IIPC DISCRETIONARY FUNDING PROGRAM 2020-2021

FINAL REPORT

IMPROVING THE DARK AND STORMY ARCHIVES FRAMEWORK BY SUMMARIZING THE COLLECTIONS OF THE NATIONAL LIBRARY OF AUSTRALIA

2021-12-17

LEAD IIPC INSTITUTION: Old Dominion University (ODU)
2ND IIPC INSTITUTION: Los Alamos National Laboratory (LANL)
OTHER INSTITUTIONS OR CONSULTANTS: National Library of Australia (NLA)
PROJECT TEAM MEMBERS:
Michael L. Nelson, ODU
Martin Klein, LANL
Michele C. Weigle, ODU
Paul Koerbin, NLA
Alex Osborne, NLA
Shawn M. Jones, LANL
Himarsha Jayanetti, ODU
AWARDED FUNDING (IN USD)**:** 50,000
PROJECT WEBSITE: https://netpreserve.org/projects/dark-and-stormy-archives/

BRIEF ABSTRACT OF THE PROJECT:

Our goal was to develop the Dark and Stormy Archives (DSA) toolkit to the extent that it is generally applicable to all Memento-compliant web archives. We piloted these modifications with the Memento-compliant archive of the National Library of Australia (NLA). The DSA toolkit improves the summarization, abstracting, understanding, and sharing of collections of archived web pages. This project has extended work begun in the (now complete) IMLS grant "Combining Social Media Storytelling With Web Archives". The five tools that make up the DSA toolkit are: AIU, Off-Topic Memento Toolkit (OTMT), MementoEmbed, Raintale, and Hypercane. Our pilot with the NLA has served as a launching point for future collaborations with other interested organizations.

The DSA framework's premise is to employ "storytelling" techniques to summarize archival collections by choosing a small number of exemplar pages from the collection that best demonstrate what the collection is "about". This approach is beneficial over providing metadata about all pages in the collection since if a collection has 100s of seeds and 100s of mementos per seed, the collection quickly overwhelms conventional UIs. Rather than deploying esoteric or custom

interfaces for summarizing the collection, we have demonstrated the potential of our approach of summarizing the collection with a small number of pages selected from the entire collection. The varying story and collection types influence which pages are selected and how they are arranged. Our collaboration with the National Library of Australia has provided additional insight necessary to develop future standards that enable sharing collection information.

PROJECT OUTPUTS / OUTCOMES:

An updated DSA software suite consisting of the below components is the core deliverable:

- AIU (formerly Archive-It Utilities) is a Python library that gathers seed URLs and metadata from web archive collections.
- The Off-Topic Memento Toolkit (OTMT) identifies off-topic mementos in a collection.
- MementoEmbed produces surrogates (e.g., cards, screenshots) of individual mementos.
- Raintale leverages MementoEmbed to produce full stories of mementos suitable for publishing via static files or social media.
- Hypercane provides tools for selecting exemplar mementos and generating metadata from a collection. These exemplars and metadata can then be fed into Raintale to produce stories that summarize collections.

Based on feedback from our collaboration with NLA, we have implemented new versions of these software packages.

AIU was initially named “Archive-It Utilities” - we renamed it because it now provides seed URLs and metadata for collections in Archive-It, Trove, and Pandora. The original name was no longer applicable. Where Archive-It only has one type of collection, NLA has three distinct types of collections: Trove collections, Pandora Subjects, and Pandora Collections. We released four new versions of AIU through this grant to address defects and support these new collection platforms. OTMT required minimal changes for this project. We created a single new release that incorporates fixes found during development.

MementoEmbed required changes to support Pandora and Trove despite the NLA's Memento protocol support. First, Trove brands each memento with logos and navigational elements. As we reported during our TSS presentation in August, MementoEmbed requires mementos' actual content, and Trove’s branding creates noise that affects its analysis. We updated MementoEmbed to be aware of this branding so its natural language processing could proceed as expected. Additionally, Trove contains some mementos that were imported from Pandora. These mementos’ metadata reflected when they were imported, not when they were initially crawled. Their metadata also listed Pandora as their original resource rather than the page that Pandora had initially captured. We updated MementoEmbed with heuristics to discover this additional metadata specifically for mementos from Pandora and Trove.

Raintale requires that an archivist create and supply a template so that the resulting story meets the branding needs of an institution. Templates also allow archivists to decide which metadata to include in stories. Early in the project, NLA opted for HTML output as a target for their Raintale stories. In addition to developing a new carousel and a wide HTML template, we helped NLA develop Raintale templates specific to their archive for sharing these stories. This collaboration helped us evaluate our Raintale documentation to ensure that future users could easily create their own templates.

To tell a story that summarizes a collection or features some aspect of it, archivists need to select a small set of exemplars from the collection. Hypercane automates this process through probabilistic and intelligent sampling algorithms. We recognized through this project that dictating a single algorithm for summarizing collections would not suffice for the many potential use cases of the DSA toolkit. Instead, we focused on creating some core algorithms while providing a set of primitives that help users create their own algorithms. Before this project, Hypercane supported three core algorithms. Now it supports 14. Hypercane provides eight primitives – sample, filter, cluster, score, order, identify, report, and synthesize. At the beginning of this project, these primitives supported 39 different operations; now, they support more than 70. Thanks to our collaboration with the NLA, we recognize that archivists may need “recipes” rather than just “algorithms” and have been incorporating these recipes as needed to benefit other DSA users.

Making the DSA’s user-facing tools more accessible to those without computer science backgrounds is one of this project’s goals. Creating a graphical interface for these tools was critical toward meeting this goal. MementoEmbed had an existing web user interface (WUI) before this project, but Raintale and Hypercane were only accessible via a command-line interface (CLI). We discovered the Wooey framework, which helps developers provide a WUI for CLI applications. Wooey provides graphical elements (e.g., textboxes, menus) so that users can supply the same values they would supply for the CLI application, but in a much more user-friendly fashion. Adopting Wooey required that we rewrite portions of Hypercane and Raintale to be compatible with this new interface, but the benefits of Wooey outweighed the challenges introduced by this rewriting. With Wooey, multiple users can log in and run concurrent Raintale or Hypercane jobs, helping institutions service many archivists’ projects simultaneously. We also ensured that we replicated the CLI functionality present in Hypercane and Raintale in their corresponding WUIs.

When the project started, installing MementoEmbed, Hypercane, or Raintale required that a user be a Python and database expert. We had only provided install packages with Docker, which some organizations cannot implement. Thanks to feedback from NLA and other members of the IIPC community, we now provide a native RHEL 8/CentOS 8 installer for each product that handles installing dependencies and setting up databases. We also have Ubuntu 21.04 installers for MementoEmbed and Raintale. We achieved our goal of helping users or administrators install and run these tools like any other professional software package.

We are still receiving feedback from NLA and incorporating fixes as our partners discover them. We will release new versions of the software as these fixes become available.

OTHER RESULTS (IF APPLICABLE):

We used this project's work to promote the DSA and IIPC whenever possible. We promoted new features enabled by this grant primarily on Twitter and included posts to Facebook and LinkedIn whenever possible. With every social media post, we thanked the IIPC for its support. The New Mexico Consortium, a non-profit body that facilitates research between LANL and other universities, also advertised this grant in February 2021.

During IIPC WAC 2021 in June, we presented our work with MementoEmbed and Raintale. We described the architecture and highlighted the work we had accomplished with NLA. This was the first time we featured some early NLA-specific Raintale templates for the audience. We also presented our status on this project in August and December as part of the IIPC's Technical Speaker Series (TSS).

Also, in June, we presented "Automatically Selecting Striking Images for Social Cards" at ACM Web Science 2021. Hypercane's reporting abilities allowed us to study which image best summarized a given document. We will be applying our results to MementoEmbed.

As part of this research, we uncovered patterns in the use of web page metadata. MementoEmbed leverages web page metadata as part of its summarization. We discovered that the social media surrogate metadata leveraged by MementoEmbed had experienced a meteoric rise compared to other metadata categories, like Dublin Core. We presented these results in "It's All About The Cards: Sharing on Social Media Probably Encouraged HTML Metadata Growth" at ACM/IEEE JCDL 2021. We also featured Hypercane in a poster at ACM/IEEE JCDL 2021.

We summarized Hypercane in the article "Hypercane: Toolkit for Summarizing Large Collections of Archived Webpages" as part of the Summer 2021 issue of the ACM SIGWEB newsletter. In that article, we highlighted the role of IIPC in funding Hypercane's development.

Perhaps the most impactful result of this grant was its ability to fund the work necessary for Shawn M. Jones to complete his doctoral dissertation "Improving Collection Understanding for Web Archives with Storytelling: Shining Light Into Dark and Stormy Archives." The dissertation features the results of the research that produced the DSA and the results of working with the NLA.

For 2022, we have at least three publications planned based on the research performed thanks to this grant.

ANECDOTAL INFORMATION:

After the initial meeting between NLA, ODU, and LANL, we mitigated timezone issues with email and Slack, as detailed in "Best Practices." In January, we began researching NLA's collections and took care of other project objectives that did not require this initial meeting.

PhD student Shawn Jones became Doctor Shawn Jones during the course of this grant. The additional time needed to prepare and defend a dissertation required some rearrangement of the schedule to ensure both Shawn's and this project's goals could be met. As noted above, his dissertation is now available and features work from this grant.

As he was no longer a student, Doctor Shawn Jones ceased being a graduate research assistant at the Los Alamos Research Library. He was awarded a new postdoc position in LANL's Information Sciences division. This also delayed some project goals as Shawn worked to balance the work from his new position with the work in this project. In the last week, Shawn was awarded a fellowship by the Information Science & Technology Institute (ISTI) and is now an ISTI Postdoctoral Fellow.

BEST PRACTICES:

At the start of this grant, we used GitHub as the desired platform for recording software issues. Unfortunately, keeping a list of issues inside each software repository was insufficient for planning because we have multiple software projects to track. Experiments with spreadsheets and long-form documents proved insufficient. Fortunately, GitHub provides a configurable Kanban board for "Projects" that helps developers coordinate issues and pull requests across multiple GitHub repositories. Once we adopted this board, we used it to organize the rest of the project. In retrospect, we feel we should have started with the Kanban board but were still getting a feel for the breadth of the work and how to manage it.

Drawing from Scrum's standup concept, Himarsha and Shawn kept weekly video meetings to discuss status and issues. Sometimes these meetings would result in a demonstration of completed work. Other times they would become pair programming sessions. We incorporated other best software engineering practices like test-driven development and continuous integration when possible. Changes in Travis-CI's and Docker's free-tier policies interrupted our continuous integration capabilities, and we are exploring alternatives. We employed Docker when testing installs on different Linux distributions, saving us the expense of renting or building many different test machines.

For less formal short-form communications, we applied Slack. It helped us overcome many of the timezone issues inherent with collaborations between both sides of the planet. Often, the North Americans would post something to Slack and get an answer once the Australians were available. Again, we should have started the project this way and will keep this in mind for future projects.

PROGRAM CONTINUITY:

As presented at the December IIPC TSS, there are many directions with which to take these tools, and now that we have engaged in this pilot, we better understand the needs of its users.

Shawn's ISTI fellowship focuses on research related to automatic summarization. This time will provide him with the flexibility needed to perform future experiments and further development with Hypercane. He will investigate new algorithms for selecting exemplars, which will likely involve adding new operations to the existing Hypercane primitives. His work will also focus on extending the summarization beyond documents, creating summaries of mixed-media corpora consisting of video, documents, images, audio, and data. He will provide the results of this work in new releases of Hypercane, as permitted by LANL.

Other parties have expressed interest in these tools. We are in discussions with the Internet Archive to adopt these tools for some of their upcoming initiatives. Our SHARI process provides daily news summaries using the DSA toolkit. The Internet Archive has offered to pilot it on their infrastructure. Recently, we have been contacted by the University of Alberta about using the Off-Topic Memento Toolkit and Hypercane in some research studies. We are continuing our collaboration with NLA. With support like this, we expect the DSA toolkit to continue to receive feedback and improvements.

Our overall goal for the DSA toolkit is to continue to improve its accessibility. We are planning installers for other versions of Ubuntu as well as macOS. The Wooey library was derived from a library named Gooey, which promises to provide a conversion from a CLI application to a native graphical desktop application. If this proves to be the case, we are in good shape to create Linux and macOS desktop apps. As we build upon these technologies, we are charting a path toward making the DSA toolkit a set of native Microsoft Windows desktop applications, expanding its user base even further.

AIU now supports Archive-It, Trove, and Pandora. We can expand it to support collections in the Croatian web archive. We can also adapt it to the collections at the Library of Congress. Once AIU supports these collection types, we can incorporate them into Hypercane, further extending its ability to process public web archives.

To make Hypercane more approachable, we had the idea for a “Recipe Builder” that would allow users to create and save their own recipes built from Hypercane’s primitives. The recipe builder would be an important step in making sure that graphical users can enjoy the same functionality that CLI users enjoy with scripts. Once the Recipe Builder exists, users should be able to share their recipes and we anticipate building a community where they can do so, further aiding in research that helps all process web archives.

Likewise, we would like to build a community of shared templates for Raintale. Our existing templates demonstrate capability, but we see a future where archivists and other users share the look and feel of their stories with each other. Much like with Twitter stories, we have experimented with Facebook as well. We can facilitate uploading Raintale’s video stories to Tiktok, YouTube, or Twitter. We are also considering updates to Raintale so that it can generate stories in non-web formats like PDF or Microsoft Powerpoint. To facilitate building stories outside of templates, Himarsha is considering the nascent idea of a “Story Builder” that would allow users to construct stories graphically in real-time, without a template, much like bloggers do with WordPress blocks.

Thus, with the interest in the DSA toolkit and a plethora of ideas, we see a bright future ahead.