Issue 53, 2022-05-09

The DSA Toolkit Shines Light Into Dark and Stormy Archives

Themed web archive collections exist to make sense of archived web pages (mementos). Some collections contain hundreds of thousands of mementos. There are many collections about the same topic. Few collections on platforms like Archive-It include standardized metadata. Reviewing the documents in a single collection thus becomes an expensive proposition. Search engines help find individual documents but do not provide an overall understanding of each collection as a whole. Visitors need to be able to understand what individual collections contain so they can make decisions about individual collections and compare them to each other. The Dark and Stormy Archives (DSA) Project applies social media storytelling to a subset of a collection to facilitate collection understanding at a glance. As part of this work, we developed the DSA Toolkit, which helps archivists and visitors leverage this capability. As part of our recent International Internet Preservation Consortium (IIPC) grant, Los Alamos National Laboratory (LANL) and Old Dominion University (ODU) piloted the DSA toolkit with the National Library of Australia (NLA). Collectively we have made numerous improvements, from better handling of NLA mementos to native Linux installers to more approachable Web User Interfaces. Our goal is to make the DSA approachable for everyone so that end-users and archivists alike can apply social media storytelling to web archives.

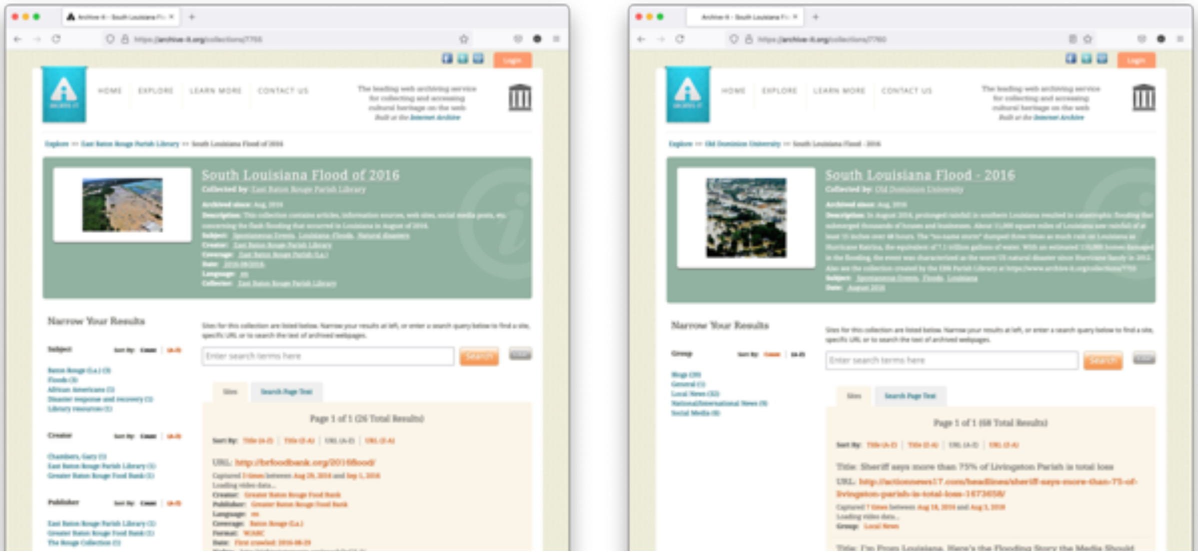
by Shawn M. Jones, Himarsha R. Jayanetti, Alex Osborne, Paul Koerbin, Martin Klein, Michele C. Weigle, Michael L. Nelson

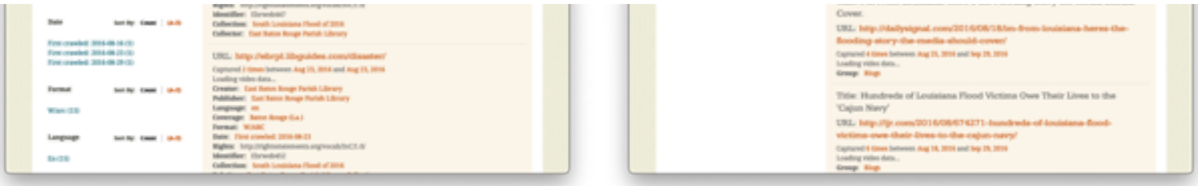
Editor's Note: This article makes use of Robust Links. Next to each hyperlink the reader will discover a menu that allows them to visit an archived version of the linked resource in case the current version has changed or is no longer available. Visit the Robust Links project for tools and more information on combating reference rot.

Web Archive Collections Are Too Large To Understand At A Glance

Web archives are invaluable for a variety of research studies. Historians have analyzed how humans interacted on extinct websites, like Geocities. Social scientists have used them to study the changes in social commerce over time. Journalists can use web archive evidence to bring attention to questionable medical practices and document changes in government policy.

Some archivists create themed **web archive collections** by selecting web pages for preservation that support a topic. Each web page, or **original resource**, can change over time. Archivists capture these original resources at specific points in time, turning each observation into a **memento**. The date and time of capture is that memento's **memento-datetime**. A **TimeMap** contains the set of mementos for an original resource. Archive-It is a popular platform for creating themed web archive collections. Themed collections also exist at the Library of Congress, Conifer, the Croatian Web Archive, the UK Web Archive, and the National Library of Australia's (NLA) PANDORA and Trove collections.





Date Sort By: Count | (A-Z)

First crawled: 2016-08-18 (1)
First crawled: 2016-08-23 (1)
First crawled: 2016-08-29 (1)

Format Sort By: Count | (A-Z)

WARC (33)

Language Sort By: Count | (A-Z)

En (33)

Identifier: [illegible]
Collection: South Louisiana Flood of 2016
Collector: East Baton Rouge Parish Library

URL: http://ebrpl.libguides.com/disaster/
Captured 2 times between Aug 23, 2016 and Aug 23, 2016
Loading video data...
Creator: East Baton Rouge Parish Library
Publisher: East Baton Rouge Parish Library
Language: en
Coverage: Baton Rouge (La.)
Format: WARC
Date: First crawled: 2016-08-23
Rights: http://rightsstatements.org/vocab/InC/1.0/
Identifier: [illegible]
Collection: South Louisiana Flood of 2016

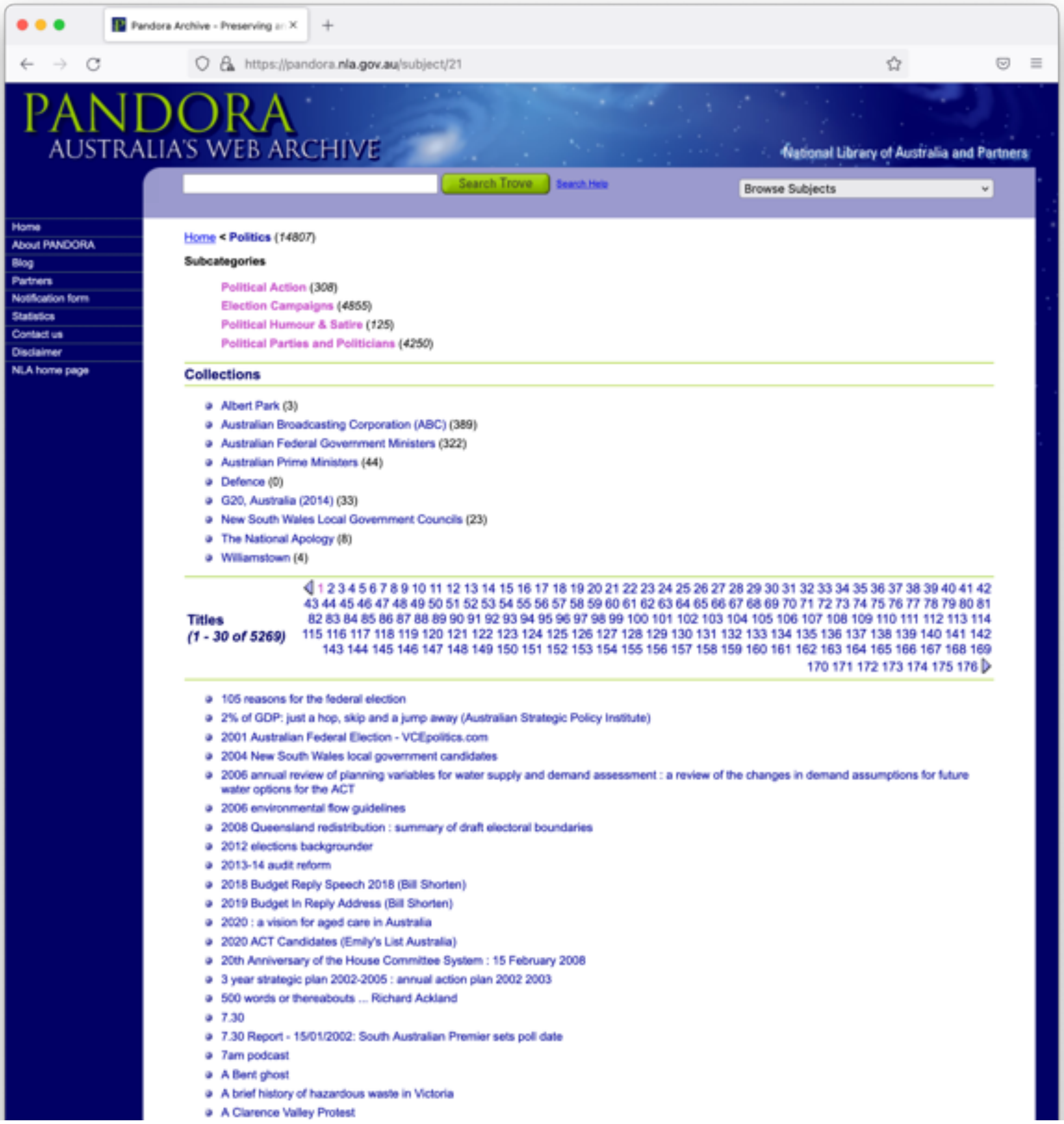
Cover.
URL: http://dailysignal.com/2016/08/18/im-from-louisiana-heres-the-flooding-story-the-media-should-cover/
Captured 4 times between Aug 23, 2016 and Sep 29, 2016
Loading video data...
Group: Blogs

Title: Hundreds of Louisiana Flood Victims Owe Their Lives to the 'Cajun Navy'
URL: http://ijr.com/2016/08/674271-hundreds-of-louisiana-flood-victims-owe-their-lives-to-the-cajun-navy/
Captured 4 times between Aug 18, 2016 and Sep 29, 2016
Loading video data...
Group: Blogs

Figure 1. Archive-It has two collections about the South Louisiana Flood in 2016.
East Baton Rouge Parish Library created the one represented by the left screenshot.
Old Dominion University created the one on the right. (Screenshots taken in June 2021.)

Such topical web archive collections are invaluable to those studying a topic, but it can be difficult for visitors to understand which collection they should begin exploring. Figure 1 shows screenshots of two Archive-It collections containing mementos about the South Louisiana Flood of 2016. Each collection contains different metadata fields. Which one should a researcher use in their project? Alternatively, if a visitor uses Archive-It's collection-level search engine to find collections about a topic like "human rights," they have 48 collections to review as of January 2022. Like *Government of Canada Publications*, some collections contain hundreds of thousands of documents but no metadata to help visitors understand the collection. In a 2019 study (*preprint version*), we evaluated all collections at Archive-It and determined that collections with more original resources have less metadata to help visitors understand them.

Metadata is more consistent in the NLA's PANDORA collections than in Archive-It's collections, but the sheer size of some collections makes it difficult to understand them at a glance. Figure 2 shows a screenshot of the page for the PANDORA subject Politics. This subject is a collection in its own right and contains subcategories and collections so that visitors can view parts of the collection. It also includes a listing of 5,269 page titles. At a minimum, each PANDORA collection contains a collection title and page titles. Some are divided into sub-collections, but a human would still need to review many documents (or at least page titles) to understand the collection.



Home < **Politics** (*14807*)

Subcategories

Political Action (*308*)
Election Campaigns (*4855*)
Political Humour & Satire (*125*)
Political Parties and Politicians (*4250*)

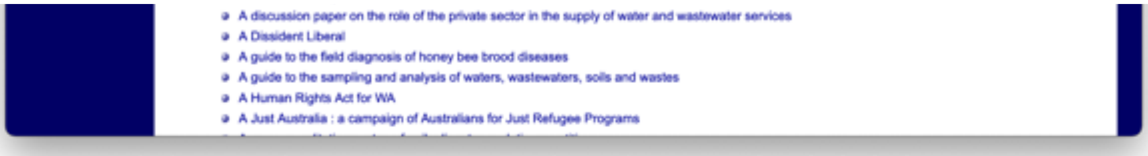
Collections

- Albert Park (3)
- Australian Broadcasting Corporation (ABC) (389)
- Australian Federal Government Ministers (322)
- Australian Prime Ministers (44)
- Defence (0)
- G20, Australia (2014) (33)
- New South Wales Local Government Councils (23)
- The National Apology (8)
- Williamstown (4)

Titles
(1 - 30 of 5269)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176

- 105 reasons for the federal election
- 2% of GDP: just a hop, skip and a jump away (Australian Strategic Policy Institute)
- 2001 Australian Federal Election - VCEpolitics.com
- 2004 New South Wales local government candidates
- 2006 annual review of planning variables for water supply and demand assessment : a review of the changes in demand assumptions for future water options for the ACT
- 2006 environmental flow guidelines
- 2008 Queensland redistribution : summary of draft electoral boundaries
- 2012 elections backgrounder
- 2013-14 audit reform
- 2018 Budget Reply Speech 2018 (Bill Shorten)
- 2019 Budget In Reply Address (Bill Shorten)
- 2020 : a vision for aged care in Australia
- 2020 ACT Candidates (Emily's List Australia)
- 20th Anniversary of the House Committee System : 15 February 2008
- 3 year strategic plan 2002-2005 : annual action plan 2002 2003
- 500 words or thereabouts ... Richard Ackland
- 7.30
- 7.30 Report - 15/01/2002: South Australian Premier sets poll date
- 7am podcast
- A Bent ghost
- A brief history of hazardous waste in Victoria
- A Clarence Valley Protest



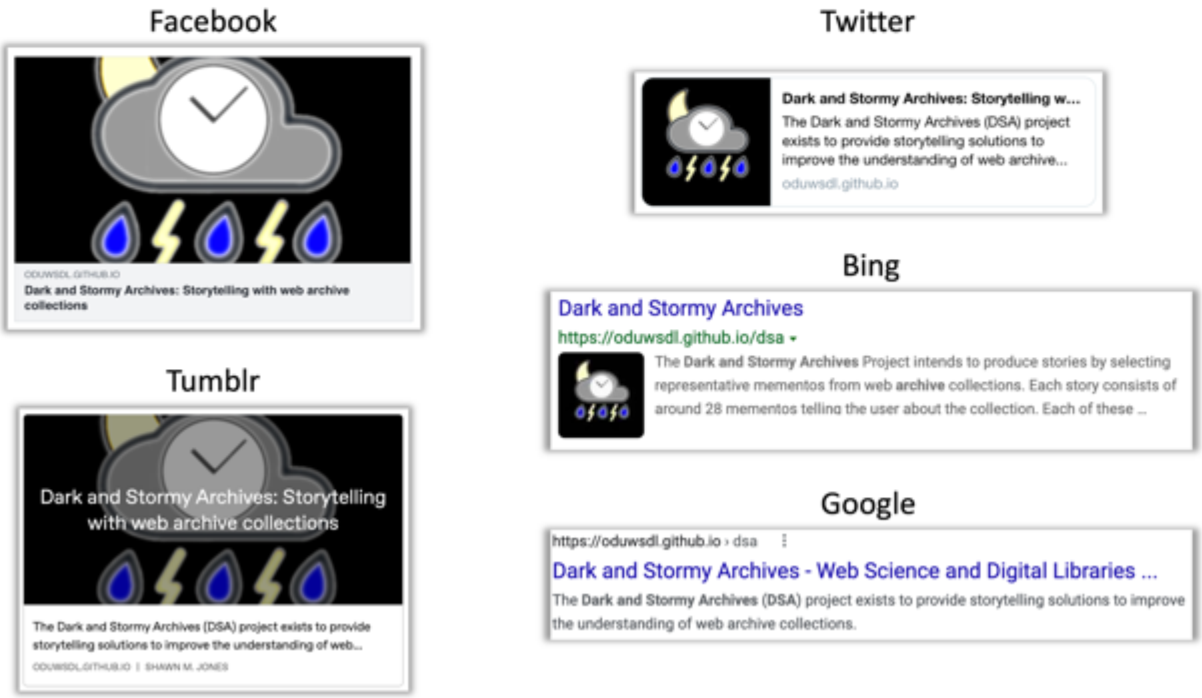
- A discussion paper on the role of the private sector in the supply of water and wastewater services
- A Dissident Liberal
- A guide to the field diagnosis of honey bee brood diseases
- A guide to the sampling and analysis of waters, wastewaters, soils and wastes
- A Human Rights Act for WA
- A Just Australia : a campaign of Australians for Just Refugee Programs

Figure 2. The PANDORA Subject *Politics* contains subcategories and collections, as well as a list of page titles.

The DSA Toolkit For Summarizing Corpora

The Dark and Stormy Archives (DSA) Project helps visitors achieve collection understanding by finding solutions that summarize web archive collections through visualizations that provide understanding at a glance. We apply social media storytelling as that visualization because most visitors are already familiar with the paradigm and thus require no additional training to understand how to consume these visualizations.

Social media storytelling consists of **surrogates** that summarize individual pages. Figure 3 shows the same page rendered as a surrogate in different web platforms. Most web users understand how to interpret the cards from search results. Many readers understand how to interpret social cards used by Facebook, Twitter, Tumblr, and other platforms. These surrogates take different shapes, but each attempts to summarize an individual document.



ODUWSDL.GITHUB.IO
Dark and Stormy Archives: Storytelling with web archive collections

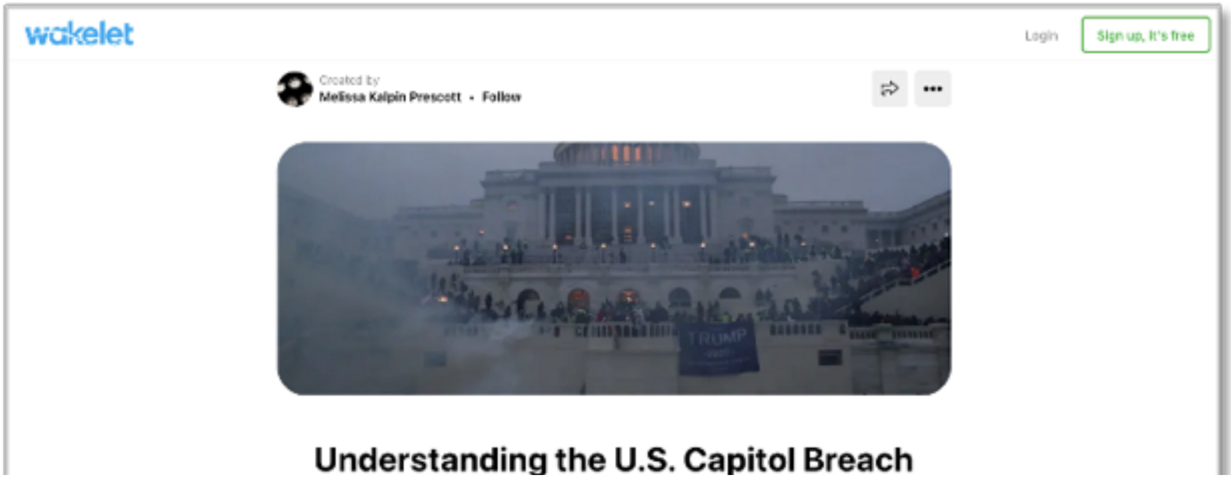
Dark and Stormy Archives: Storytelling w...
The Dark and Stormy Archives (DSA) project exists to provide storytelling solutions to improve the understanding of web archive...
oduwsdl.github.io

Dark and Stormy Archives
https://oduwsdl.github.io/dsa ▾
The Dark and Stormy Archives Project intends to produce stories by selecting representative mementos from web archive collections. Each story consists of around 28 mementos telling the user about the collection. Each of these ...

Dark and Stormy Archives: Storytelling with web archive collections
The Dark and Stormy Archives (DSA) project exists to provide storytelling solutions to improve the understanding of web...
ODUWSDL.GITHUB.IO | SHAWN M. JONES

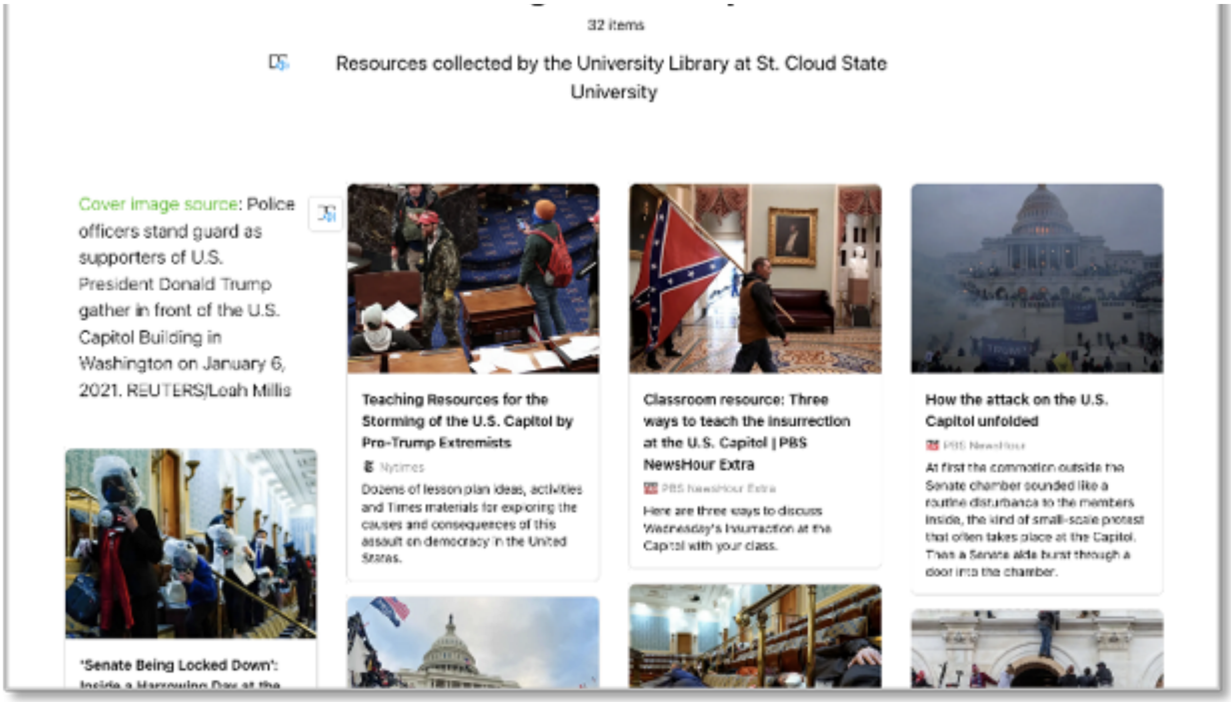
https://oduwsdl.github.io › dsa ⋮
Dark and Stormy Archives - Web Science and Digital Libraries ...
The Dark and Stormy Archives (DSA) project exists to provide storytelling solutions to improve the understanding of web archive collections.

Figure 3. Different surrogates for the same web page rendered by different platforms.



Created by
Melissa Kalpin Prescott • Follow

Understanding the U.S. Capitol Breach



32 items

Resources collected by the University Library at St. Cloud State University

Cover image source: Police officers stand guard as supporters of U.S. President Donald Trump gather in front of the U.S. Capitol Building in Washington on January 6, 2021. REUTERS/Leah Millis

Teaching Resources for the Storming of the U.S. Capitol by Pro-Trump Extremists

Nytimes

Dozens of lesson plan ideas, activities and Times materials for exploring the causes and consequences of this assault on democracy in the United States.

Classroom resource: Three ways to teach the insurrection at the U.S. Capitol | PBS NewsHour Extra

PBS NewsHour Extra

Here are three ways to discuss Wednesday's insurrection at the Capitol with your class.

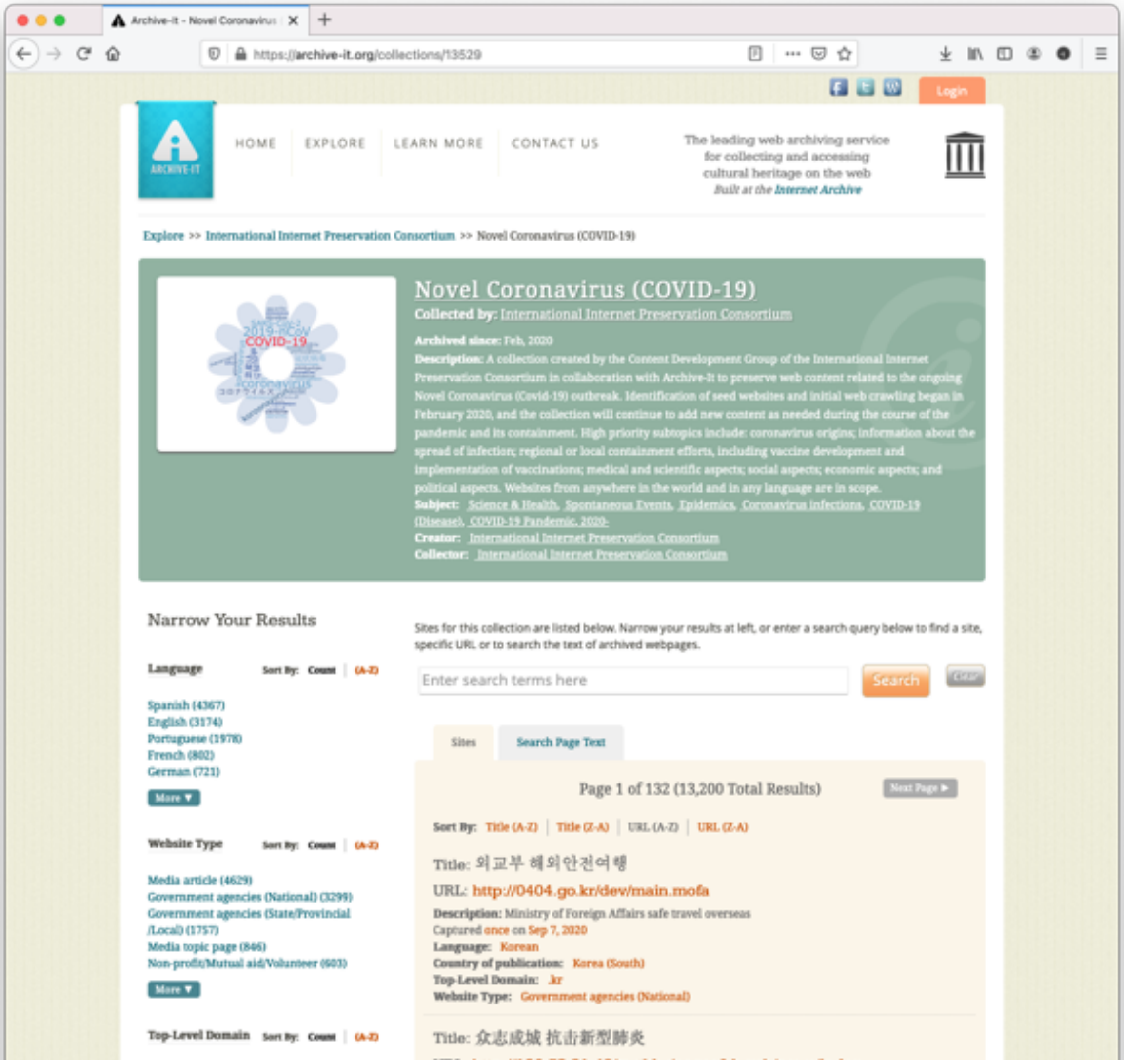
How the attack on the U.S. Capitol unfolded

PBS NewsHour

At first the commotion outside the Senate chamber sounded like a routine disturbance to the members inside, the kind of small-scale protest that often takes place at the Capitol. Then a Senate aide burst through a door into the chamber.

'Senate Being Locked Down': Inside a Harrowing Day at the

Figure 4. A social media story on Wakelet about the 2021 US Capitol Breach.



The leading web archiving service for collecting and accessing cultural heritage on the web
Built at the Internet Archive

Explore >> International Internet Preservation Consortium >> Novel Coronavirus (COVID-19)

Novel Coronavirus (COVID-19)

Collected by: International Internet Preservation Consortium

Archived since: Feb, 2020

Description: A collection created by the Content Development Group of the International Internet Preservation Consortium in collaboration with Archive-It to preserve web content related to the ongoing Novel Coronavirus (Covid-19) outbreak. Identification of seed websites and initial web crawling began in February 2020, and the collection will continue to add new content as needed during the course of the pandemic and its containment. High priority subtopics include: coronavirus origins; information about the spread of infection; regional or local containment efforts, including vaccine development and implementation of vaccinations; medical and scientific aspects; social aspects; economic aspects; and political aspects. Websites from anywhere in the world and in any language are in scope.

Subject: Science & Health, Spontaneous Events, Epidemics, Coronavirus infections, COVID-19 (Disease), COVID-19 Pandemic, 2020-

Creator: International Internet Preservation Consortium

Collector: International Internet Preservation Consortium

Narrow Your Results

Language Sort By: Count | (A-Z)

Spanish (4367)
English (3174)
Portuguese (1978)
French (802)
German (721)
More ▼

Website Type Sort By: Count | (A-Z)

Media article (4629)
Government agencies (National) (3299)
Government agencies (State/Provincial /Local) (1757)
Media topic page (846)
Non-profit/Mutual aid/Volunteer (603)
More ▼

Top-Level Domain Sort By: Count | (A-Z)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here Search Clear

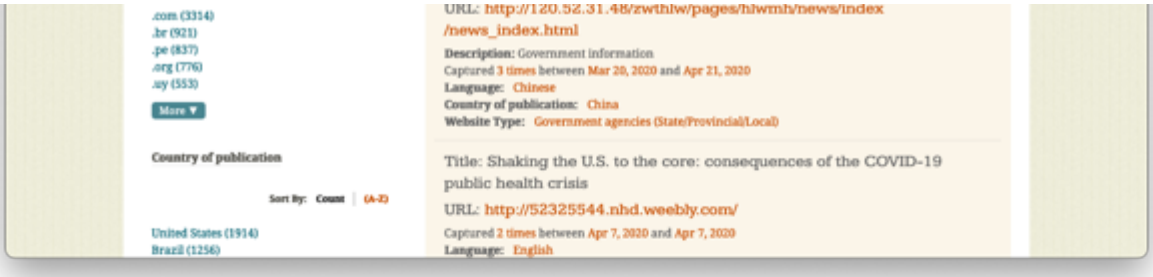
Sites | Search Page Text

Page 1 of 132 (13,200 Total Results) Next Page ▶

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: 외교부 해외안전여행
URL: http://0404.go.kr/dev/main.mofa
Description: Ministry of Foreign Affairs safe travel overseas
Captured once on Sep 7, 2020
Language: Korean
Country of publication: Korea (South)
Top-Level Domain: .kr
Website Type: Government agencies (National)

Title: 众志成城 抗击新型肺炎



.com (3314)
.br (921)
.pe (837)
.org (776)
.uy (553)

Country of publication

Sort By: Count | (A-Z)

United States (1914)
Brazil (1256)

URL: http://120.52.31.48/zwthlw/pages/hlwmh/news/index/news_index.html

Description: Government information
Captured 3 times between Mar 20, 2020 and Apr 21, 2020
Language: Chinese
Country of publication: China
Website Type: Government agencies (State/Provincial/Local)

Title: Shaking the U.S. to the core: consequences of the COVID-19 public health crisis

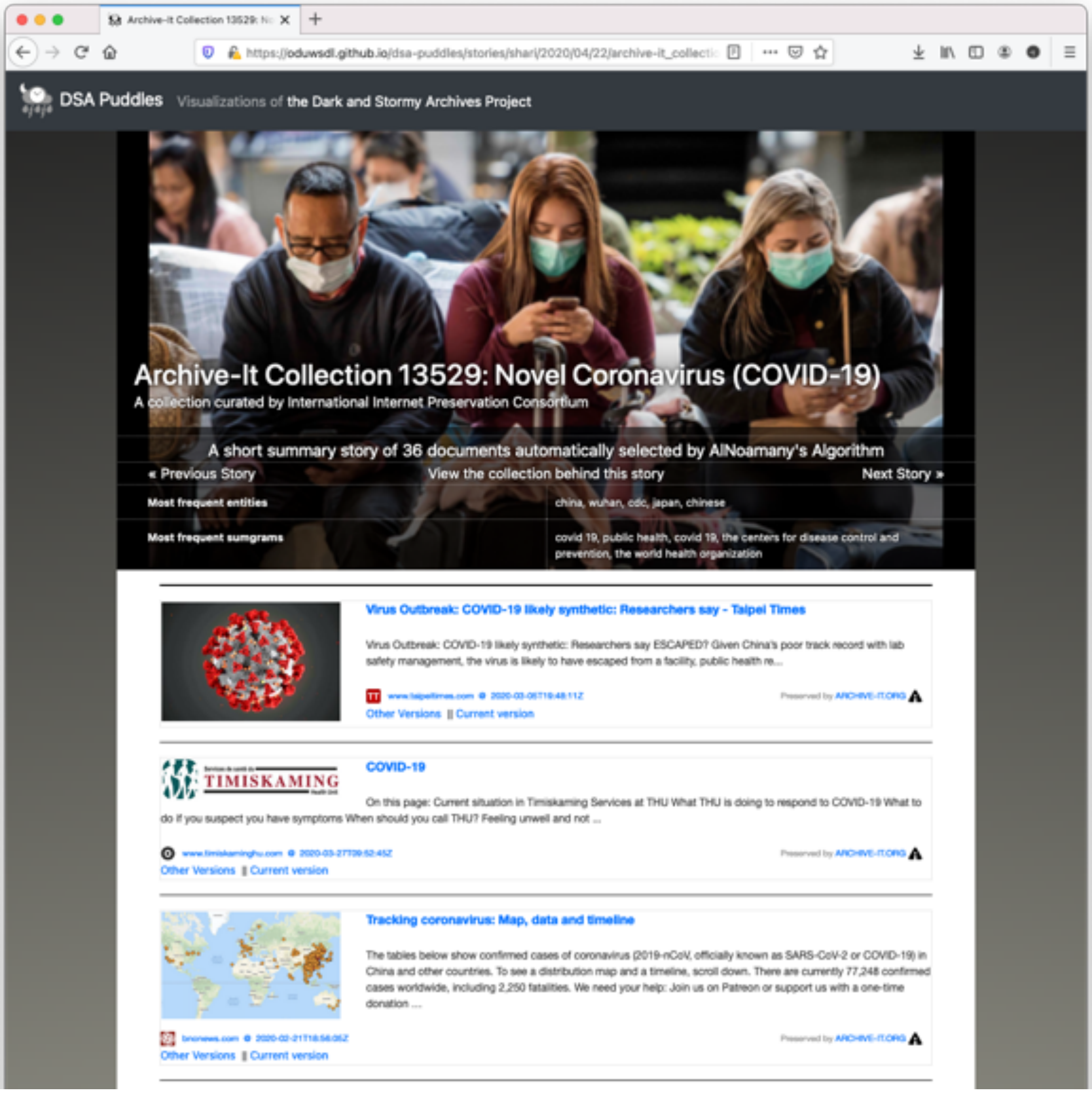
URL: http://52325544.nhd.weebly.com/

Captured 2 times between Apr 7, 2020 and Apr 7, 2020
Language: English

Figure 5. A screenshot of the IIPC's Archive-It Collection *Novel Coronavirus (COVID-19)*.

When we combine a group of surrogates together as a unit, we create a **story** that summarizes a **topic**. Figure 4 shows a story that summarizes the 2021 US Capitol Breach. This story contains surrogates, in the form of social cards, each describing a different web resource discussing the attack. Archive-It and PANDORA have their own surrogates. In Archive-It's case, the surrogate consists of the URL captured and any metadata added by the archivist. In PANDORA's case, each surrogate is a page title. We want to extend this idea of the social media story further than is already accomplished by these web archiving platforms.

Figure 5 shows a 2019 screenshot of the International Internet Preservation Consortium's (IIPC) Archive-It Collection *Novel Coronavirus (COVID-19)*. This collection contains more than 23,000 mementos, far more than a single human can review, let alone understand at a glance. We applied the DSA toolkit to produce a story summarizing this collection (Figure 6). The Archive-it collection contains metadata painstakingly provided by the archivist. We see page titles, countries of publication, languages, and, in some cases, descriptions of individual resources. There are even facets to help users explore different aspects of the collection. The story generated by the DSA Toolkit has people in masks, pictures of the virus, headlines, dates, sources, names, places, page summaries, maps showing the virus spreading across the world, and links back to the collection so visitors can explore that collection further.



Archive-It Collection 13529: N
+
https://oduwsdl.github.io/dsa-puddles/stories/shari/2020/04/22/archive-it_collecti
DSA Puddles
Visualizations of the Dark and Stormy Archives Project

Archive-It Collection 13529: Novel Coronavirus (COVID-19)

A collection curated by International Internet Preservation Consortium

A short summary story of 36 documents automatically selected by AlNoamany's Algorithm

« Previous Story   View the collection behind this story   Next Story »

Most frequent entities   china, wuhan, cdc, japan, chinese

Most frequent sumgrams   covid 19, public health, covid 19, the centers for disease control and prevention, the world health organization

Virus Outbreak: COVID-19 likely synthetic: Researchers say - Taipei Times

Virus Outbreak: COVID-19 likely synthetic: Researchers say ESCAPED? Given China's poor track record with lab safety management, the virus is likely to have escaped from a facility, public health re...

www.taipeitimes.com ● 2020-03-05T19:48:11Z   Preserved by ARCHIVE-IT.ORG
Other Versions || Current version

COVID-19

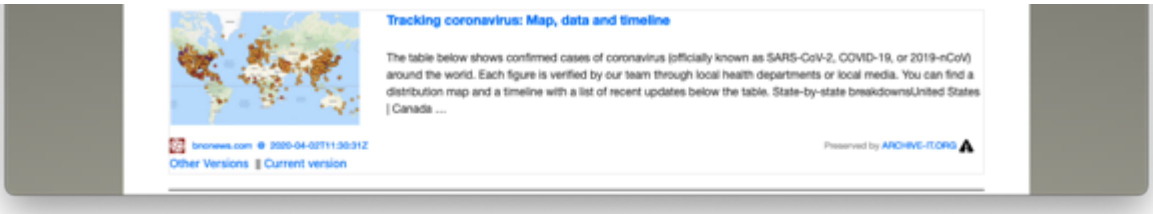
On this page: Current situation in Timiskaming Services at THU What THU is doing to respond to COVID-19 What to do if you suspect you have symptoms When should you call THU? Feeling unwell and not ...

www.timiskaminghu.com ● 2020-03-27T09:52:45Z   Preserved by ARCHIVE-IT.ORG
Other Versions || Current version

Tracking coronavirus: Map, data and timeline

The tables below show confirmed cases of coronavirus (2019-nCoV, officially known as SARS-CoV-2 or COVID-19) in China and other countries. To see a distribution map and a timeline, scroll down. There are currently 77,248 confirmed cases worldwide, including 2,250 fatalities. We need your help: Join us on Patreon or support us with a one-time donation ....

bnonews.com ● 2020-02-21T18:56:05Z   Preserved by ARCHIVE-IT.ORG
Other Versions || Current version



Tracking coronavirus: Map, data and timeline

The table below shows confirmed cases of coronavirus (officially known as SARS-CoV-2, COVID-19, or 2019-nCoV) around the world. Each figure is verified by our team through local health departments or local media. You can find a distribution map and a timeline with a list of recent updates below the table. State-by-state breakdownsUnited States | Canada ...

bnonews.com ● 2020-04-02T11:30:31Z

Other Versions || Current version

Preserved by ARCHIVE-IT.ORG

Figure 6. A story by the DSA Toolkit summarizing IIPC's Archive-It Collection *Novel Coronavirus (COVID-19)* using sampling algorithms and social media storytelling.

The DSA Toolkit is built on our model of five storytelling processes, as shown in Figure 7. A user follows these processes to tell a story that summarizes a corpus.

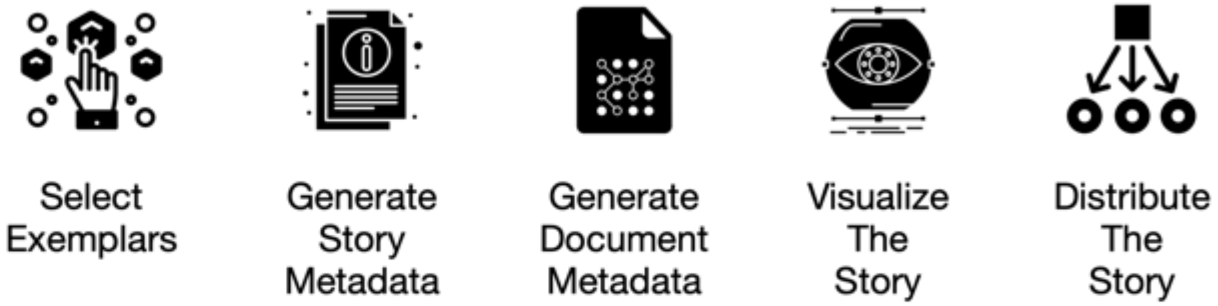


Figure 7. The five processes for telling a story that summarizes a corpus.

As a summarization, our story provides a subset of the collection to a user. This subset consists of **exemplars** – documents that represent the collection well. Thus, we need to **select exemplars** from the collection to tell our story. Selecting exemplars can be done by humans or by various sampling algorithms. We can choose exemplars that attempt to summarize the collection as a whole or select those that feature a particular aspect of the collection, such as a specific time period, person, or source.

We **generate story metadata** to enrich our story for the visitor. Story metadata can consist of the collection name, who created the collection, and other collection metadata. The story from Figure 6 shows that story metadata can also contain entities, terms, images, and other content extracted from the collection.

Once our exemplars are selected, we **generate document metadata** to summarize each exemplar. In Figure 6, each of the social cards shown in the story visualizes this document metadata. Document metadata can take many forms, as needed by the storyteller.

We can then **visualize the story** in the desired medium by applying our story and document metadata. Most users will want a story that visitors can view on the web, and hence we have focused on providing visualizations that use HTML. We can also visualize stories in other media, such as video.

Finally, we can **distribute the story** for others to consume our visualization. Distribution can be from an author's website. It could also be via social media. Alternatively, a user could potentially print the story and manually hand it to someone. The point is that visitors cannot consume a story if they cannot access it.

As shown in Figure 8, the DSA Toolkit provides tools to help meet the goals of this model. **Hypercane** helps users select exemplars and generate story metadata. **MementoEmbed** generates document metadata. **Raintale** helps users visualize the story by accepting input from Hypercane and MementoEmbed. Not shown in Figure 8 are two additional supporting packages. The library **AIU** helps programmers identify the mementos and collection metadata from a web archive collection. The **OTMT** (short for Off-Topic Memento Toolkit) identifies off-topic mementos in a collection. This paper highlights these tools and the results of their recent pilot with the NLA, thanks to a grant from the IIPC.

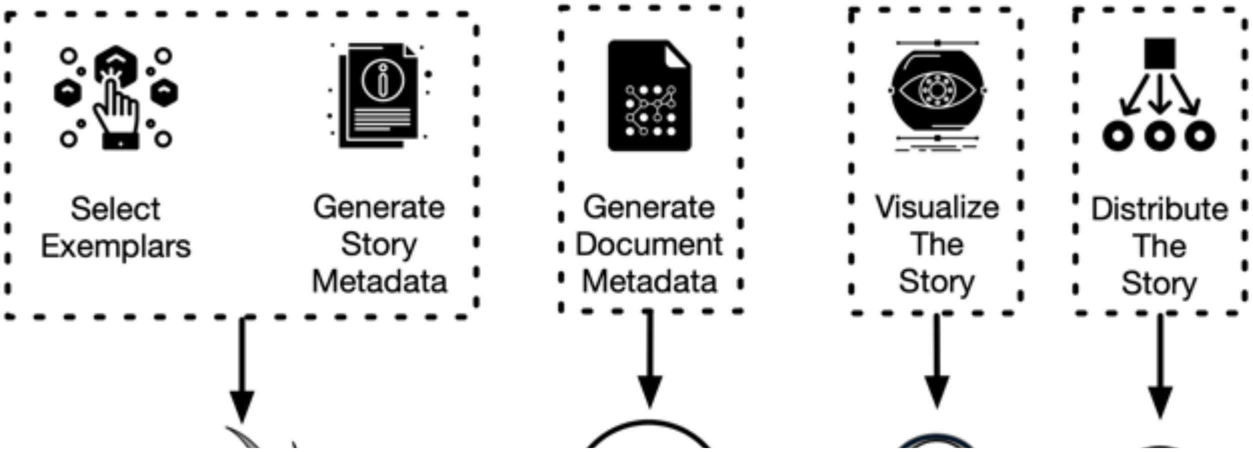


Figure 8. How the DSA Toolkit fits with the storytelling model shown in Figure 7.

Other Efforts For Understanding Web Archive Collections

The DSA Toolkit is one of many tools for helping users understand web archive collections. For users with access to the Web ARChive (WARC) files that make up a web archive collection, ArchiveSpark (by Holzmann et al.) and the Archives Unleashed Toolkit (by Ruest et al.) offer an environment for exploring collections.

A user can load WARCs into ArchiveSpark and then perform a set of operations for generating data, such as extracting titles, calculating term distributions, discovering named entities, building a graph of outgoing hyperlinks, and extracting images. ArchiveSpark typically outputs a JSON file consisting of the desired data. The output of one operation in ArchiveSpark can feed into another, allowing users to chain them together and produce the desired result. It is up to the end-user to further process this data with a third-party tool.

Archives Unleashed Toolkit (AUT) is built on Warcbase (Lin et al.). Based on work with web archive researchers, Lin et al. developed four steps for those trying to work with web archives: filter, analyze, aggregate, and visualize. AUT implements these four steps. With AUT, a user can filter based on various features, such as original resource URL, language, domain name, memento-datetime, or URL pattern. An AUT user can also generate reports on top-level domains, named entities, or link structure. Archives Unleashed Cloud (AUK) integrates a subset of this functionality with Archive-It, providing different reports for archivists to apply to their own collections. If an AUT user wishes to visualize the output, they must use a third-party tool.

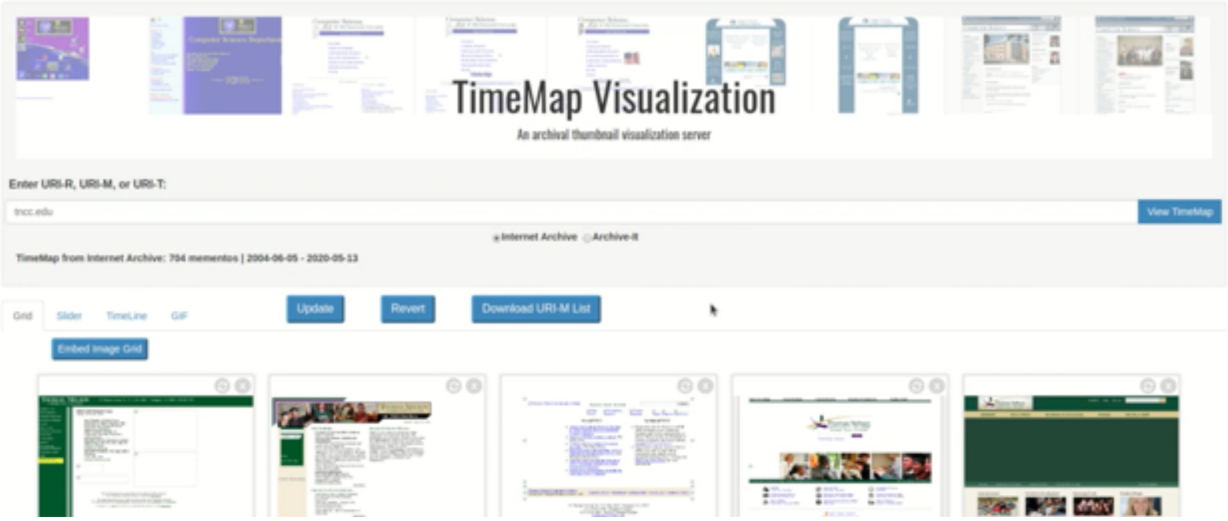
These tools inspired the development of the DSA Toolkit. Hypercane allows users to chain together several web archive collection operations, just like ArchiveSpark and AUT. Raintale accepts the output of Hypercane's operations and visualizes it.

AUT and ArchiveSpark can potentially help archivists select exemplars and generate story metadata. They can also extract or generate some document metadata. Their focus, however, is not storytelling, so they rely on third-party tools for visualization and distribution.

The input type is another difference between the DSA Toolkit and these tools. Where ArchiveSpark and AUT accept WARCs as input, the DSA Toolkit does not. Because its goal is storytelling through summarization and visualization, the DSA Toolkit expects that all resources to be shared are accessible to the visitor of the story. Also, we wanted the DSA Toolkit to be usable by those who did not have access to a collection's WARCs. Thus, the DSA Toolkit uses memento URLs instead of WARC files as its input.

Jatowt et al. may have been the first to visualize web pages over time. Their Page History Explorer downloaded mementos from web archives and generated screenshots ordered by memento-datetime as part of a timeline. Similarly, the TMVis project (by Mabe et al.) accepts a single original resource URL, applies AlSum's Algorithm to find the most novel mementos of that page, allows the user to reduce that set further manually, and then renders those mementos as a set of screenshots. These efforts work with a single original resource URL, not a collection of many different ones. However, they satisfy parts of our storytelling model by selecting exemplars, generating story metadata, generating document metadata (in the form of a screenshot), and visualizing the story – with Jatowt using a timeline and TMVis providing several different visual arrangements. With TMVis, users can distribute parts of their stories in various forms such as embeds, animated GIFs, and image sliders.

Padia et al. generated a set of visualizations for web archive collections. Padia's visualizations did not focus on a single original resource URL but instead tried to consider different features of the mementos in the collection. They produced visualizations consisting of heat maps, timelines, word clouds, bubble charts, and treemaps. When possible, they leveraged the metadata provided by the collection, allowing them to cluster individual mementos as part of these visualizations. While they successfully created a set of visualizations to convey meaning about aspects of the collection, these visualizations required training for the viewer to understand them. Thus, they did not satisfy our goal of conveying meaning at a glance.



Enter URI-R, URI-M, or URI-T: tncc.edu

TimeMap from Internet Archive: 704 mementos | 2004-06-05 - 2020-05-13

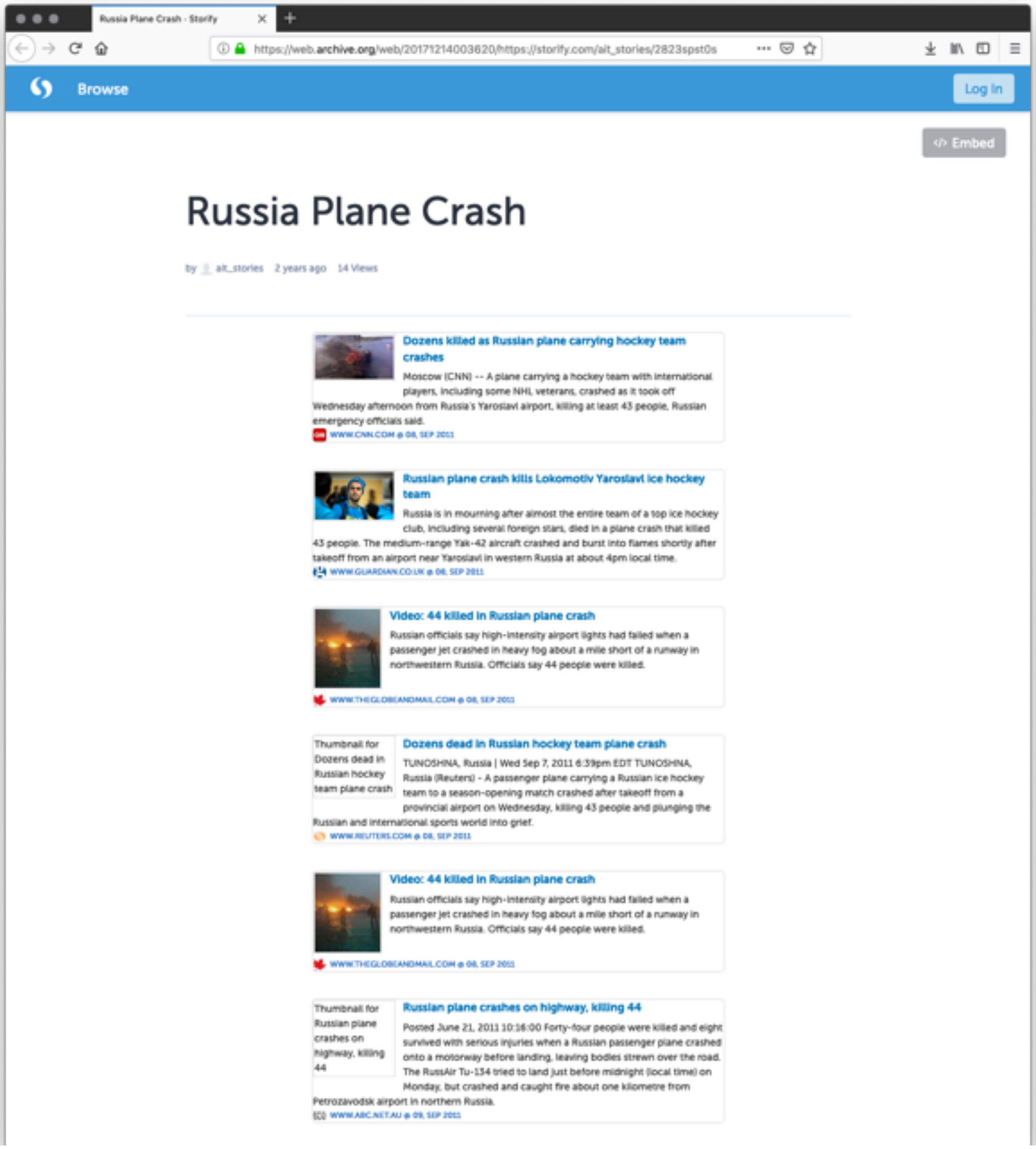
Datetime: 2004-06-05, 04:39:49
Datetime: 2004-08-24, 12:54:33
Datetime: 2006-06-16, 03:19:46
Datetime: 2008-10-11, 02:00:11
Datetime: 2011-04-17, 02:23:10
THOMAS NELSON
THOMAS NELSON MyTNCC
THOMAS NELSON
THOMAS NELSON
THOMAS NELSON 50

Figure 9. With TMVis, users can explore a single original resource over time and visualize thumbnails for that resource.

AlNoamany et al. pioneered combining storytelling with web archive collections. At the time of her work, the most popular social media storytelling service was the now-defunct Storify. She analyzed Storify stories and compared them with Archive-It collections to understand their similarities and differences. She identified that popular stories contain a median of 28 links, giving automated storytelling platforms a target number of items to select from a collection.

She identified some of the operations necessary to select exemplars automatically. Her algorithm filtered off-topic mementos, then duplicate mementos, and then non-English content from the collection. She also applied clustering, both by memento-datetime and by content. Finally, she scored the resulting mementos by Padia's web page category, McCown's URL path depth, and Brunelle's memento damage. Her algorithm, which the DSA Toolkit implements as *DSA1*, is the first of many possible methods for selecting exemplars. She demonstrated (*preprint version*) that participants *could* determine the difference between randomly generated stories and those created by human archivists through a user study. Equally important, she confirmed that participants *could not* distinguish between stories produced by her visualization algorithm and those created by human archivists.

Figure 10 shows a story generated by AlNoamany's proof-of-concept with Storify. Her story metadata was the collection name. Her proof-of-concept generated document metadata to fit into this platform. Storify handled distributing the story to visitors. Her research project was named *Dark and Stormy Archives*, and the current DSA project is a continuation of her work.


Russia Plane Crash - Storify
https://web.archive.org/web/20171214003620/https://storify.com/ait_stories/2823spst0s
Browse
Log In
</> Embed

Russia Plane Crash

by ait_stories 2 years ago 14 Views

Dozens killed as Russian plane carrying hockey team crashes

Moscow (CNN) -- A plane carrying a hockey team with international players, including some NHL veterans, crashed as it took off Wednesday afternoon from Russia's Yaroslavl airport, killing at least 43 people, Russian emergency officials said.

WWW.CNN.COM @ 08, SEP 2011

Russian plane crash kills Lokomotiv Yaroslavl ice hockey team

Russia is in mourning after almost the entire team of a top ice hockey club, including several foreign stars, died in a plane crash that killed 43 people. The medium-range Yak-42 aircraft crashed and burst into flames shortly after takeoff from an airport near Yaroslavl in western Russia at about 4pm local time.

WWW.GUARDIAN.CO.UK @ 08, SEP 2011

Video: 44 killed in Russian plane crash

Russian officials say high-intensity airport lights had failed when a passenger jet crashed in heavy fog about a mile short of a runway in northwestern Russia. Officials say 44 people were killed.

WWW.THEGLOBEANDMAIL.COM @ 08, SEP 2011

Thumbnail for Dozens dead in Russian hockey team plane crash

Dozens dead in Russian hockey team plane crash

TUNOSHNA, Russia | Wed Sep 7, 2011 6:39pm EDT TUNOSHNA, Russia (Reuters) - A passenger plane carrying a Russian ice hockey team to a season-opening match crashed after takeoff from a provincial airport on Wednesday, killing 43 people and plunging the Russian and international sports world into grief.

WWW.REUTERS.COM @ 08, SEP 2011

Video: 44 killed in Russian plane crash

Russian officials say high-intensity airport lights had failed when a passenger jet crashed in heavy fog about a mile short of a runway in northwestern Russia. Officials say 44 people were killed.

WWW.THEGLOBEANDMAIL.COM @ 08, SEP 2011

Thumbnail for Russian plane crashes on highway, killing 44

Russian plane crashes on highway, killing 44

Posted June 21, 2011 10:16:00 Forty-four people were killed and eight survived with serious injuries when a Russian passenger plane crashed onto a motorway before landing, leaving bodies strewn over the road. The RussAir Tu-134 tried to land just before midnight (local time) on Monday, but crashed and caught fire about one kilometre from Petrozavodsk airport in northern Russia.

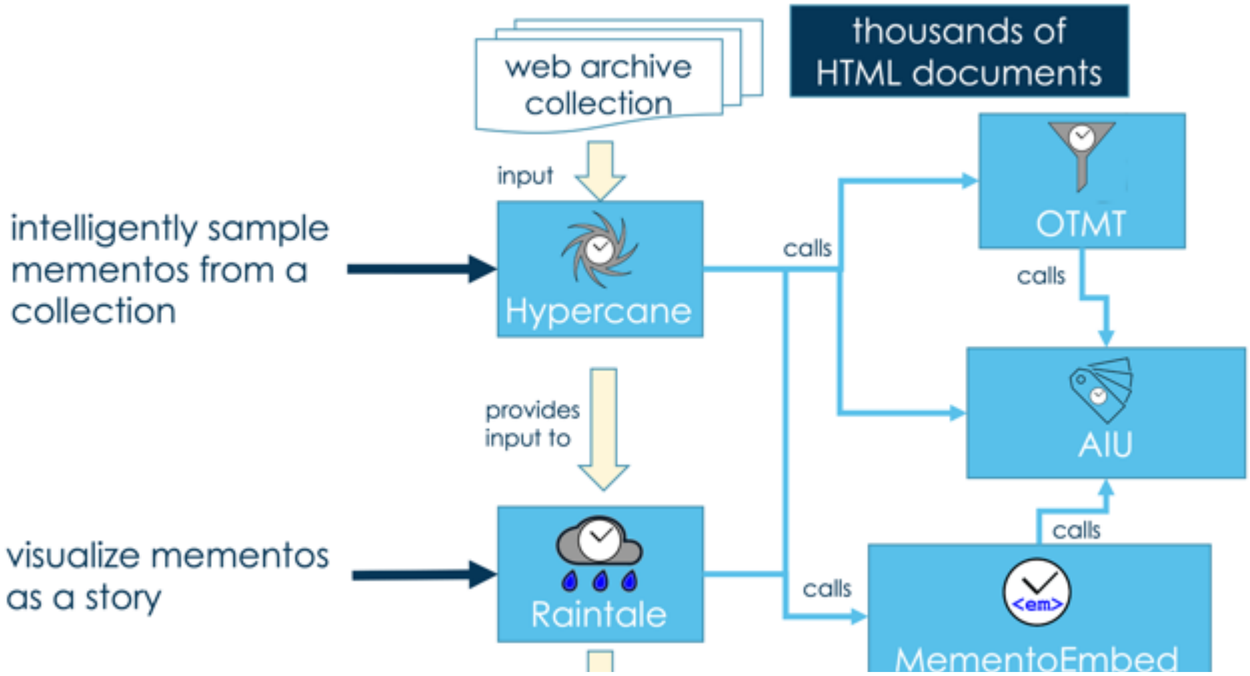
WWW.ABC.NET.AU @ 09, SEP 2011

Figure 10. A story produced by AlNoamany's Dark and Stormy Archives proof-of-concept from the Archive-It collection *Russia Plane Crash*.

Table 1. Different web archive collection analysis efforts and suitability for summarizing collections with storytelling.

		ArchiveSpark	**Archives Unleashed Toolkit**	**Page History Explorer**	**TMVis**	**Padia's Visualizations**	**AlNoamany's Dark and Stormy Archives**	**DSA Toolkit (Current)**
Storytelling Model Support	**Select Exemplars**	yes	yes	yes	yes	yes	yes	yes
	Generate Story Metadata	yes	yes	yes	yes	yes	yes	yes
	Generate Document Metadata	yes	yes	yes	yes	yes	yes	yes
	Visualize the Story	no	no	yes	yes	yes	yes	yes
	Distribute the Story	no	no	no	yes	not currently public, but could be	yes	yes
Can User Customize Part of Storytelling Model?	**Select Exemplars**	yes	yes	no	yes	no	yes	yes
	Generate Story Metadata	yes	yes	no	no	no	no	yes
	Generate Document Metadata	yes	yes	no	no	no	no	yes
	Visualize the Story	N/A	N/A	no	yes	no	no	yes
	Distribute the Story	N/A	N/A	no	no	no	no	yes
Supports More Than One Original Resource		yes	yes	no	no	yes	yes	yes
Allows Analysis Without WARCs		no	no	yes	yes	yes	yes	yes
Web Archive Platforms Supported		any WARC files	any WARC files	1	3	1	1	any Memento-compliant archive

The DSA Toolkit



output
story
~28 representative mementos visualized as surrogates

Figure 11. The DSA Toolkit workflow for producing a story relies heavily on Hypercane and Raintale.

The DSA Toolkit consists of five different software projects that satisfy our storytelling processes. Figure 11 provides an overview of the workflow for these tools. A user executes Hypercane to select exemplars and generate story metadata. Then they feed that output into Raintale to visualize their story. Both Hypercane and Raintale consult the other components as needed. Here we introduce each tool and highlight our recent improvements thanks to the pilot with NLA.

AIU

AIU (formerly Archive-It Utilities) is a Python library containing several classes that provide information about web archive collections. A programmer can invoke one of these classes and supply a collection identifier. The class includes several methods providing information like collection title, the URLs of the original resources in the collection, and other metadata, if present. AIU applies APIs and scraping HTML to gather its information.

AIU initially provided this information only for Archive-It collections. Archive-It collections contain metadata and a list of preserved original resources. From these original resources, we find TimeMaps. From these TimeMaps, we discover mementos. AIU follows this path to find the mementos in an Archive-It collection.

Below is an example iPython session where a user can extract information about an Archive-It collection (we measure the length of the list returned by the `list_seed_uris()` method for brevity, but it does produce a list of original resource URLs for the collection):

```
In [1]: from aiu import ArchiveItCollection

In [2]: aic = ArchiveItCollection(5728)

In [3]: aic.get_collection_name()
Out[3]: 'Social Media'

In [4]: aic.get_collectedby()
Out[4]: 'Willamette University'

In [5]: aic.get_description()
Out[5]: 'Social media content created by Willamette University.'

In [6]: aic.get_collection_uri()
Out[6]: 'https://archive-it.org/collections/5728'

In [7]: len(aic.list_seed_uris())
Out[7]: 113
```

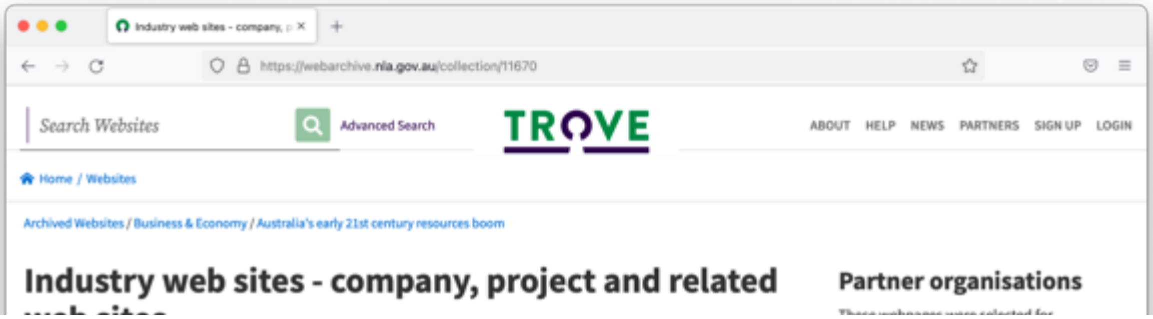
As part of our IIPC grant, we extended AIU to service the three types of collections at the NLA. as presented through their PANDORA and Trove websites. The PANDORA Archive is the curated selective web archive component of the larger Australian Web Archive (AWA). The NLA's entire web archive corpus (including PANDORA) is accessible through the Trove discovery service. While the AWA includes PANDORA's thematic sub-collections, the more extensive PANDORA subject listings are only viewable through the stand-alone PANDORA Archive website.

PANDORA subjects contain other PANDORA subjects listed as "subcategories." Each PANDORA subject includes a list of page titles, as seen in Figure 2. Each page title represents an original resource and links to a Title Entry Page (TEP). The TEP serves as a TimeMap that links to all mementos for that original resource. AIU will follow this chain to find similar metadata to Archive-It, including original resource and memento URLs.

PANDORA subjects also contain PANDORA collections. Like subjects, each PANDORA collection can include collections and page titles. Each page title links to a TEP which functions the same as with PANDORA subjects. AIU will follow these links to find metadata and URLs for PANDORA collections.

Trove collections are slightly different. Each Trove collection fans out into subcollections. Subcollections contain direct links to mementos, as shown in Figure 12. Now AIU includes a class that will follow these links, gather metadata, and capture URLs as needed.

Please view AIU's documentation for more information on applying it to gather collection information.

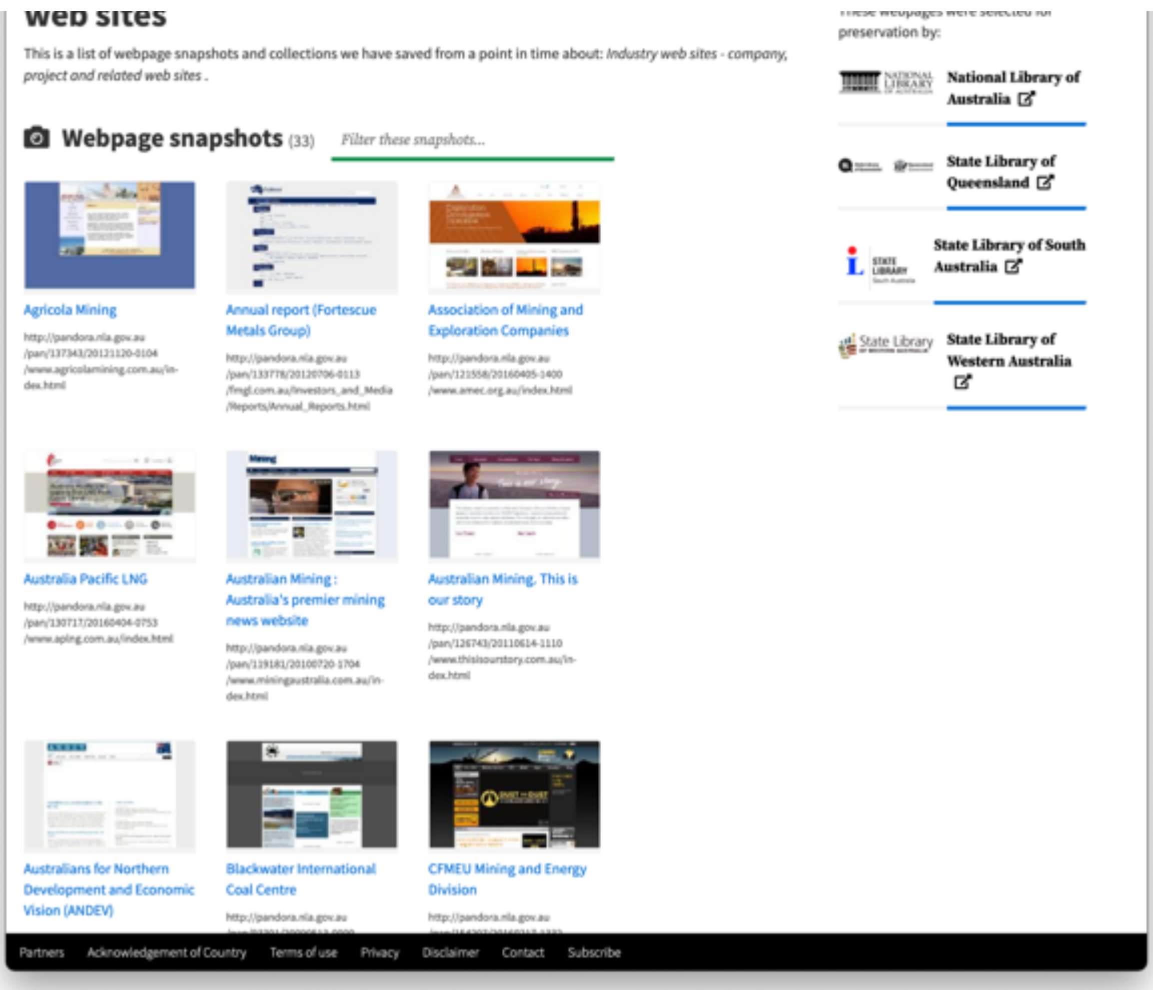


https://webarchive.nla.gov.au/collection/11670

Archived Websites / Business & Economy / Australia's early 21st century resources boom

Industry web sites - company, project and related

Partner organisations



web sites

This is a list of webpage snapshots and collections we have saved from a point in time about: *Industry web sites - company, project and related web sites .*

Webpage snapshots (33) *Filter these snapshots...*

Agricola Mining
http://pandora.nla.gov.au/pan/137343/20121120-0104/www.agricolamining.com.au/index.html

Annual report (Fortescue Metals Group)
http://pandora.nla.gov.au/pan/133778/20120706-0113/fmgl.com.au/Investors_and_Media/Reports/Annual_Reports.html

Association of Mining and Exploration Companies
http://pandora.nla.gov.au/pan/121558/20160405-1400/www.amec.org.au/index.html

Australia Pacific LNG
http://pandora.nla.gov.au/pan/130717/20160404-0753/www.aplng.com.au/index.html

Australian Mining : Australia's premier mining news website
http://pandora.nla.gov.au/pan/119181/20100720-1704/www.miningaustralia.com.au/index.html

Australian Mining, This is our story
http://pandora.nla.gov.au/pan/126743/20110614-1110/www.thisisourstory.com.au/index.html

Australians for Northern Development and Economic Vision (ANDEV)

Blackwater International Coal Centre
http://pandora.nla.gov.au

CFMEU Mining and Energy Division
http://pandora.nla.gov.au

These webpages were selected for preservation by:

National Library of Australia

State Library of Queensland

State Library of South Australia

State Library of Western Australia

Figure 12. A screenshot of a Trove collection.

OTMT

The Off-Topic Memento Toolkit (OTMT) is an application and library that supports different text similarity measures for identifying which mementos in a collection are on-topic or off-topic. It is used as a library by Hypercane to perform this same task. The OTMT received bug fixes as part of our IIPC grant work.

Hypercane

Hypercane is a complex application that provides the user with many ways to select exemplars and generate story metadata automatically. Hypercane exists in the command line application `hc`. A user selects exemplars with the `sample` action of that application. For example, to select exemplars from Archive-It collection 694 using the DSA1 algorithm, a user types:

```
# hc sample dsa1 -i archiveit -a 694 -o story-mementos.tsv
```

Generating story metadata is handled by Hypercane's `report` action. For example, to generate a list of named entities from a file containing a list of mementos, a user types:

```
# hc report entities -i mementos -a story-mementos.tsv \
   -o entity-report.tsv
```

Finally, to create a rich story, a user can `synthesize` the output of other Hypercane commands into other formats. To create a rich JSON story file for Raintale, a user types:

```
# hc synthesize raintale-story -i mementos -a story-mementos.tsv \
    --imagedata ${image_report} \
    --title "My Story for Collection X" \
    --termdata terms-report.tsv \
    --entitydata entity-report.tsv \
    -o mystory.json
```

The `sample`, `report`, and `synthesize` actions provide Hypercane's top-level functionality. Their capabilities are not limited to what we show in the examples above. The `sample` action currently supports fourteen different sampling algorithms, denoted by their name – e.g., a user can execute stratified random sampling with `hc sample stratified-random`, or they can run the DSA3 algorithm from our previous work with `hc sample dsa3`. Likewise, the `report` action supports ten different types of reports that can produce helpful metadata for storytelling, like extracted collection metadata, image analysis, or the list of top phrases for the input. The `synthesize` action gives the user different output format options, like a Raintale story file or WARCs.

As seen above, each Hypercane command supports the arguments `-i` (for input type) and `-a` (for input argument). The input type instructs Hypercane on the nature of the input argument. As seen above, if the input type is `archiveit`, then the input argument must be an Archive-It collection identifier. If the input type is `mementos`, then the input argument must be a file containing a list of memento URLs.

Hypercane currently supports the following input types:

- `mementos` – for a file containing a list of memento URLs
- `timemaps` – the input argument is a file containing a list of TimeMap URLs
- `original-resources` – the input argument is a file containing a list of original resource URLs
- `archiveit` – the input argument is an Archive-It collection identifier
- `pandora-subject` – the input argument is a PANDORA subject identifier, added as part of the IIPC grant work
- `pandora-collection` – the input argument is PANDORA collection identifier, added as part of the IIPC grant work
- `trove` – the input argument is a Trove collection identifier, added as part of the IIPC grant work

Hypercane supports all Memento-compliant archives. Thus, a user need not rely upon mementos from Archive-It or NLA. The `mementos` and `timemaps` input types allow users to submit URLs from other archives, such as the Internet Archive's Wayback Machine, Archive.Today, or Arquivo.pt.

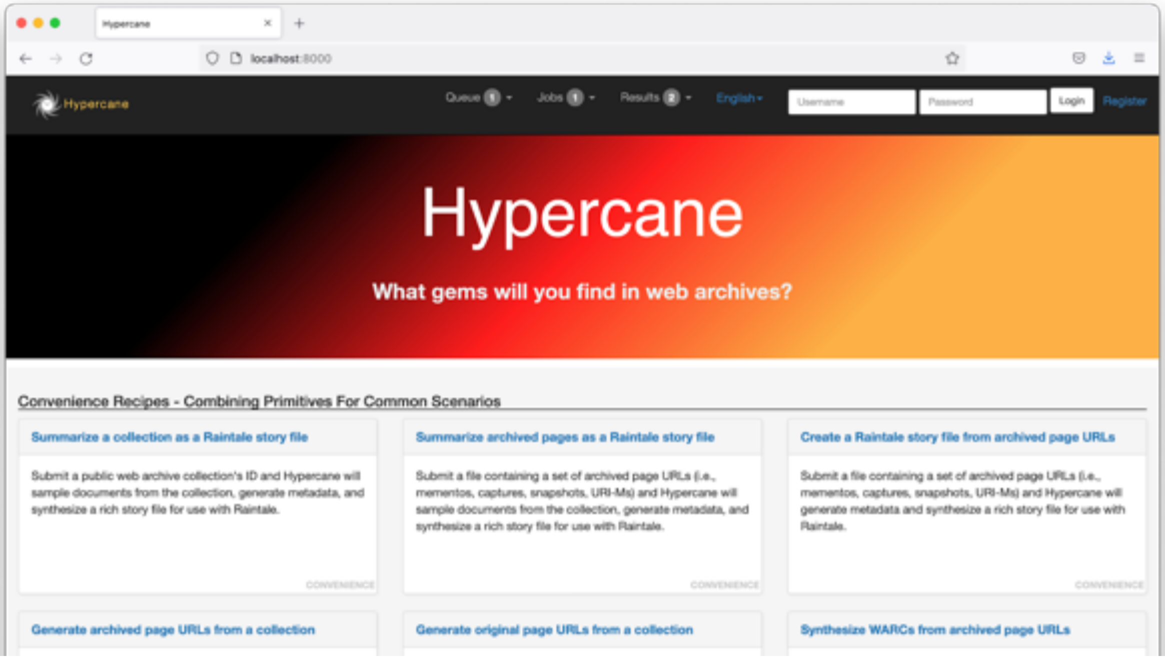
Every Hypercane command also supports the `-o` argument to specify an output file. In most cases, the output file is a list of memento URLs. This feature allows the user to chain Hypercane commands together so that the output of one command can feed into another.

Hypercane's sample action helps users execute existing sampling algorithms, but users also have access to the following algorithmic primitives that allow them to build their samples:

- `identify` – identifies the memento, TimeMap, or original-resource URLs for a given input type
- `filter` – produces a list of mementos that match the provided criteria (e.g., are on-topic, have unique content, are written in a specific language)
- `cluster` – clusters the collection via a given algorithm, such as through LDA topic modeling or K-means clustering by memento-datetime
- `score` – scores each memento in the input by a given function
- `order` – sorts the input by a given feature, like memento-datetime

By combining these primitives, users can create powerful scripts for sampling from collections, allowing them to tell many types of stories. We will not describe all of these primitives in detail for brevity but instead, direct interested readers to Hypercane's official documentation and our recent publications.

As part of the IIPC grant, we also sought to make Hypercane more approachable to administrators and end-users. Hypercane is a Python application and also depends on MongoDB for caching. We are evaluating native Linux installers for CentOS 8 and Ubuntu 21.10 that handle installing dependencies and provide convenience scripts for administering the application.



Hypercane

What gems will you find in web archives?

Convenience Recipes - Combining Primitives For Common Scenarios

Summarize a collection as a Raintale story file

Submit a public web archive collection's ID and Hypercane will sample documents from the collection, generate metadata, and synthesize a rich story file for use with Raintale.

CONVENIENCE

Summarize archived pages as a Raintale story file

Submit a file containing a set of archived page URLs (i.e., mementos, captures, snapshots, URI-Ms) and Hypercane will sample documents from the collection, generate metadata, and synthesize a rich story file for use with Raintale.

CONVENIENCE

Create a Raintale story file from archived page URLs

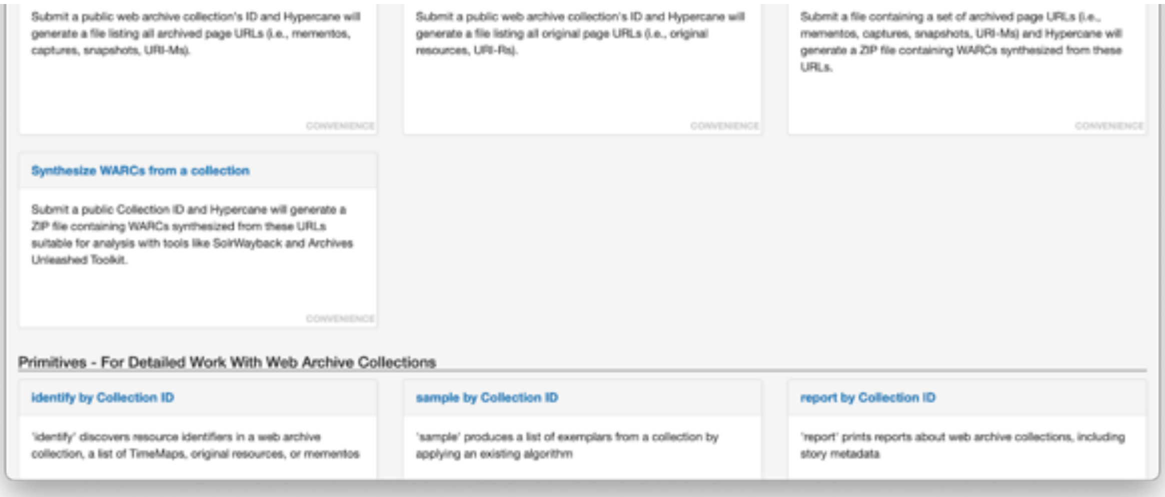
Submit a file containing a set of archived page URLs (i.e., mementos, captures, snapshots, URI-Ms) and Hypercane will generate metadata and synthesize a rich story file for use with Raintale.

CONVENIENCE

Generate archived page URLs from a collection

Generate original page URLs from a collection

Synthesize WARCs from archived page URLs



Submit a public web archive collection's ID and Hypercane will generate a file listing all archived page URLs (i.e., mementos, captures, snapshots, URI-Ms).

CONVENIENCE

Submit a public web archive collection's ID and Hypercane will generate a file listing all original page URLs (i.e., original resources, URI-Rs).

CONVENIENCE

Submit a file containing a set of archived page URLs (i.e., mementos, captures, snapshots, URI-Ms) and Hypercane will generate a ZIP file containing WARCs synthesized from these URLs.

CONVENIENCE

Synthesize WARCs from a collection

Submit a public Collection ID and Hypercane will generate a ZIP file containing WARCs synthesized from these URLs suitable for analysis with tools like SolrWayback and Archives Unleashed Toolkit.

CONVENIENCE

Primitives - For Detailed Work With Web Archive Collections

identify by Collection ID

'identify' discovers resource identifiers in a web archive collection, a list of TimeMaps, original resources, or mementos

sample by Collection ID

'sample' produces a list of exemplars from a collection by applying an existing algorithm

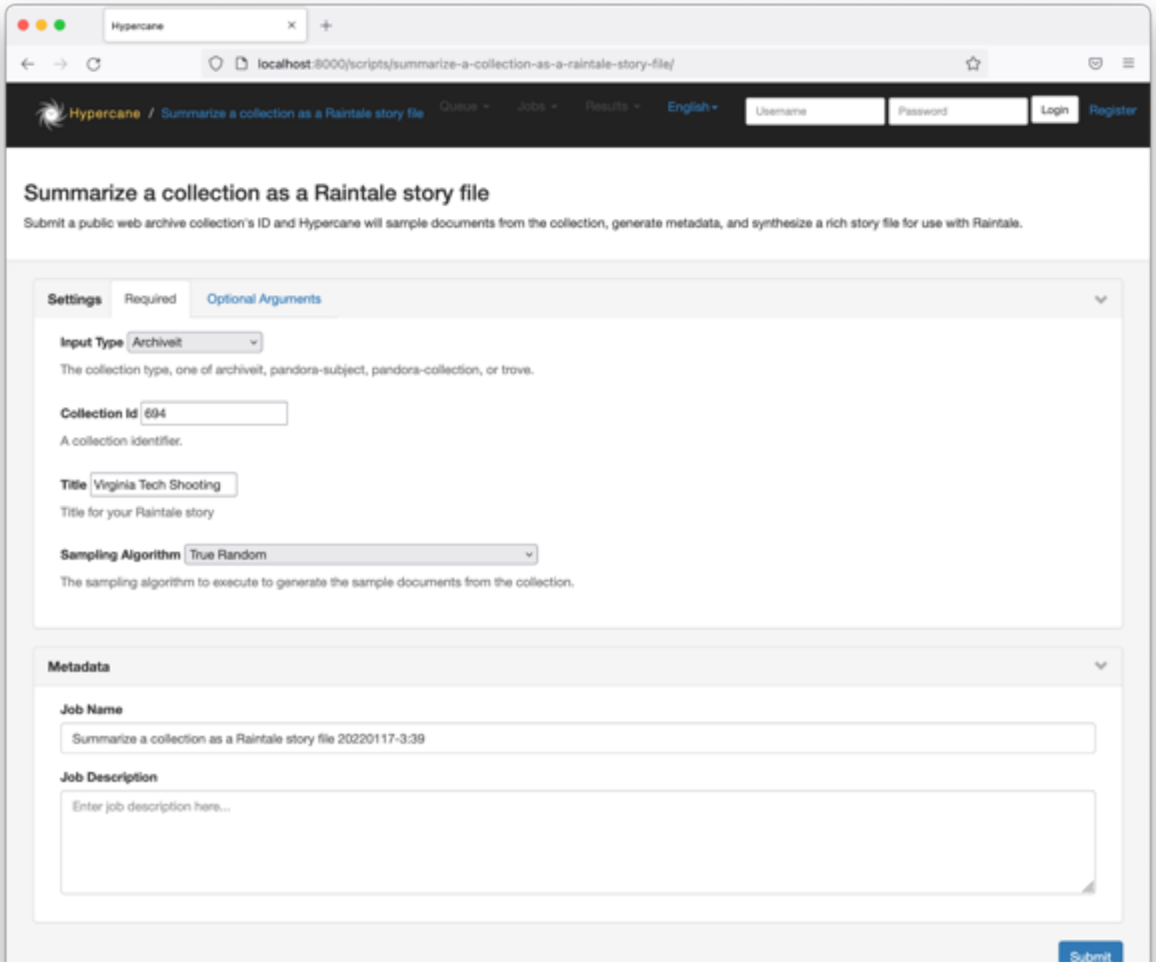
report by Collection ID

'report' prints reports about web archive collections, including story metadata

Figure 13. A screenshot of Hypercane's Web User Interface landing page.

We developed a new Web User Interface (WUI) so that Hypercane was approachable for users not familiar with scripting or working with command-line interfaces. Figure 13 shows a screenshot of this WUI. We wanted to create a web user interface while still preserving the existence and capabilities of the command-line application for scripting. We built Hypercane's WUI with the Wooey library, which creates web user interfaces from command-line applications, satisfying our needs. With Wooey, a user can fill out a web form to submit a Hypercane command as a job and come back later to retrieve the results. Wooey also provides authentication and separation of jobs to protect users' privacy as a Django application.

Hypercane's `sample` action allows users to run pre-existing sampling algorithms to get a list of mementos for their story. As noted above, it supports ten sampling algorithms, including AlNoamany's. But what if a user wants to run `sample`, execute all `report` actions necessary to generate different types of story metadata, and finally `synthesize` a JSON file for use in a rich Raintale story? Before the WUI, we suggested that they write a script that executes these commands in order. The WUI presents **convenience recipes** that help users run complex tasks with minimal input and no scripting. Hypercane's WUI currently supports seven convenience recipes as web forms.



Hypercane
localhost:8000/scripts/summarize-a-collection-as-a-raintale-story-file/
Hypercane / Summarize a collection as a Raintale story file
Queue
Jobs
Results
English
Username
Password
Login
Register

Summarize a collection as a Raintale story file

Submit a public web archive collection's ID and Hypercane will sample documents from the collection, generate metadata, and synthesize a rich story file for use with Raintale.

Settings | Required | Optional Arguments

Input Type: Archiveit

The collection type, one of archiveit, pandora-subject, pandora-collection, or trove.

Collection Id: 694

A collection identifier.

Title: Virginia Tech Shooting

Title for your Raintale story

Sampling Algorithm: True Random

The sampling algorithm to execute to generate the sample documents from the collection.

Metadata

Job Name

Summarize a collection as a Raintale story file 20220117-3:39

Job Description

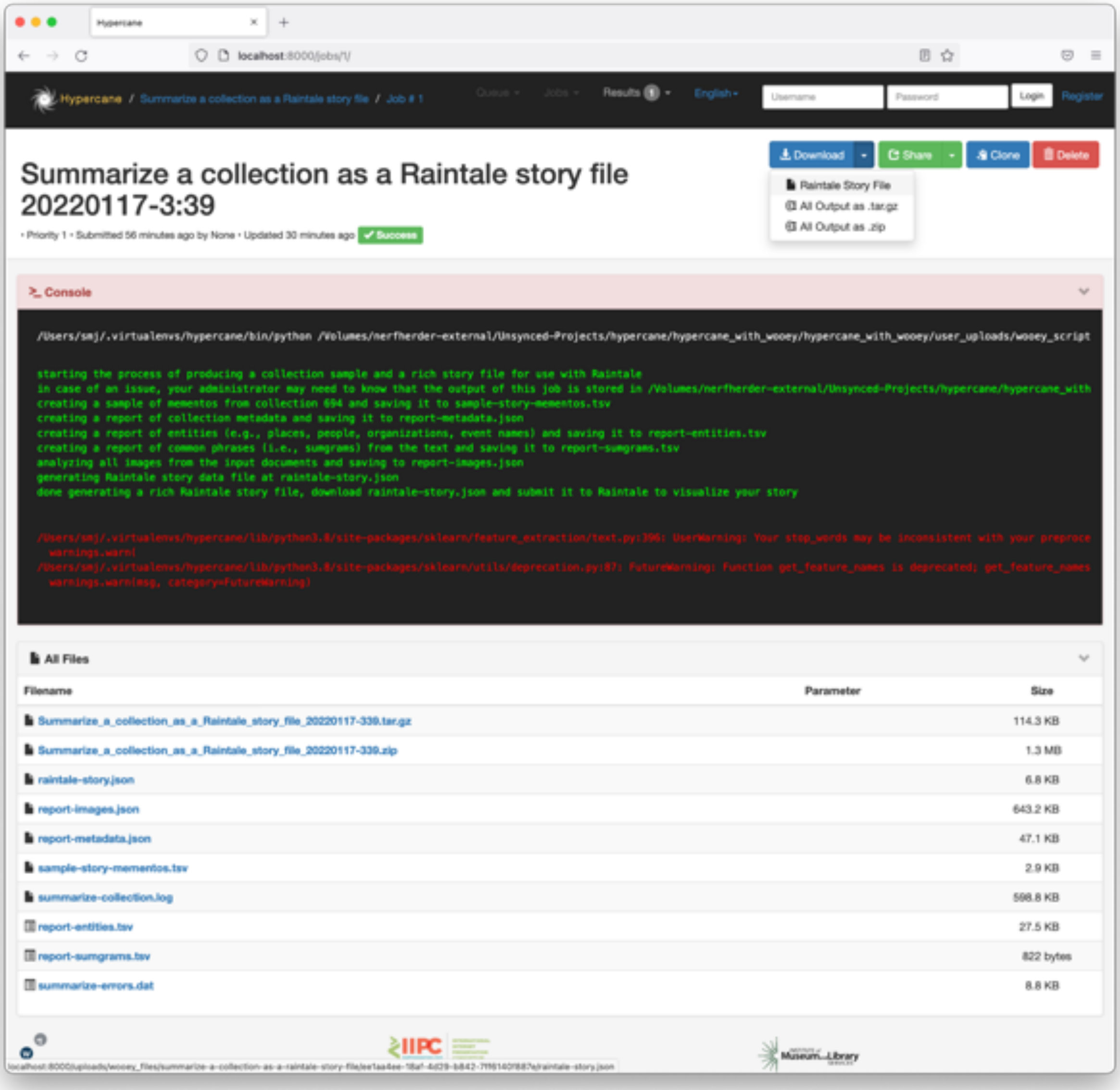
Enter job description here...

Submit

This work was funded in part by a grant from the International Internet Preservation Consortium and grant LG-71-15-0077-15 from the Institute of Museum and Library Services.

This software is licensed under a MIT License by the Old Dominion University Web Science and Digital Libraries Research Group 2021.

Figure 14. A screenshot of the Hypercane web form for the recipe "Summarize a collection as a Raintale story file."



Summarize a collection as a Raintale story file 20220117-3:39

• Priority 1 • Submitted 56 minutes ago by None • Updated 30 minutes ago ✓ Success

Console

```
/Users/smj/.virtualenvs/hypercane/bin/python /Volumes/nerfherder-external/Unsynced-Projects/hypercane/hypercane_with_wooey/hypercane_with_wooey/user_uploads/wooey_script


starting the process of producing a collection sample and a rich story file for use with Raintale
in case of an issue, your administrator may need to know that the output of this job is stored in /Volumes/nerfherder-external/Unsynced-Projects/hypercane/hypercane_with
creating a sample of mementos from collection 694 and saving it to sample-story-mementos.tsv
creating a report of collection metadata and saving it to report-metadata.json
creating a report of entities (e.g., places, people, organizations, event names) and saving it to report-entities.tsv
creating a report of common phrases (i.e., sumgrams) from the text and saving it to report-sumgrams.tsv
analyzing all images from the input documents and saving to report-images.json
generating Raintale story data file at raintale-story.json
done generating a rich Raintale story file, download raintale-story.json and submit it to Raintale to visualize your story


/Users/smj/.virtualenvs/hypercane/lib/python3.8/site-packages/sklearn/feature_extraction/text.py:396: UserWarning: Your stop_words may be inconsistent with your preproce
  warnings.warn(
/Users/smj/.virtualenvs/hypercane/lib/python3.8/site-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function get_feature_names is deprecated; get_feature_names
  warnings.warn(msg, category=FutureWarning)
```

All Files

Filename	Parameter	Size
Summarize_a_collection_as_a_Raintale_story_file_20220117-339.tar.gz		114.3 KB
Summarize_a_collection_as_a_Raintale_story_file_20220117-339.zip		1.3 MB
raintale-story.json		6.8 KB
report-images.json		643.2 KB
report-metadata.json		47.1 KB
sample-story-mementos.tsv		2.9 KB
summarize-collection.log		598.8 KB
report-entities.tsv		27.5 KB
report-sumgrams.tsv		822 bytes
summarize-errors.dat		8.8 KB

localhost:8000/uploads/wooey_files/summarize-a-collection-as-a-raintale-story-file/ee1aa4ee-18af-4d29-b842-7f16140f887e/raintale-story.json

Figure 15. A screenshot of a completed Hypercane job. The user can click the Download button (top right) and save the Raintale story file. (The red text only indicates a warning.)

Most users just wish to leverage Hypercane's storytelling capabilities. The recipe "Summarize a collection as a Raintale story file" accepts a collection ID for Archive-It, PANDORA Subject, PANDORA Collection, or Trove and executes the necessary `sample`, `report`, and `synthesize` commands to produce a Raintale story file. A user will need to supply the collection type, collection ID, sampling algorithm, and story title, as shown in Figure 14. Figure 15 shows the interface once the job is complete, allowing the user to download their Raintale story file.

Some users already have a list of memento URLs or wish to use mementos from a Memento-compliant archive not listed above. The recipe "Summarize archived pages as a Raintale story file" accepts an input file containing a newline-separated list of memento URLs and does the same as the recipe above.

Other users have already selected exemplars from web archives and do not need to execute the `sample` step. The recipe "Create a Raintale story file from archived page URLs" accepts an input file containing a newline-separated list of memento URLs and executes the necessary `report` and `synthesize` commands to produce a Raintale story file.

Sometimes archivists or researchers just need a list of the memento page URLs in a collection. The "Generate archived page URLs from a collection" recipe accepts a collection ID for Archive-It, PANDORA Subject, PANDORA Collection, or Trove and executes the necessary `identify` command to find all memento page URLs in the collection. Similarly, a user may want the list of original resource URLs in a collection. The "Generate original page URLs from a collection" recipe performs the same steps but outputs a list of original resource URLs. The user could also run the "identify by Collection ID" web form, but these recipes present simplified language and options.

Additionally, there are cases where a user will desire WARCs for a third-party application. The recipe "Synthesize WARCs from a collection" accepts a collection ID and will execute the necessary `identify` and `synthesize` steps to convert the mementos in the collection into WARCs that closely resemble what the archive initially crawled. Hypercane leverages the Memento protocol and its knowledge of accessing raw memento content from different web archives to make this happen. These WARCs are an approximation built from what Hypercane can access via the web and are not identical to the web archive's holdings. Still, they are sufficient for experimentation and analysis with tools like AUT. Likewise, a user can produce WARCs from a list of memento URLs with the recipe "Synthesize WARCs from archive page URLs."

These recipes continue to make Hypercane more approachable for end-users. Though Hypercane primarily exists to feed content to Raintale, it has evolved into a much more promising toolkit of its own.

MementoEmbed

MementoEmbed is a surrogate generation tool that can create four different types of surrogates. It produces social cards like those seen on Facebook and Twitter, page screenshots (also called browser thumbnails), word clouds, and **imagereels**. Imagereels are animated GIFs (GIF version 89a) of the top images found in a memento. Figure 16 displays one of MementoEmbed's social cards.

MementoEmbed also supports a web API for generating document metadata. Through this web API, a client can create surrogates. Clients can also request specific document metadata, such as page titles, memento-datetimes, image ranking, and automatic text summaries.

We developed MementoEmbed after analyzing more than 50 platforms and finding that none support mementos correctly. Some platforms do not understand how to extract the original resource URL from the memento. Others could not differentiate between a web archive's content and the memento content. Still, others refused to generate surrogates for mementos.

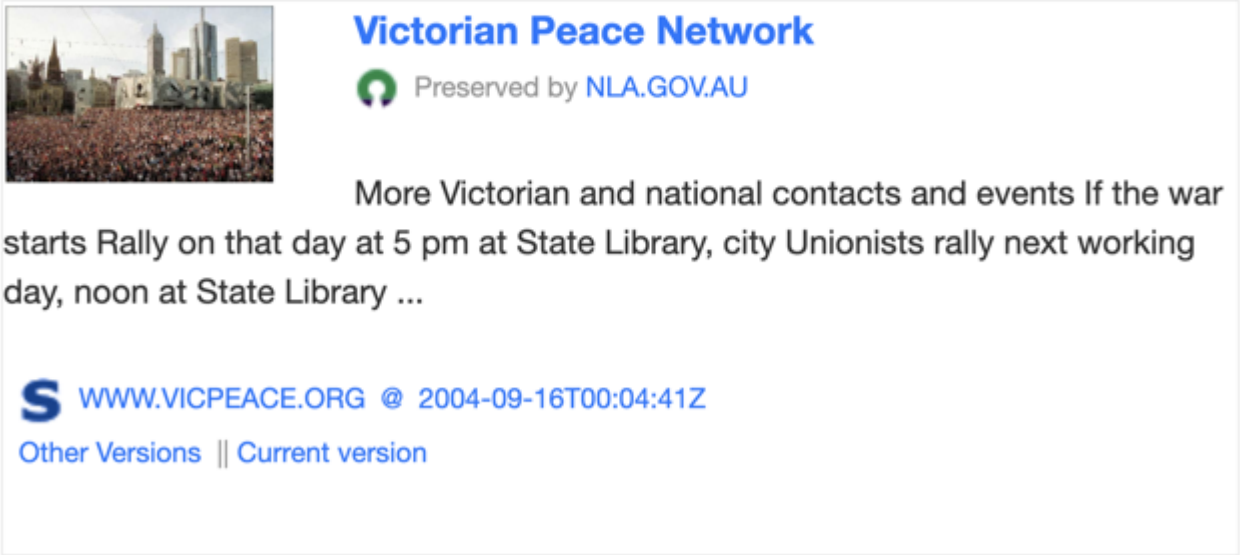
MementoEmbed is critical to the DSA Toolkit. Hypercane applies MementoEmbed's functions that discover images (*preprint version*) and raw memento content. Raintale is a client of MementoEmbed's web API.

Most web archives augment their mementos with navigation elements and branding to aid visitors. Unfortunately, these augmentations confuse natural language processing technology, like that implemented by Hypercane and MementoEmbed. For example, when comparing frequent terms among a set of mementos, the word "Trove" appears to be the most frequent word, even though it does not represent the actual memento content. For Wayback and OpenWayback web archives, like Archive-It, MementoEmbed simply inserts an `id_` flag into their memento URL to visit a page lacking these augmentations. Trove mementos required different handling.

Figure 17a displays a screenshot of a Trove memento. Figure 17b blurs its augmentations to emphasize the actual memento's content. As part of our pilot with the NLA, we had to update MementoEmbed to discard these augmentations when processing Trove's mementos. Finding the actual memento content required the following steps:

1. change the domain name in the memento URL from webarchive.nla.gov.au to web.archive.org.au
2. insert the `id_` flag into the memento URL

From here, MementoEmbed can now process the raw memento content.



Victorian Peace Network

Preserved by NLA.GOV.AU

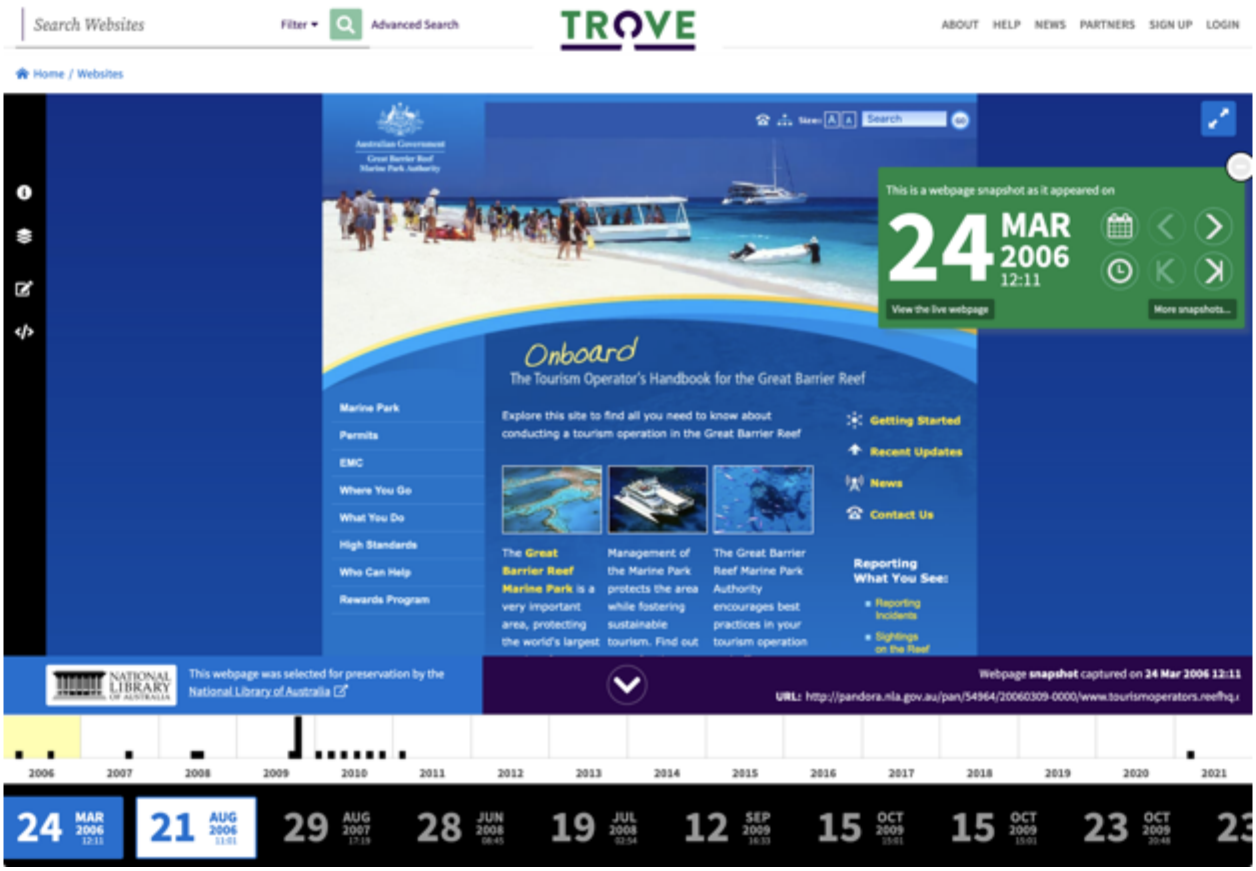
More Victorian and national contacts and events If the war starts Rally on that day at 5 pm at State Library, city Unionists rally next working day, noon at State Library ...

WWW.VICPEACE.ORG @ 2004-09-16T00:04:41Z

Other Versions || Current version

Figure 16. A screenshot of a MementoEmbed social card containing a striking image, title, description, original resource domain, and favicons, as well as information about the archive, when the archive preserved the page, and links to other mementos.

a. memento from Trove



This is a webpage snapshot as it appeared on 24 MAR 2006 12:11

View the live webpage

More snapshots...

Onboard

The Tourism Operator's Handbook for the Great Barrier Reef

Explore this site to find all you need to know about conducting a tourism operation in the Great Barrier Reef

The Great Barrier Reef Marine Park is a very important area, protecting the world's largest

Management of the Marine Park protects the area while fostering sustainable tourism. Find out

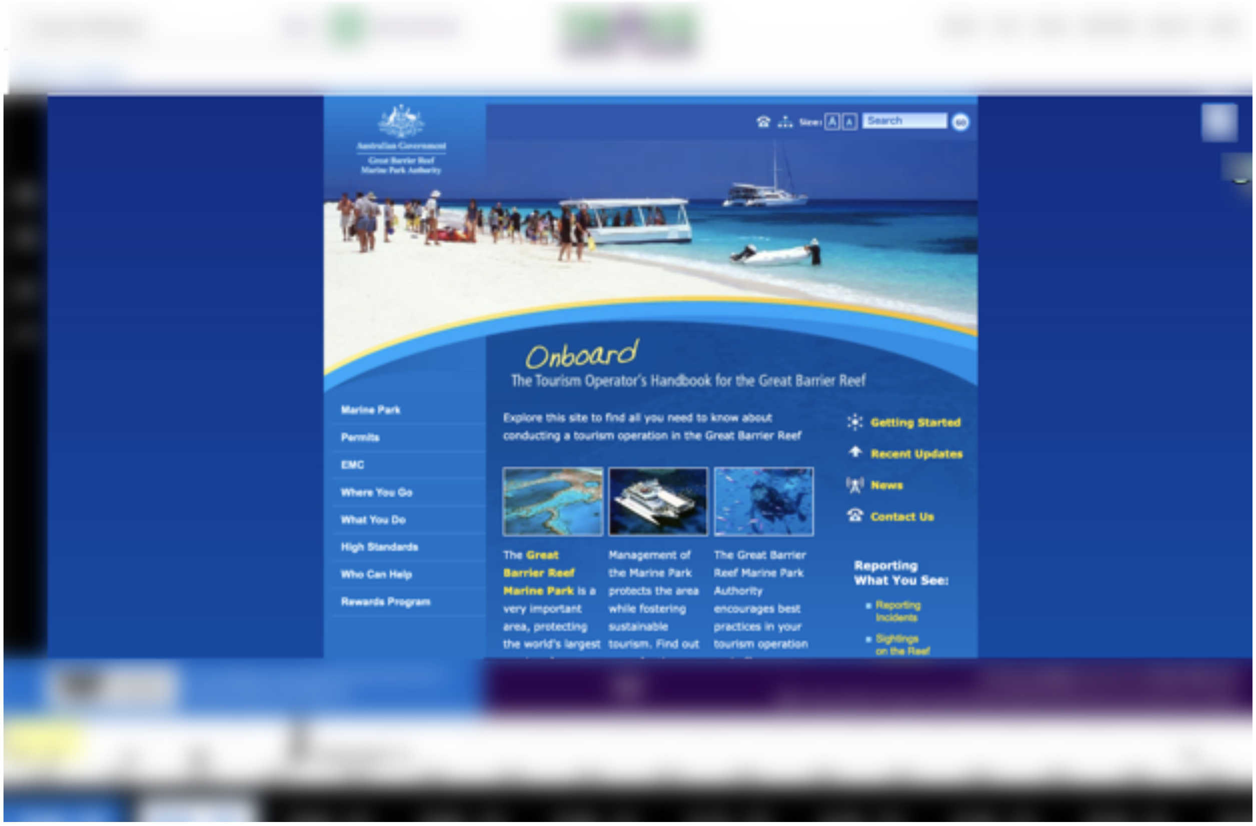
The Great Barrier Reef Marine Park Authority encourages best practices in your tourism operation

This webpage was selected for preservation by the National Library of Australia

Webpage snapshot captured on 24 Mar 2006 12:11

URL: http://pandora.nla.gov.au/pan/54964/20060309-0000/www.tourismoperators.reefhq.e

b. The content of the memento exists in the unblurred area.



Onboard

The Tourism Operator's Handbook for the Great Barrier Reef

Explore this site to find all you need to know about conducting a tourism operation in the Great Barrier Reef

The Great Barrier Reef Marine Park is a very important area, protecting the world's largest

Management of the Marine Park protects the area while fostering sustainable tourism. Find out

The Great Barrier Reef Marine Park Authority encourages best practices in your tourism operation

Figure 17. Trove mementos consist of augmentations that aid users in navigation and provide additional information that confuse natural language processing systems.

We also updated MementoEmbed to better handle creating screenshot surrogates of Trove's mementos. The additional augmentations required changing the timeouts on MementoEmbed's screenshot capabilities.

As with Hypercane, we created native Linux installers for MementoEmbed for CentOS 8 and Ubuntu 21.10 to help system administrators quickly stand up MementoEmbed. For more information on MementoEmbed, please see its documentation.

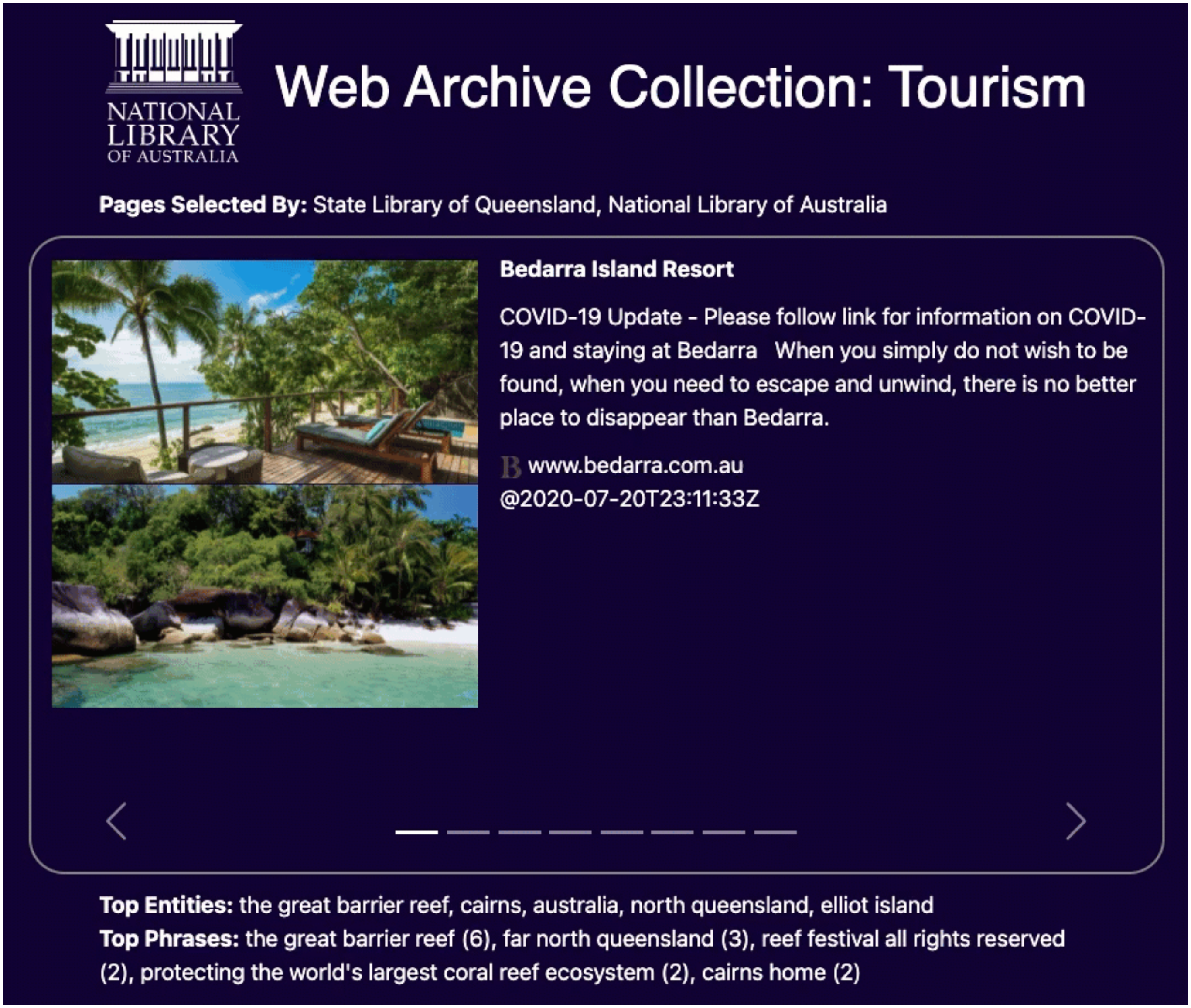
Raintale

Hypercane focuses on selecting exemplars and generating story metadata. MementoEmbed focuses on generating document metadata for a single memento. Raintale takes in all of this information and produces a story in a format that the storyteller desires. Raintale is a command-line application executed by the `tellstory` command. To apply the rich story file generated in the Hypercane section above, a user types:

```
# tellstory -i mystory.json --storyteller template \
    --story-template mytemplate.html -o mystory.html
```

The `-i` argument specifies the input file containing the information for the story. The `--storyteller` argument indicates the type of story being told. Because the above example suggests that we wish to use a template, the `--story-template` argument is needed to specify a file containing a template that Raintale will use to format the story. Finally, the `-o` argument indicates where to store the resulting story.

The user heavily controls the format of Raintale stories through templates. Figure 18 shows a story that applies a Bootstrap carousel template to a set of mementos to summarize the Trove collection *Tourism*. Figure 18 shows a story that applies an NLA-authored template to a group of hand-selected mementos about Australian Zoos, Wildlife Sanctuaries & Aquariums.



NATIONAL LIBRARY OF AUSTRALIA

Web Archive Collection: Tourism

Pages Selected By: State Library of Queensland, National Library of Australia

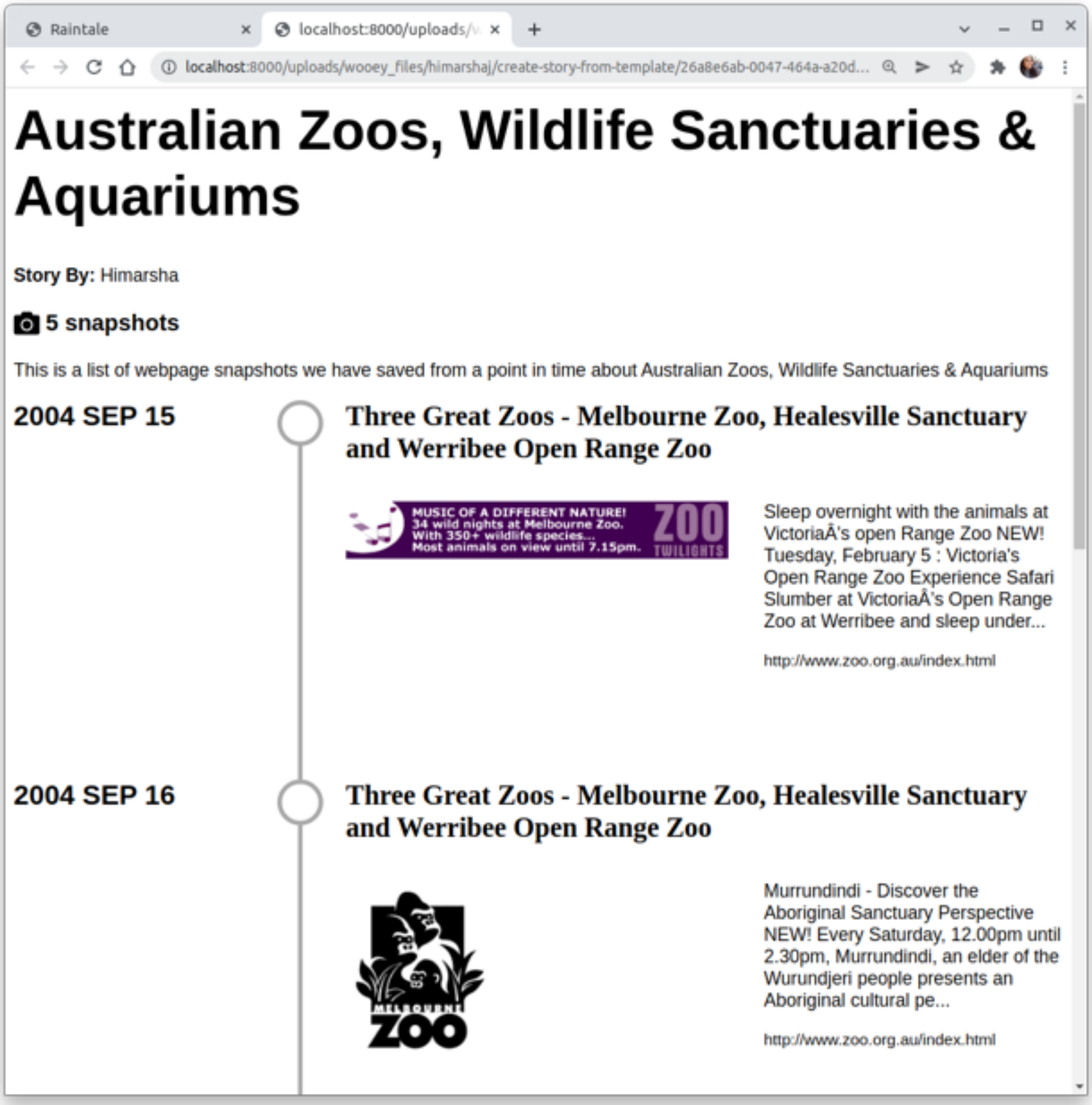
Bedarra Island Resort

COVID-19 Update - Please follow link for information on COVID-19 and staying at Bedarra   When you simply do not wish to be found, when you need to escape and unwind, there is no better place to disappear than Bedarra.

B www.bedarra.com.au
@2020-07-20T23:11:33Z

Top Entities: the great barrier reef, cairns, australia, north queensland, elliot island
Top Phrases: the great barrier reef (6), far north queensland (3), reef festival all rights reserved (2), protecting the world's largest coral reef ecosystem (2), cairns home (2)

Figure 18. A Raintale story generated using a Bootstrap carousel.


Raintale
localhost:8000/uploads/w
localhost:8000/uploads/wooey_files/himarshaj/create-story-from-template/26a8e6ab-0047-464a-a20d...
MUSIC OF A DIFFERENT NATURE! 34 wild nights at Melbourne Zoo. With 350+ wildlife species... Most animals on view until 7.15pm.
ZOO TWILIGHTS
MELBOURNE ZOO

Australian Zoos, Wildlife Sanctuaries & Aquariums

Story By: Himarsha

5 snapshots

This is a list of webpage snapshots we have saved from a point in time about Australian Zoos, Wildlife Sanctuaries & Aquariums

2004 SEP 15

Three Great Zoos - Melbourne Zoo, Healesville Sanctuary and Werribee Open Range Zoo

Sleep overnight with the animals at VictoriaÂ's open Range Zoo NEW! Tuesday, February 5 : Victoria's Open Range Zoo Experience Safari Slumber at VictoriaÂ's Open Range Zoo at Werribee and sleep under...

http://www.zoo.org.au/index.html

2004 SEP 16

Three Great Zoos - Melbourne Zoo, Healesville Sanctuary and Werribee Open Range Zoo

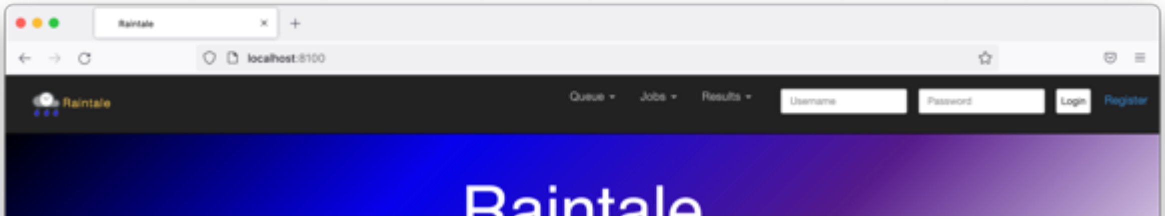
Murrundindi - Discover the Aboriginal Sanctuary Perspective NEW! Every Saturday, 12.00pm until 2.30pm, Murrundindi, an elder of the Wurundjeri people presents an Aboriginal cultural pe...

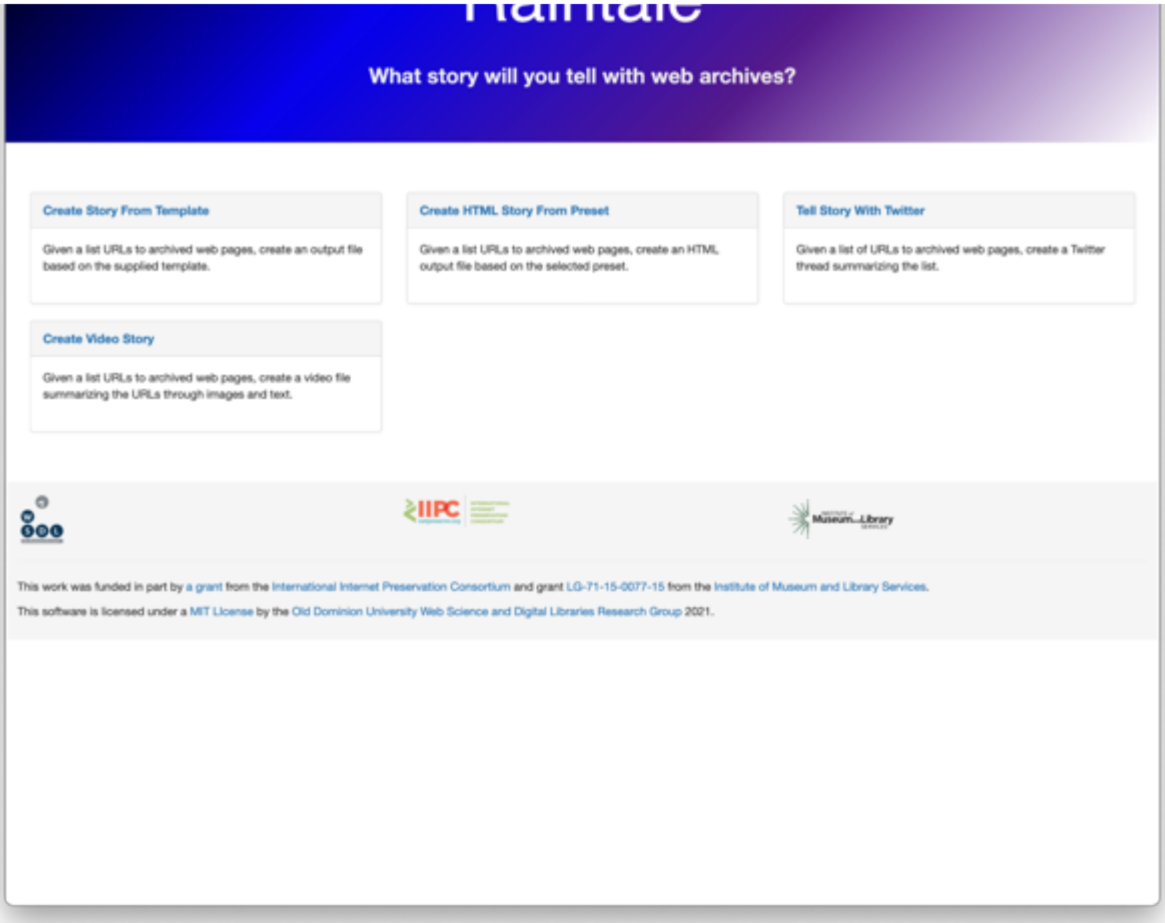
http://www.zoo.org.au/index.html

Figure 19. A Raintale story generated using a template developed by the National Library of Australia.

Because Raintale supports templates, users can construct stories that take many forms, from the COVID-19 story in Figure 6 to the stories shown in Figures 17 and 18. Raintale applies a modified version of the Jinja2 template language that supports additional options for extracting specific document metadata for the output. This template language does not limit Raintale to HTML. It permits many textual formats, like Markdown, XML, JSON, and Jekyll.

As with Hypercane, we recognized that a command-line application is not always approachable for all software users. As part of our pilot, we leveraged Wooey to create a WUI for Raintale as well. Figures 20–22 show screenshots of a user executing “Create Story From Template,” which produces a templated story just like we demonstrated with the `tellstory` command above. The “Create Story From Preset” option lets users forgo a template in favor of built-in story templates. With “Tell Story With Twitter,” Raintale will create a Twitter thread from the mementos in the input. Finally, “Create Video Story” generates an MP4 file consisting of the top-ranked images and text from each memento in the story.





Raintale

What story will you tell with web archives?

Create Story From Template

Given a list URLs to archived web pages, create an output file based on the supplied template.

Create HTML Story From Preset

Given a list URLs to archived web pages, create an HTML output file based on the selected preset.

Tell Story With Twitter

Given a list of URLs to archived web pages, create a Twitter thread summarizing the list.

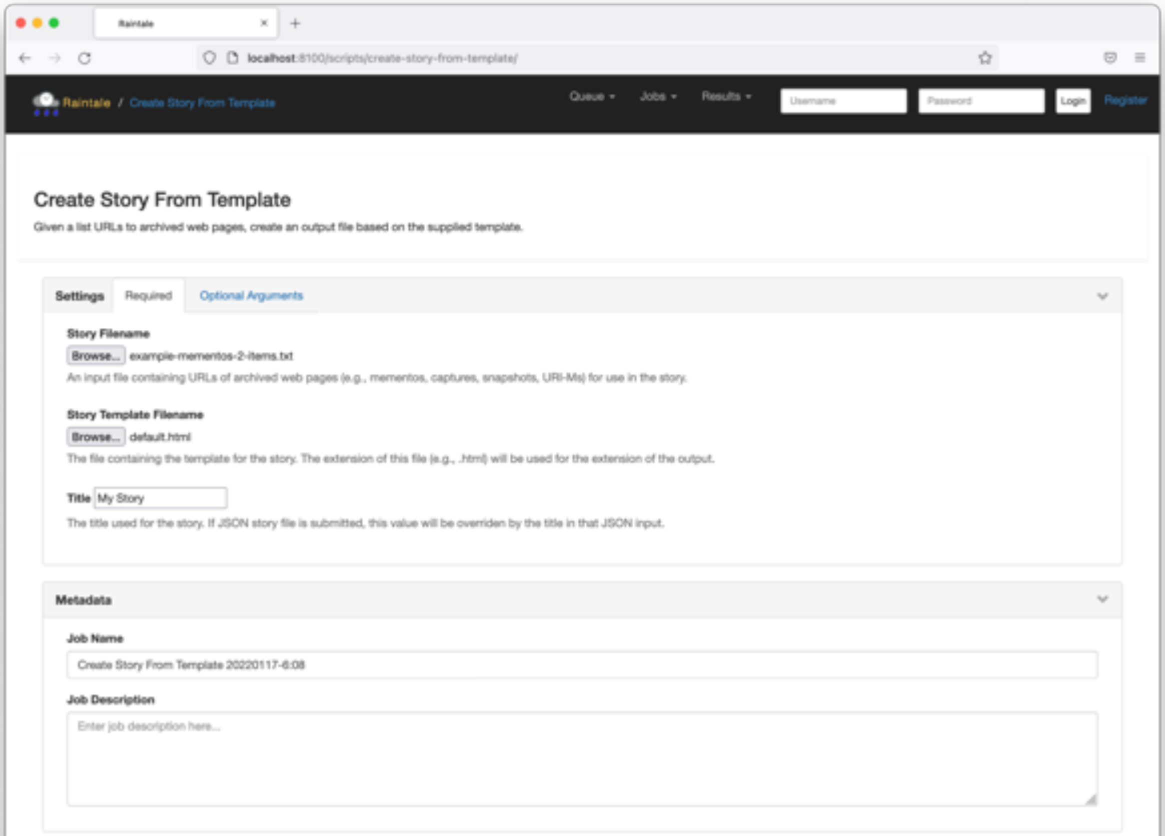
Create Video Story

Given a list URLs to archived web pages, create a video file summarizing the URLs through images and text.

This work was funded in part by a grant from the International Internet Preservation Consortium and grant LG-71-15-0077-15 from the Institute of Museum and Library Services.

This software is licensed under a MIT License by the Old Dominion University Web Science and Digital Libraries Research Group 2021.

Figure 20. A screenshot of the landing page of the Raintale WUI.



Raintale
localhost:8100/scripts/create-story-from-template/
Raintale / Create Story From Template
Queue
Jobs
Results
Username
Password
Login
Register

Create Story From Template

Given a list URLs to archived web pages, create an output file based on the supplied template.

Settings
Required
Optional Arguments

Story Filename

Browse... example-mementos-2-items.txt

An input file containing URLs of archived web pages (e.g., mementos, captures, snapshots, URI-Ms) for use in the story.

Story Template Filename

Browse... default.html

The file containing the template for the story. The extension of this file (e.g., .html) will be used for the extension of the output.

Title My Story

The title used for the story. If JSON story file is submitted, this value will be overriden by the title in that JSON input.

Metadata

Job Name

Create Story From Template 20220117-6:08

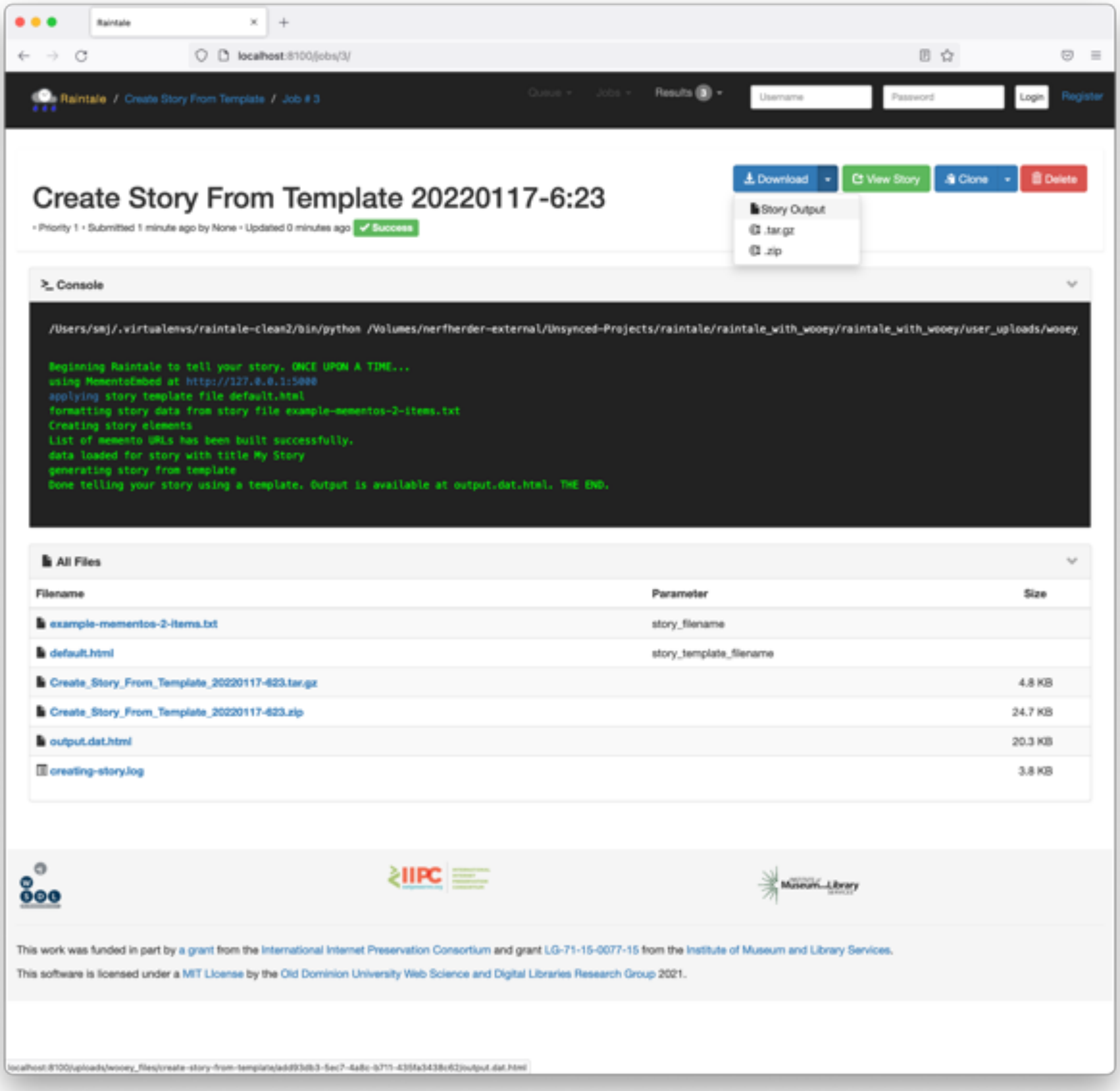
Job Description

Enter job description here...

This work was funded in part by a grant from the International Internet Preservation Consortium and grant LG-71-15-0077-15 from the Institute of Museum and Library Services.

This software is licensed under a MIT License by the Old Dominion University Web Science and Digital Libraries Research Group 2021.

Figure 21. A screenshot of the web form allowing a user to "Create Story From Template."



Create Story From Template 20220117-6:23

• Priority 1 • Submitted 1 minute ago by None • Updated 0 minutes ago

```
/Users/smj/.virtualenvs/raintale-clean2/bin/python /Volumes/nerfherder-external/Unsynced-Projects/raintale/raintale_with_wooey/raintale_with_wooey/user_uploads/wooey

Beginning Raintale to tell your story. ONCE UPON A TIME...
using MementoEmbed at http://127.0.0.1:5000
applying story template file default.html
formatting story data from story file example-mementos-2-items.txt
Creating story elements
List of memento URLs has been built successfully.
data loaded for story with title My Story
generating story from template
Done telling your story using a template. Output is available at output.dat.html. THE END.
```

Filename	Parameter	Size
example-mementos-2-items.txt	story_filename	
default.html	story_template_filename	
Create_Story_From_Template_20220117-623.tar.gz		4.8 KB
Create_Story_From_Template_20220117-623.zip		24.7 KB
output.dat.html		20.3 KB
creating-story.log		3.8 KB

This work was funded in part by a grant from the International Internet Preservation Consortium and grant LG-71-15-0077-15 from the Institute of Museum and Library Services.

This software is licensed under a MIT License by the Old Dominion University Web Science and Digital Libraries Research Group 2021.

localhost:8100/uploads/wooey_files/create-story-from-template/add93db3-5ec7-4a8c-b711-435fa3438c62/output.dat.html

Figure 22. A screenshot showing a completed Raintale job where the user can download the story.

As with Hypercane and MementoEmbed, we developed native CentOS 8 and Ubuntu 21.10 installers for Raintale. We presented Raintale at the WADL 2020 workshop and IIPC WAC 2021. An early version of Raintale helped us conduct user testing that showed that social cards probably provide a better understanding of web archive collections when compared to alternatives like page screenshots. For more information on installing and leveraging Raintale, please consult its documentation.

The Future of Storytelling With Web Archives

We see a lot of potential for further development of these tools. This pilot project has helped us better understand the needs of our users. We continue to improve the error handling of these tools as we learn more from the community. Overall, we want to improve the approachability of these tools.

Our native installers and Web User Interfaces are a start but only apply to Linux. Thanks to the experience of creating an installer for Linux, we have some ideas for a macOS DMG file that supports the applications of Hypercane, MementoEmbed, and Raintale.

With Wooey, we could preserve our existing command-line interfaces for scripting while also providing new web user interfaces. The authors of Gooey promise this same functionality using the same methods but with a native non-browser graphic user interface. We are hopeful that our efforts with Wooey might make it easier to apply Gooey and provide a proper native GUI interface.

As we build upon these technologies, we are charting a path toward making the DSA Toolkit a set of Microsoft Windows desktop applications. As a set of Python applications, the DSA Toolkit will have to contend with library installation issues on Windows. Once we resolve those issues, we will have reached the maximum number of users with the current code.

We are also exploring additional functionality. AIU, and, by proxy, Hypercane now support Archive-It, Trove, and PANDORA. We can expand it to support web archive collections from the Croatian Web Archive, the UK Web Archive, the Library of Congress, and Conifer. This way, users can apply our tools to summarize more collection types. Of course, we would continue to support sets of memento URLs as input because that would extend the reach to web archives that do not currently support collections.

The Hypercane recipes inspired by our collaboration are just the beginning. We are exploring the idea of a “Recipe Builder” where users can create their own Hypercane scripts but in a much more user-friendly fashion than shell scripting. Once this is in place, we hope that users will share their recipes. We could even facilitate a recipe exchange for interested archives that want to share Hypercane solutions.

Raintale supports templates so that archivists can customize their stories. We see a future where archivists can share the look and feel of their stories with each other. In addition to templates, we are considering the idea of a “Story Builder” allowing users to create stories graphically in real-time, much like bloggers do with WordPress blocks.

Our collaboration between LANL, ODU, and the NLA has been fruitful but was only the beginning. Either by executing these tools on your own or visiting someone else’s stories, what gems will you discover in a web archive collection?

Acknowledgements

Our many thanks go to the IIPC for funding the DSA pilot documented here. We also thank the IMLS for funding the DSA project through their storytelling (LG-71-15-0077-15) and CEDWARC (RE-70-18-0005-18) grants. This research was supported by the Information Science and Technology Institute and by the Laboratory Directed Research and Development program of Los Alamos National Laboratory (LANL) under project number 20210529CR. LANL is operated by Triad National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy (Contract No. 89233218CNA000001).

About the Authors

Shawn M. Jones (0000-0002-4372-870X), Los Alamos National Laboratory, is an Information Science & Technology Institute Postdoctoral Fellow working for LANL’s Information Sciences (CCS-3) division. Shawn recently received his Ph.D. from Old Dominion University and was advised by Dr. Michael L. Nelson. He is also an alumnus of ODU’s Web Science and Digital Libraries research group. He has contributed tools, data, and analysis to projects and frameworks such as Memento, Signposting, and Robust Links. In addition, Shawn is the lead of the Dark and Stormy Archives project, an initiative to summarize web archive collections through visualization techniques common in social media. More information about Shawn is available at: https://www.shawnmjones.org/.

Himarsha R. Jayanetti (0000-0003-4748-9176) is a Ph.D. student at Old Dominion University working under the supervision of Dr. Michele C. Weigle. She is also a member of the Web Science and Digital Libraries research group as a graduate research assistant. Her research interests are in digital preservation, digital libraries, and social media. She graduated from Gujarat Technological University in India with a Bachelor’s degree in Computer Engineering in 2017. More information about Himarsha is available at: https://himarshaj.github.io/.

Alex Osborne develops open source tools for web archiving and maintains the infrastructure of the Australian Web Archive at the National Library of Australia.

Paul Koerbin is Assistant-Director Web Archiving at the National Library of Australia. He has been involved with the development and operation of the NLA’s web archiving program since its inception in the 1990s, including being part of the team that developed one of the first workflow systems for selective web archiving. He has published papers and given many presentations, in Australia and overseas, on web archiving operations and practice. He holds a graduate qualification in library and information studies from the University of Tasmania and a Ph.D. from the University of Western Sydney.

Martin Klein (0000-0003-0130-2097), Los Alamos National Laboratory, is a scientist and lead of the Prototyping Team in LANL’s Research Library. In this role, he focuses on research and development efforts in the realm of web archiving, scholarly communication, digital system interoperability, and data management. He is involved in standards and frameworks such as Memento, ResourceSync, Signposting, and Robust Links. Martin holds a Diploma in Computer Science from the University of Applied Sciences Berlin, Germany, and a Ph.D. in Computer Science from Old Dominion University.

Michele C. Weigle (0000-0002-2787-7166) is a Professor of Computer Science at Old Dominion University. Her research interests include web science, social media, digital preservation, and information visualization. She has published over 115 articles in peer-reviewed conferences and journals and has served as PI or Co-PI on external research grants totaling almost $6M from a wide range of funders, including the National Science Foundation, the National Endowment for the Humanities, the Institute of Museum and Library Services, and the Andrew W. Mellon Foundation. She currently serves on the editorial boards of the *Journal of the Association for Information Science and Technology* (JASIST) and the *International Journal on Digital Libraries* (IJDL). Dr. Weigle received her Ph.D. in computer science from the University of North Carolina in 2003.

Michael L. Nelson (0000-0003-3749-8116) is a professor at Old Dominion University and the Virginia Modeling, Analysis, and Simulation Center (VMASC), and he co-leads the Web Science and Digital Libraries Research Group. Prior to joining ODU, he worked at NASA Langley Research Center from 1991-2002, where he created the NASA Technical Report Server (NTRS). More information about Dr. Nelson can be found at www.cs.odu.edu/~mln/ and twitter.com/phonedude_mln.

Subscribe to comments: For this article | For all articles