and documentation do we need to have to ensure that we can manage and preserve a collection throughout its life? What constitutes end-of-life for a collection or project, and what should happen next? These are some of the questions that we are grappling with as we plan our future digital preservation program.

VI. Acknowledgements

The authors would like to acknowledge everyone involved in this collection migration project, including Sarah Graham, Susan Murray, Susan Brazel, Phil Jones, Marthe Follestad, Piyachat Ratana, Ryan Stoker, Arin Bryant-Munoz, Dora Zhang, Wayne Zhang, Keerat Judge, and Rengen Parlane. Thanks to Nicholas Keyzer and Anthony Green from Schaeffer Library for providing us with images and metadata for the Sydney College of the Arts Archive. We would also particularly like to thank Jim Nicholls and Kim Williams, who both played major roles in the project and provided us with information and feedback on this paper.

1. REFERENCES

[1] J. Smithies, C. Westling, A.-M. Sichani, P. Mellen, and A. Ciula, “Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King’s Digital Lab,” *Digit. Humanit. Q.*, vol. 13, no. 1, 2019, [Online]. Available: http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html

[2] “SiteSucker for macOS.” https://ricks-apps.com/osx/sitesucker/index.html (accessed Mar. 03, 2023).

[3] “Simple mass downloader - Chrome browser extension.” https://chrome.google.com/webstore/detail/simple-mass-downloader/abdkkegmcbiomijcbdaodaflgehfffed (accessed Mar. 03, 2023).

[4] “Digital Collections | University of Sydney Library.” https://digital.library.sydney.edu.au/ (accessed Mar. 03, 2023).

[5] “Recollect - Collection Management and Community Engagement Software.” https://www.recollectcms.com/ (accessed Mar. 03, 2023).

[6] “Sydney eScholarship Repository.” https://ses.library.usyd.edu.au/ (accessed Mar. 03, 2023).

[7] “DSpace.” https://dspace.lyrasis.org/ (accessed Mar. 03, 2023).

[8] “John Anderson Archive.” https://digital.library.sydney.edu.au/nodes/view/6932 (accessed Mar. 03, 2023).

[9] “NSW Archaeology Online.” https://digital.library.sydney.edu.au/nodes/view/6929 (accessed Mar. 03, 2023).

[10] “Sydney College of the Arts Archive.” https://digital.library.sydney.edu.au/nodes/view/6927 (accessed Mar. 03, 2023).