Summer 2021

Improving Collection Understanding for Web Archives with Storytelling: Shining Light Into Dark and Stormy Archives

Shawn M. Jones
Old Dominion University, jones.shawn.m@gmail.com

Recommended Citation

Jones, Shawn M.. "Improving Collection Understanding for Web Archives with Storytelling: Shining Light Into Dark and Stormy Archives" (2021). Doctor of Philosophy (PhD), Dissertation, Computer Science, Old Dominion University, DOI: 10.25777/zts6-v512
https://digitalcommons.odu.edu/computerscience_etds/131

IMPROVING COLLECTION UNDERSTANDING FOR WEB ARCHIVES WITH STORYTELLING: SHINING LIGHT INTO DARK AND STORMY ARCHIVES

by

Shawn M. Jones
M.S. December 2015, Old Dominion University
B.S. December 1999, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
August 2021

Approved by:

Michael L. Nelson (Director)

Michele C. Weigle (Member)

Sampath Jayarathna (Member)

Jian Wu (Member)

Jose Padilla (Member)

Martin Klein (Member)

ABSTRACT

IMPROVING COLLECTION UNDERSTANDING FOR WEB ARCHIVES WITH STORYTELLING: SHINING LIGHT INTO DARK AND STORMY ARCHIVES

Shawn M. Jones
Old Dominion University, 2021
Director: Dr. Michael L. Nelson

Collections are the tools that people use to make sense of an ever-increasing number of archived web pages. As collections themselves grow, we need tools to make sense of them. Tools that work on the general web, like search engines, are not a good fit for these collections because search engines do not currently represent multiple document versions well. Web archive collections are vast, some containing hundreds of thousands of documents. Thousands of collections exist, many of which cover the same topic. Few collections include standardized metadata. Too many documents from too many collections with insufficient metadata makes collection understanding an expensive proposition.

This dissertation establishes a five-process model to assist with web archive collection understanding. This model aims to produce a social media story – a visualization with which most web users are familiar. Each social media story contains surrogates which are summaries of individual documents. These surrogates, when presented together, summarize the topic of the story. After applying our storytelling model, they summarize the topic of a web archive collection.

We develop and test a framework to select the best exemplars that represent a collection. We establish that algorithms produced from these primitives select exemplars that are otherwise undiscoverable using conventional search engine methods. We generate story metadata to improve the information scent of a story so users can understand it better. After an analysis showing that existing platforms perform poorly for web archives and a user study establishing the best surrogate type, we generate document metadata for the exemplars with machine learning. We then visualize the story and document metadata together and distribute it to satisfy the information needs of multiple personas who benefit from our model.

Our tools serve as a reference implementation of our *Dark and Stormy Archives* storytelling model. *Hypercane* selects exemplars and generates story metadata. *MementoEmbed* generates document metadata. *Raintale* visualizes and distributes the story based on the story metadata and the document metadata of these exemplars. By providing understanding immediately, our stories

save users the time and effort of reading thousands of documents and, most importantly, help them understand web archive collections.

Dedicated to the my wife, Valentina, who continues to improve my soul.

ACKNOWLEDGEMENTS

Valentina Neblitt-Jones was vital to making sure this work happened. She sacrificed a position at Old Dominion University Libraries so that we could move to Los Alamos National Laboratory and work on this goal. She has carried me financially, spiritually, and emotionally through this endeavor. Her contributions did not stop there. With her combination of Computer Science expertise and Library Science experience, she helped me understand different Library Science perspectives on curation, collection management, metadata, surrogates, best practices, and more. She endured many hours of discussing these concepts so that I better understood all of the people we were helping with this work.

Michael L. Nelson was crucial to helping me change from the software engineer that the US Navy had made me into the researcher that I needed to be. He engaged in many hours of brainstorming the ideas that would bring this work to fruition. His ability to apply metaphor helped me grasp many new concepts. His hard work putting together a curriculum that combines information retrieval, web archiving, and web science helped ground me so I could accomplish this goal. Through our weekly meetings, he helped extend my vision beyond software engineering into new frontiers by advising me on how to review and incorporate the research of others.

Integral to helping me accomplish the work of getting this dissertation done was Michele C. Weigle. She kept me on track with the information necessary to keep this work moving. Her expertise in information visualization helped me understand some of the reasons why surrogates work. Without her networking and information visualization courses, I would not have the foundation necessary to fully appreciate what we accomplish here. She also committed to many hours of discussion, review, criticism, and support. Without her attention to detail, I would have missed mistakes. I hope that I can carry these lessons forward.

I would not have been able to accomplish this without the significant effort of Martin Klein. In addition to helping to analyze ideas in this dissertation, he also included me in projects that helped me expand my contributions and furthered my understanding of the landscape at the intersection of scholarly communications and web archiving. I also appreciate his support in helping me make additional contributions to the Memento ecosystem, some of which are the tools in this dissertation. Martin Klein, along with Harihar Shankar and Lyudmila Balakireva, helped me understand where research and prototyping fit into the spectrum of software engineering.

I would not have the opportunity to work at the intersection of scholarly communications and web archiving without the opportunities presented by Herbert Van de Sompel. Our collaborations continue to be some of my most highly cited works. He took a chance on hiring me as a member

of the Research Library Prototyping Team at Los Alamos National Laboratory. Thanks to Herbert, I have become a better researcher.

Before all of this, my friends Michael Olson, Kara Olson, Rick Hughes, and Scott Ainsworth were instrumental in converting me from a fearful, impetuous, reckless, and cavalier character into someone who thinks logically and can be successful in the professional world. My friends Kelly Durkin Ruth and Corey Ruth were also essential to helping me keep my sanity during my time at Los Alamos as well as during the COVID-19 pandemic. I also want to acknowledge the backing of my mother Susan Jones, my mother-in-law Valentina M. Neblitt, my father-in-law Frederick Neblitt, Sr., my sisters Christine Lewis, Stephanie Jones, and Cathleen Fareed, and my brother-in-law Frederick Neblitt, Jr. Without their efforts and support, I would not be able to achieve this goal and meet all of the wonderful people along the way.

I would not have been able to get this far without the efforts of Irwin Levinstein, Ravi Mukkamala, Mohammad Zubair, Sampath Jayarathna, Jian Wu, Larry Wilson, Hussein Abdel-Wahab, Chester Grosch, and Desh Ranjan. All of them took some time to help brainstorm ideas or address issues in my academic career. I will miss those who have passed on while I completed my graduate work. I know they would have wanted to see my defense.

Without the hard work of the staff at both the Old Dominion University Library and the Los Alamos National Laboratory Research Library, I would not have had access to the resources necessary to complete this dissertation. I would like to recognize Karen Vaughan for collaborating with Martin, Michael, Herbert, and me on our Robust Links work.

Finally, the input and encouragement of the ODU Computer Science WS-DL team was essential to my success. I could not have done this without the support and collaboration of Alexander Nwala, Mat Kelly, Sawood Alam, Mohamed Aturban, Ahmed AlSum, Yasmin AlNoamany, Justin Brunelle, Hany SalahEldeen, Lulwah Alkwai, and Himarsha Jayanetti.

This work was supported in part by ACM SIGIR Travel Grants, IMLS RE-70-18-0005-18, IMLS LG-71-15-0077-15, Andrew W. Mellon Foundation #11600663, and IIPC.

“I rather think that archives exist to keep things safe, but not secret.”
– Kevin Young

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

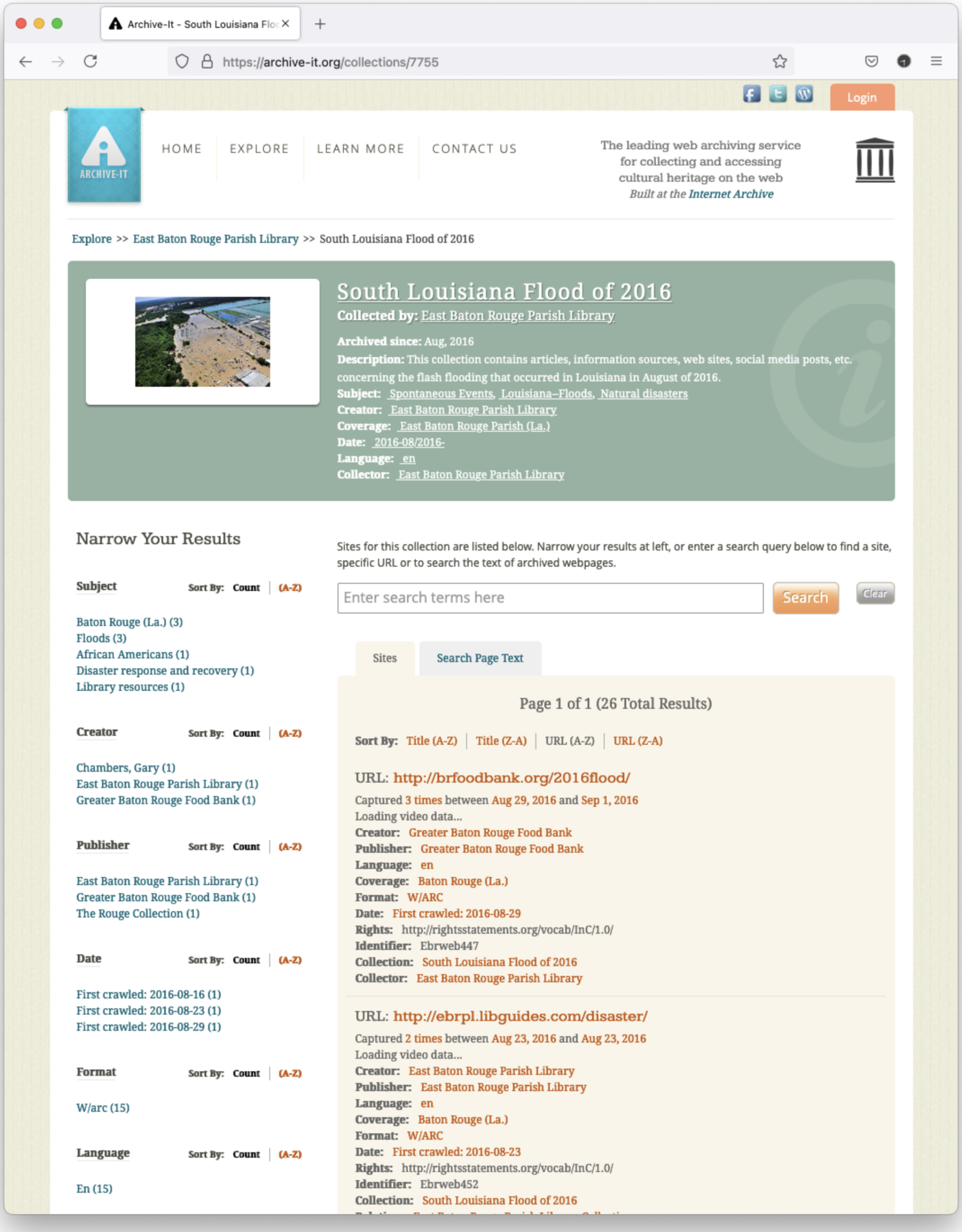
CHAPTER 1

INTRODUCTION

Natasha is a historian focusing on how disasters shape cultures. She is currently studying the South Louisiana Flood of 2016. Even though resources, like Wikipedia [323], exist that summarize the event, Natasha is more interested in the news reporting at the time of the event [100, 104, 235]. Attempting to discover contemporary news articles is difficult. Most authors update news articles with the most current information. As articles age, they slide further down in search engine results [246], becoming more challenging to discover. Fortunately, Natasha knows about web archives. Web archives capture the content of web pages as they existed at a specific point in time. Natasha could search these archives to find news stories about this flood, but others had the foresight to create themed web archive collections of news articles and other web pages that existed at the time of the flood. Archivists created these collections with Archive-It, a web archive collection service initiated by the Internet Archive [217].

With Archive-It, a curator can select web pages, or **seeds**, and then instruct the system to capture versions of these seeds as **mementos**. These collections are a fantastic resource for Natasha. Each memento is an observation from a specific time and hence is a version of the seed. There can be many mementos per seed. As seen in Figure 1, Archive-It has two collections about the South Louisiana Flood of 2016. Using Archive-It's user interface, she sees that one collection has 26 seeds and the other has 68. Due to time constraints, she can only choose one for her research. To understand the difference between these collections, she would need to review the hundreds of mementos generated from these seeds. Hers is not the only case where multiple collections exist for the same topic. As shown in Figure 2, there are at least 44 collections on Archive-It related to “human rights”. Choosing the wrong collection will result in time wasted. Which should she review?

Rustam is a journalist who needs to study the events of the Boston Marathon Bombing as they unfolded. Rustam is working with an Archive-It collection specifically about the event. He is interested in how specific pages changed throughout the event. Olayinka has a similar problem but in a different dimension. She is reviewing news and wants to know what different news sources discussed at a given date and time. Where Rustam wants to see how a page changed over time, Olayinka wants to understand how news articles covered a specific timespan. How do they do this? How do they not only select these items but also visualize them for themselves and others?



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

Explore >> East Baton Rouge Parish Library >> South Louisiana Flood of 2016

South Louisiana Flood of 2016

Collected by: East Baton Rouge Parish Library

Archived since: Aug, 2016
Description: This collection contains articles, information sources, web sites, social media posts, etc. concerning the flash flooding that occurred in Louisiana in August of 2016.
Subject: Spontaneous Events, Louisiana–Floods, Natural disasters
Creator: East Baton Rouge Parish Library
Coverage: East Baton Rouge Parish (La.)
Date: 2016-08/2016-
Language: en
Collector: East Baton Rouge Parish Library

Narrow Your Results

Subject Sort By: Count | (A-Z)

Baton Rouge (La.) (3)
Floods (3)
African Americans (1)
Disaster response and recovery (1)
Library resources (1)

Creator Sort By: Count | (A-Z)

Chambers, Gary (1)
East Baton Rouge Parish Library (1)
Greater Baton Rouge Food Bank (1)

Publisher Sort By: Count | (A-Z)

East Baton Rouge Parish Library (1)
Greater Baton Rouge Food Bank (1)
The Rouge Collection (1)

Date Sort By: Count | (A-Z)

First crawled: 2016-08-16 (1)
First crawled: 2016-08-23 (1)
First crawled: 2016-08-29 (1)

Format Sort By: Count | (A-Z)

W/arc (15)

Language Sort By: Count | (A-Z)

En (15)

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

Page 1 of 1 (26 Total Results)

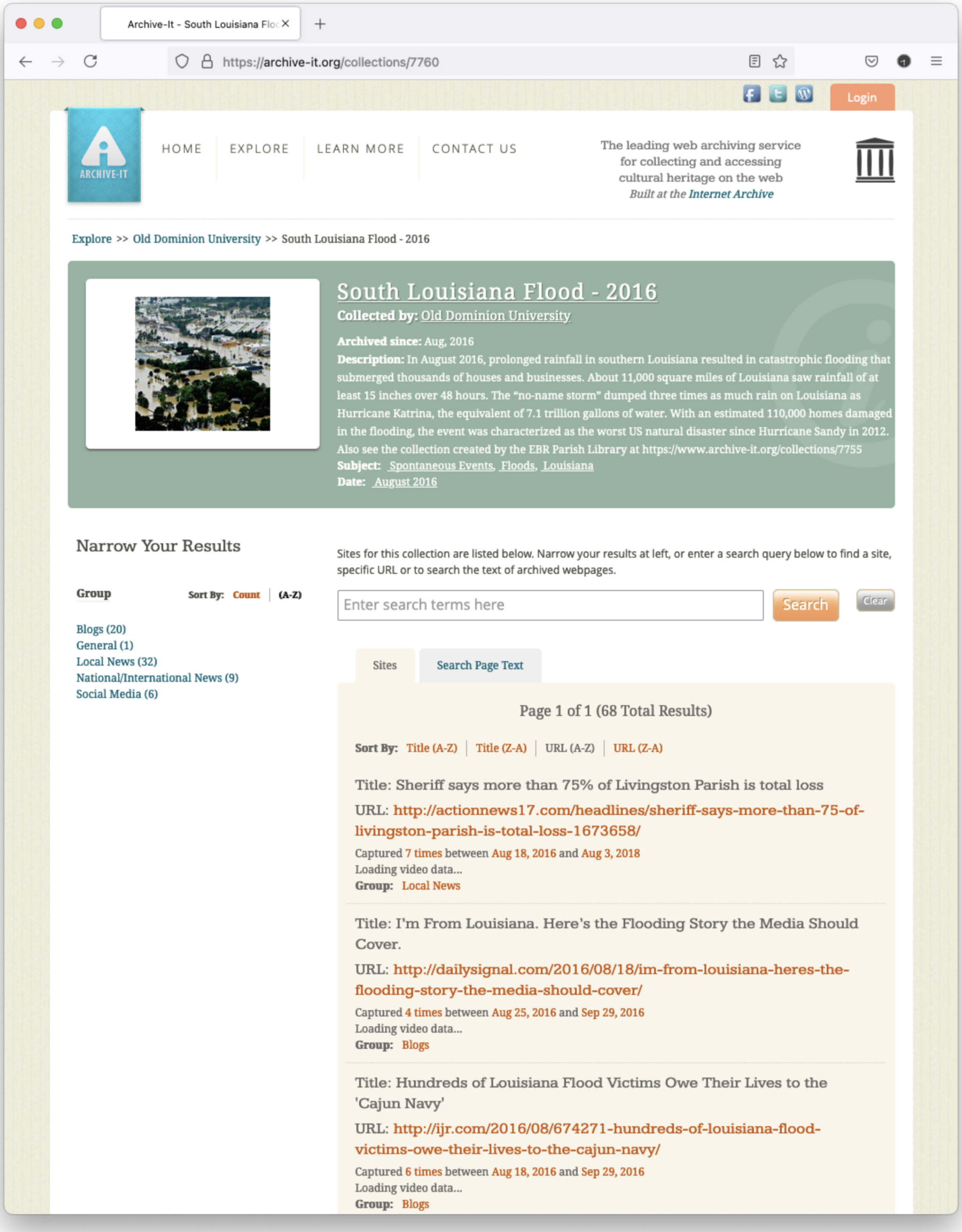
Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

URL: http://brfoodbank.org/2016flood/
Captured 3 times between Aug 29, 2016 and Sep 1, 2016
Loading video data...
Creator: Greater Baton Rouge Food Bank
Publisher: Greater Baton Rouge Food Bank
Language: en
Coverage: Baton Rouge (La.)
Format: W/ARC
Date: First crawled: 2016-08-29
Rights: http://rightsstatements.org/vocab/InC/1.0/
Identifier: Ebrweb447
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: W/ARC
Date: First crawled: 2016-08-23
Rights: http://rightsstatements.org/vocab/InC/1.0/
Identifier: Ebrweb452
Collection: South Louisiana Flood of 2016

(a) Archive-It collection 7755: *South Louisiana Flood of 2016*

Fig. 1. Archive-It has two collections – with identifiers 7755 and 7760 – about the Louisiana Flood of 2016 - which one is right for Natasha? (Screenshots taken in June 2021.)


Archive-It - South Louisiana Flo
https://archive-it.org/collections/7760
Login
ARCHIVE-IT
HOME
EXPLORE
LEARN MORE
CONTACT US
The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive

Explore >> Old Dominion University >> South Louisiana Flood - 2016

South Louisiana Flood - 2016

Collected by: Old Dominion University

Archived since: Aug, 2016

Description: In August 2016, prolonged rainfall in southern Louisiana resulted in catastrophic flooding that submerged thousands of houses and businesses. About 11,000 square miles of Louisiana saw rainfall of at least 15 inches over 48 hours. The “no-name storm” dumped three times as much rain on Louisiana as Hurricane Katrina, the equivalent of 7.1 trillion gallons of water. With an estimated 110,000 homes damaged in the flooding, the event was characterized as the worst US natural disaster since Hurricane Sandy in 2012. Also see the collection created by the EBR Parish Library at https://www.archive-it.org/collections/7755

Subject: Spontaneous Events, Floods, Louisiana

Date: August 2016

Narrow Your Results

Group **Sort By:** **Count** | **(A-Z)**

Blogs (20)
General (1)
Local News (32)
National/International News (9)
Social Media (6)

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.

Page 1 of 1 (68 Total Results)

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

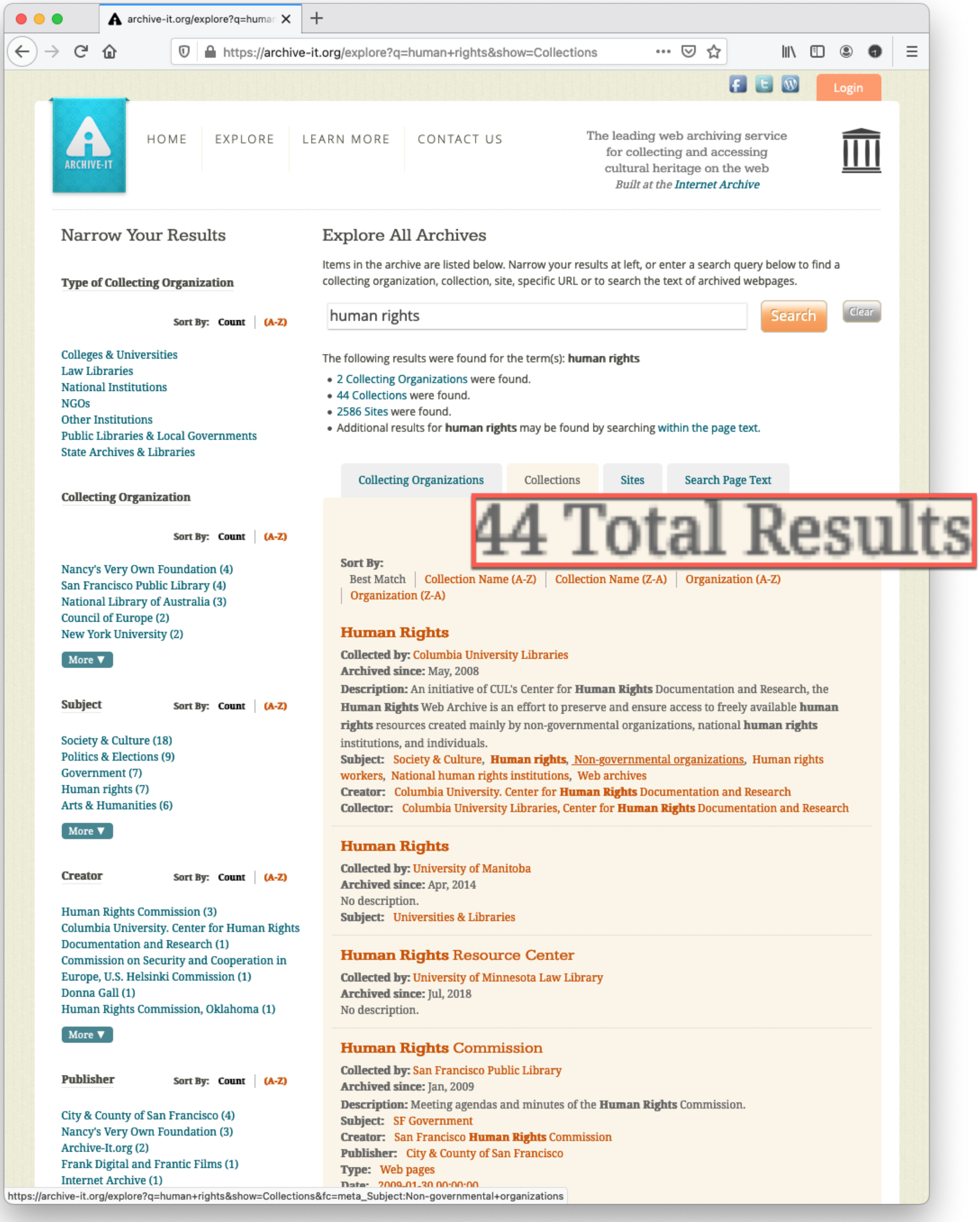
Title: Sheriff says more than 75% of Livingston Parish is total loss
URL: http://actionnews17.com/headlines/sheriff-says-more-than-75-of-livingston-parish-is-total-loss-1673658/
Captured 7 times between Aug 18, 2016 and Aug 3, 2018
Loading video data...
Group: Local News

Title: I'm From Louisiana. Here's the Flooding Story the Media Should Cover.
URL: http://dailysignal.com/2016/08/18/im-from-louisiana-heres-the-flooding-story-the-media-should-cover/
Captured 4 times between Aug 25, 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 6 times between Aug 18, 2016 and Sep 29, 2016
Loading video data...
Group: Blogs

(b) Archive-It collection 7760: *South Louisiana Flood - 2016*

Fig. 1. (Continued) Archive-It has two collections – with identifiers 7755 and 7760 – about the Louisiana Flood of 2016 - which one is right for Natasha? (Screenshots taken in June 2021.)


archive-it.org/explore?q=human
https://archive-it.org/explore?q=human+rights&show=Collections
Login
ARCHIVE-IT
HOME
EXPLORE
LEARN MORE
CONTACT US

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

Narrow Your Results

Type of Collecting Organization

Sort By: Count | (A-Z)

Colleges & Universities
Law Libraries
National Institutions
NGOs
Other Institutions
Public Libraries & Local Governments
State Archives & Libraries

Collecting Organization

Sort By: Count | (A-Z)

Nancy's Very Own Foundation (4)
San Francisco Public Library (4)
National Library of Australia (3)
Council of Europe (2)
New York University (2)

More ▼

Subject Sort By: Count | (A-Z)

Society & Culture (18)
Politics & Elections (9)
Government (7)
Human rights (7)
Arts & Humanities (6)

More ▼

Creator Sort By: Count | (A-Z)

Human Rights Commission (3)
Columbia University. Center for Human Rights Documentation and Research (1)
Commission on Security and Cooperation in Europe, U.S. Helsinki Commission (1)
Donna Gall (1)
Human Rights Commission, Oklahoma (1)

More ▼

Publisher Sort By: Count | (A-Z)

City & County of San Francisco (4)
Nancy's Very Own Foundation (3)
Archive-It.org (2)
Frank Digital and Frantic Films (1)
Internet Archive (1)

Explore All Archives

Items in the archive are listed below. Narrow your results at left, or enter a search query below to find a collecting organization, collection, site, specific URL or to search the text of archived webpages.

human rights Search Clear

The following results were found for the term(s): **human rights**

- 2 Collecting Organizations were found.
- 44 Collections were found.
- 2586 Sites were found.
- Additional results for **human rights** may be found by searching within the page text.

Collecting Organizations | Collections | Sites | Search Page Text

44 Total Results

Sort By:
Best Match | Collection Name (A-Z) | Collection Name (Z-A) | Organization (A-Z) | Organization (Z-A)

Human Rights
Collected by: Columbia University Libraries
Archived since: May, 2008
Description: An initiative of CUL's Center for **Human Rights** Documentation and Research, the **Human Rights** Web Archive is an effort to preserve and ensure access to freely available **human rights** resources created mainly by non-governmental organizations, national **human rights** institutions, and individuals.
Subject: Society & Culture, **Human rights**, Non-governmental organizations, Human rights workers, National human rights institutions, Web archives
Creator: Columbia University. Center for **Human Rights** Documentation and Research
Collector: Columbia University Libraries, Center for **Human Rights** Documentation and Research

Human Rights
Collected by: University of Manitoba
Archived since: Apr, 2014
No description.
Subject: Universities & Libraries

Human Rights Resource Center
Collected by: University of Minnesota Law Library
Archived since: Jul, 2018
No description.

Human Rights Commission
Collected by: San Francisco Public Library
Archived since: Jan, 2009
Description: Meeting agendas and minutes of the **Human Rights** Commission.
Subject: SF Government
Creator: San Francisco **Human Rights** Commission
Publisher: City & County of San Francisco
Type: Web pages
Date: 2009-01-30 00:00:00

https://archive-it.org/explore?q=human+rights&show=Collections&fc=meta_Subject:Non-governmental+organizations

Fig. 2. There are 44 results (magnified for emphasis) for a search of "human rights" in Archive-It, meaning that there are 44 collections, each consisting of many seeds and mementos, to review for this query. (Screenshot taken May 2021.)

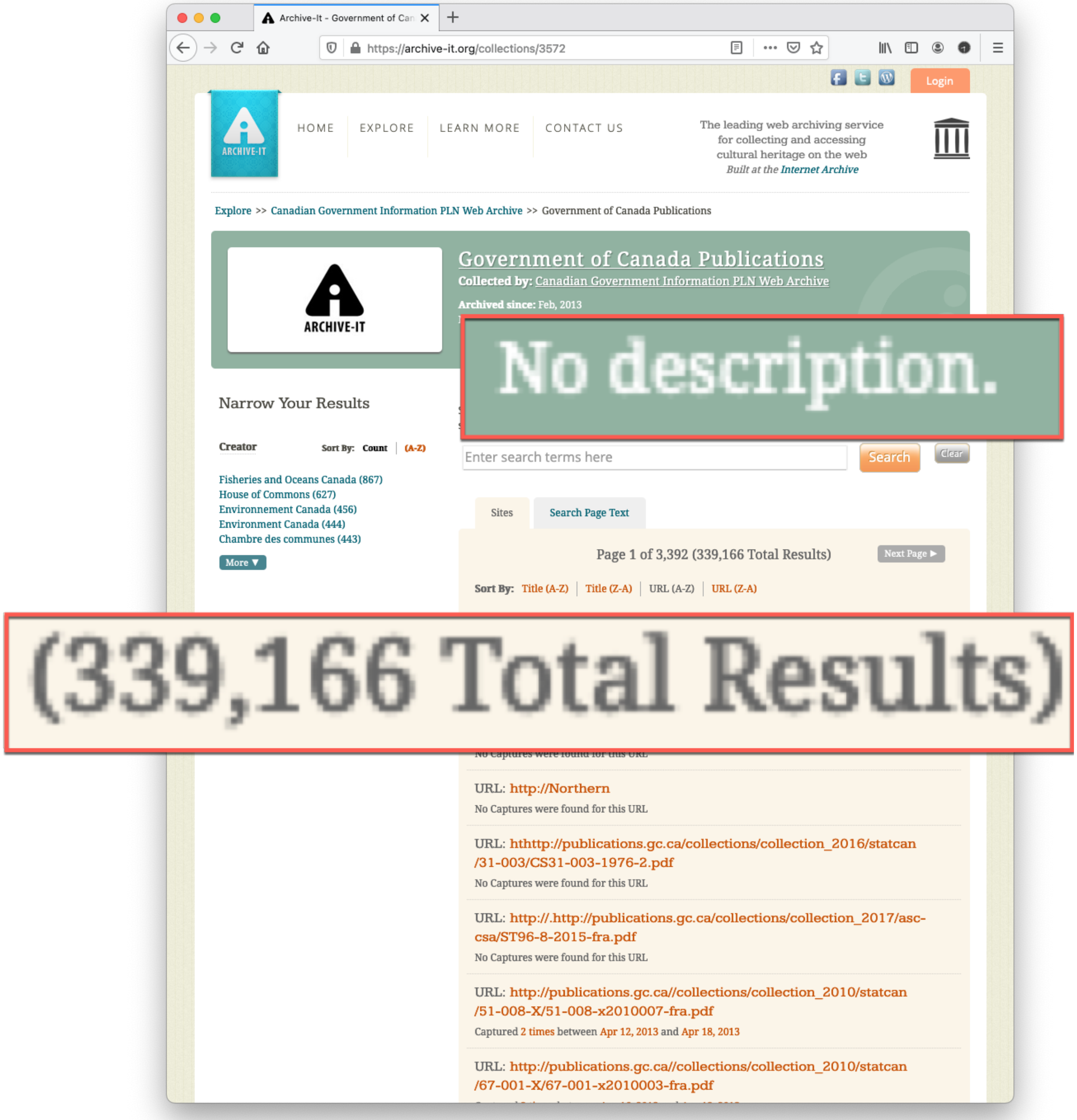
These researchers are not the only ones who need to interpret web archive collections. Elbert is an archivist who wants to promote his collection to help people like Natasha, Rustam, and Olayinka notice it and use it for their research. He wants to create a visualization that is enticing while ensuring that people understand the collection with minimal time and effort. Ling is an archivist who has inherited a web archive collection from another archivist. Like the collection shown in Figure 3, it contains thousands of documents, but no metadata. She needs to know what it contains so she can make decisions about it. How can these archivists meet their goals?

All of these people require a faster method of conveying **collection understanding**. A user could try to find individual pages on their topic in a general web archive, like the Internet Archive. Curators provide the benefit of bounding a set of mementos into a themed collection, imbuing the whole with meaning, thus saving these users time by reducing the number of potential resources to review. The curator of a web archive collection is not necessarily the only user of a web archive collection. The continued value of web archive collections comes when others like Natasha, Rustam, Olayinka, Elbert, and Ling can reuse these same collections for their work. How do these researchers and archivists process and present their collections so that they can meet their goals?

For other works, like books or museum pieces, metadata is a useful tool for evaluating an item. Metadata exists for Archive-It collections as magnified in Figure 4. Different curators create different web archive collections for different purposes. Some curators create collections to fulfill regulatory requirements, while others are trying to document unfolding events in real-time. The metadata is inconsistently applied, likely due to differences in content standards and rules of interpretation among these curators. A user cannot reliably compare metadata fields to understand the differences between collections. To provide context for each seed, a curator must generate metadata for that seed. For collections with many seeds, manually generating this metadata is a taxing and potentially unrealistic proposition. Archive-It reflects this. As shown in Figure 5, collections consisting of hundreds of thousands of seeds often contain no seed level metadata. As the need for metadata rises (i.e., more seeds), its availability falls, likely due to the effort needed to create it manually.

If metadata is valuable, then why not have someone generate metadata for these collections sometime later? In addition to the issues of scale within a collection, as of June 2021, there were more than 14,000 collections on Archive-It. Furthermore, archivists create new Archive-It collections every year, as shown in Figure 6, and the rate of creation is accelerating.

Natasha is faced with the problem of understanding if a single collection is worth her time. Rustam and Olayinka want to view a specific aspect of interest to them. All are faced with collections consisting of thousands of documents and the vastness of Archive-It's 14,000+ collections.



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

Explore >> Canadian Government Information PLN Web Archive >> Government of Canada Publications

Government of Canada Publications
Collected by: Canadian Government Information PLN Web Archive
Archived since: Feb, 2013

No description.

Narrow Your Results

Creator Sort By: Count | (A-Z)

Fisheries and Oceans Canada (867)
House of Commons (627)
Environnement Canada (456)
Environment Canada (444)
Chambre des communes (443)

Enter search terms here

Page 1 of 3,392 (339,166 Total Results)

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

(339,166 Total Results)

No Captures were found for this URL

URL: http://Northern
No Captures were found for this URL

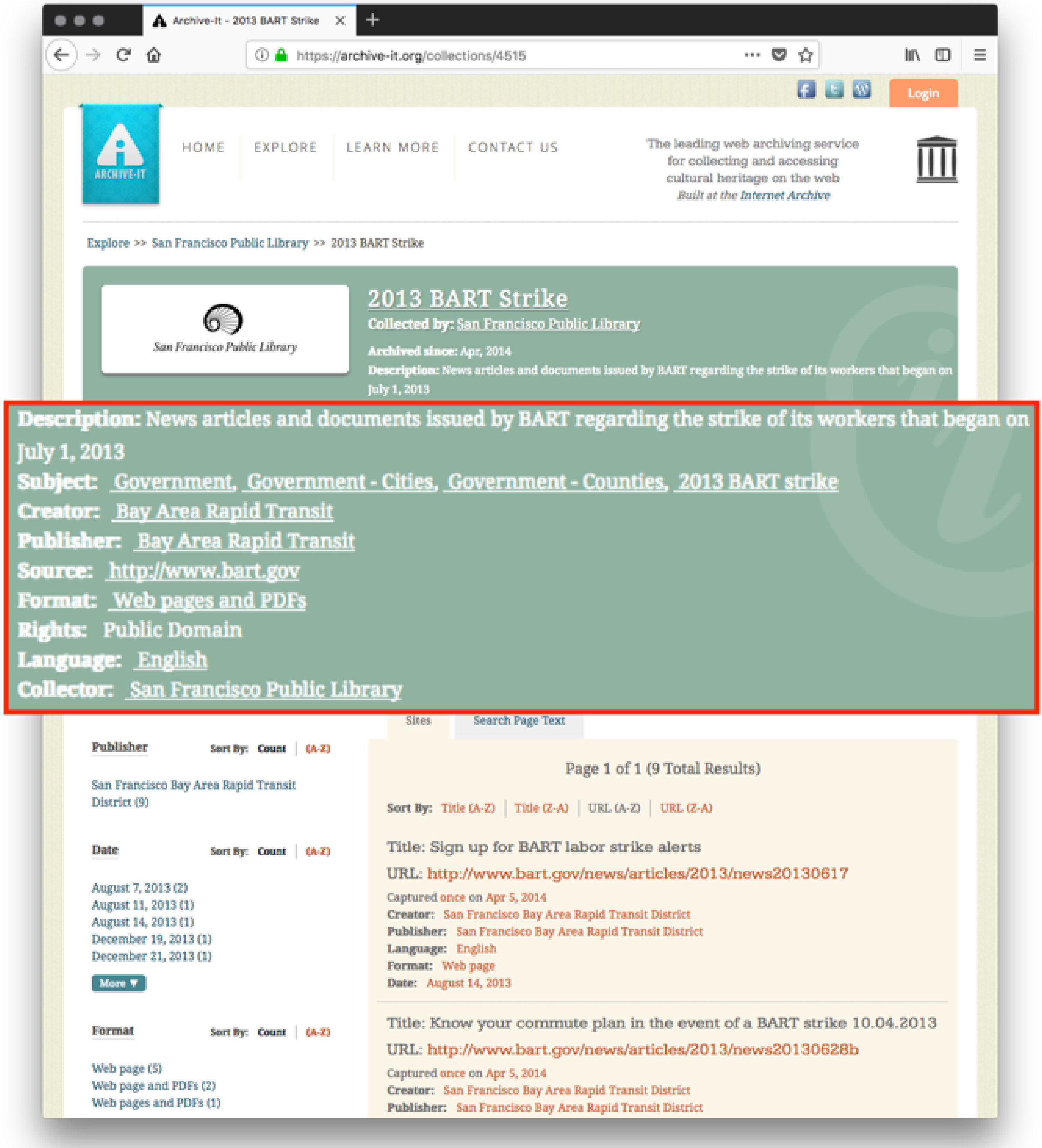
URL: hthttp://publications.gc.ca/collections/collection_2016/statcan/31-003/CS31-003-1976-2.pdf
No Captures were found for this URL

URL: http://.http://publications.gc.ca/collections/collection_2017/asc-csa/ST96-8-2015-fra.pdf
No Captures were found for this URL

URL: http://publications.gc.ca//collections/collection_2010/statcan/51-008-X/51-008-x2010007-fra.pdf
Captured 2 times between Apr 12, 2013 and Apr 18, 2013

URL: http://publications.gc.ca//collections/collection_2010/statcan/67-001-X/67-001-x2010003-fra.pdf

Fig. 3. The *Government of Canada Publications* consists of 339,166 seeds (magnified for emphasis) and some seeds have multiple mementos. There is also no metadata supplied on this collection. Human review of this collection will require a lot of time. (Screenshot taken May 2021.)


Archive-It - 2013 BART Strike
https://archive-it.org/collections/4515
Login
ARCHIVE-IT
HOME
EXPLORE
LEARN MORE
CONTACT US
The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive
San Francisco Public Library
Sites
Search Page Text
More ▼

Explore >> San Francisco Public Library >> 2013 BART Strike

2013 BART Strike

Collected by: San Francisco Public Library

Archived since: Apr, 2014

Description: News articles and documents issued by BART regarding the strike of its workers that began on July 1, 2013

Description: News articles and documents issued by BART regarding the strike of its workers that began on July 1, 2013

Subject: Government, Government - Cities, Government - Counties, 2013 BART strike
Creator: Bay Area Rapid Transit
Publisher: Bay Area Rapid Transit
Source: http://www.bart.gov
Format: Web pages and PDFs
Rights: Public Domain
Language: English
Collector: San Francisco Public Library

Publisher **Sort By:** **Count** | (A-Z)

San Francisco Bay Area Rapid Transit District (9)

Date **Sort By:** **Count** | (A-Z)

August 7, 2013 (2)
August 11, 2013 (1)
August 14, 2013 (1)
December 19, 2013 (1)
December 21, 2013 (1)

Format **Sort By:** **Count** | (A-Z)

Web page (5)
Web page and PDFs (2)
Web pages and PDFs (1)

Page 1 of 1 (9 Total Results)

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

Title: Sign up for BART labor strike alerts
URL: http://www.bart.gov/news/articles/2013/news20130617
Captured once on Apr 5, 2014
Creator: San Francisco Bay Area Rapid Transit District
Publisher: San Francisco Bay Area Rapid Transit District
Language: English
Format: Web page
Date: August 14, 2013

Title: Know your commute plan in the event of a BART strike 10.04.2013
URL: http://www.bart.gov/news/articles/2013/news20130628b
Captured once on Apr 5, 2014
Creator: San Francisco Bay Area Rapid Transit District
Publisher: San Francisco Bay Area Rapid Transit District

Fig. 4. This screenshot of Archive-It collection 4515 displays its collection-wide metadata, magnified for emphasis. (Screenshot taken in 2018.)

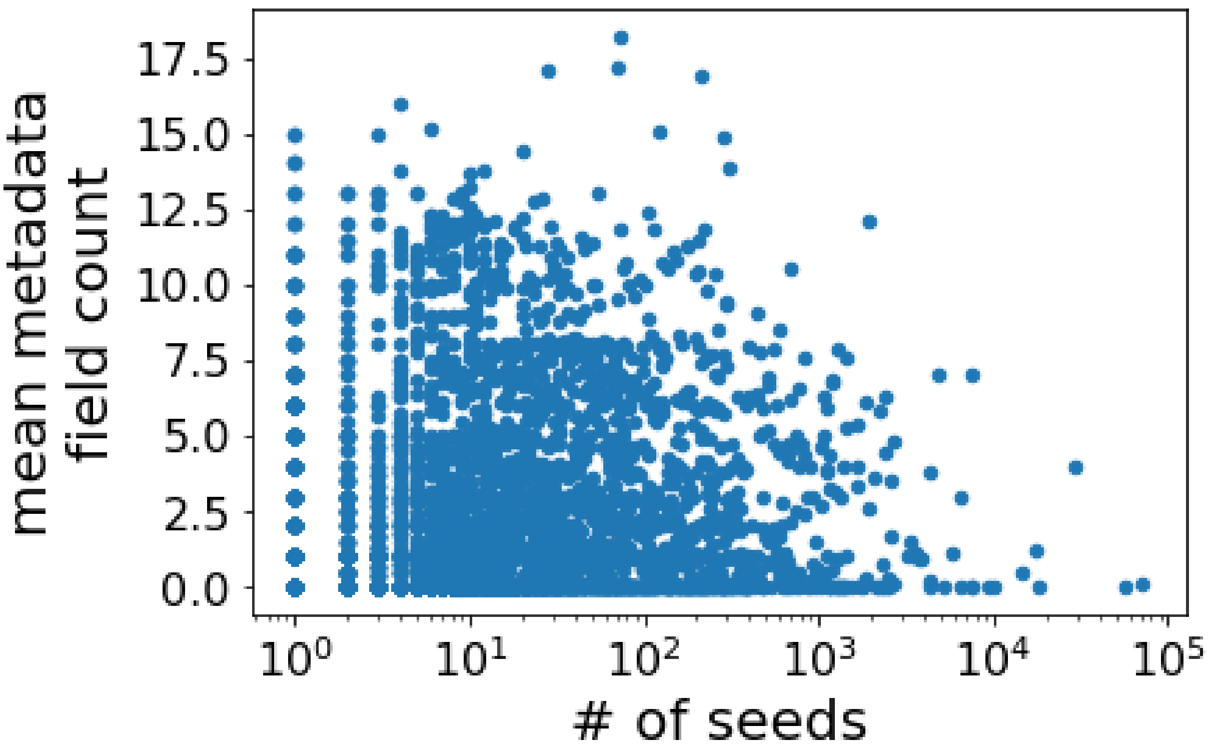


Fig. 5. A visualization of metadata field usage across all Archive-It seeds [165]. Generally, as the number of seeds in a collection goes up, the amount of metadata available goes down. Thus, the more metadata a reader needs to understand a collection, the less they have available.

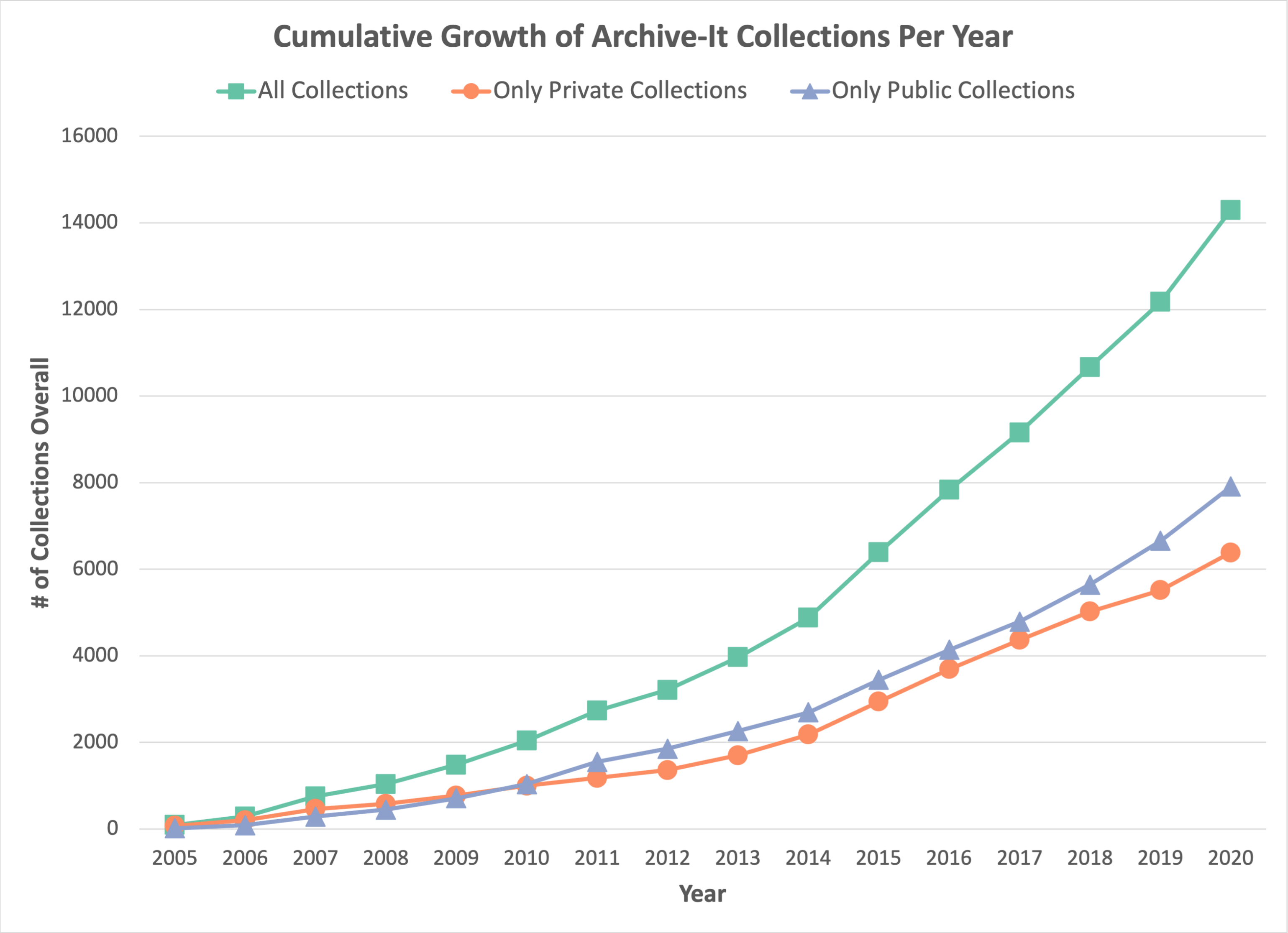
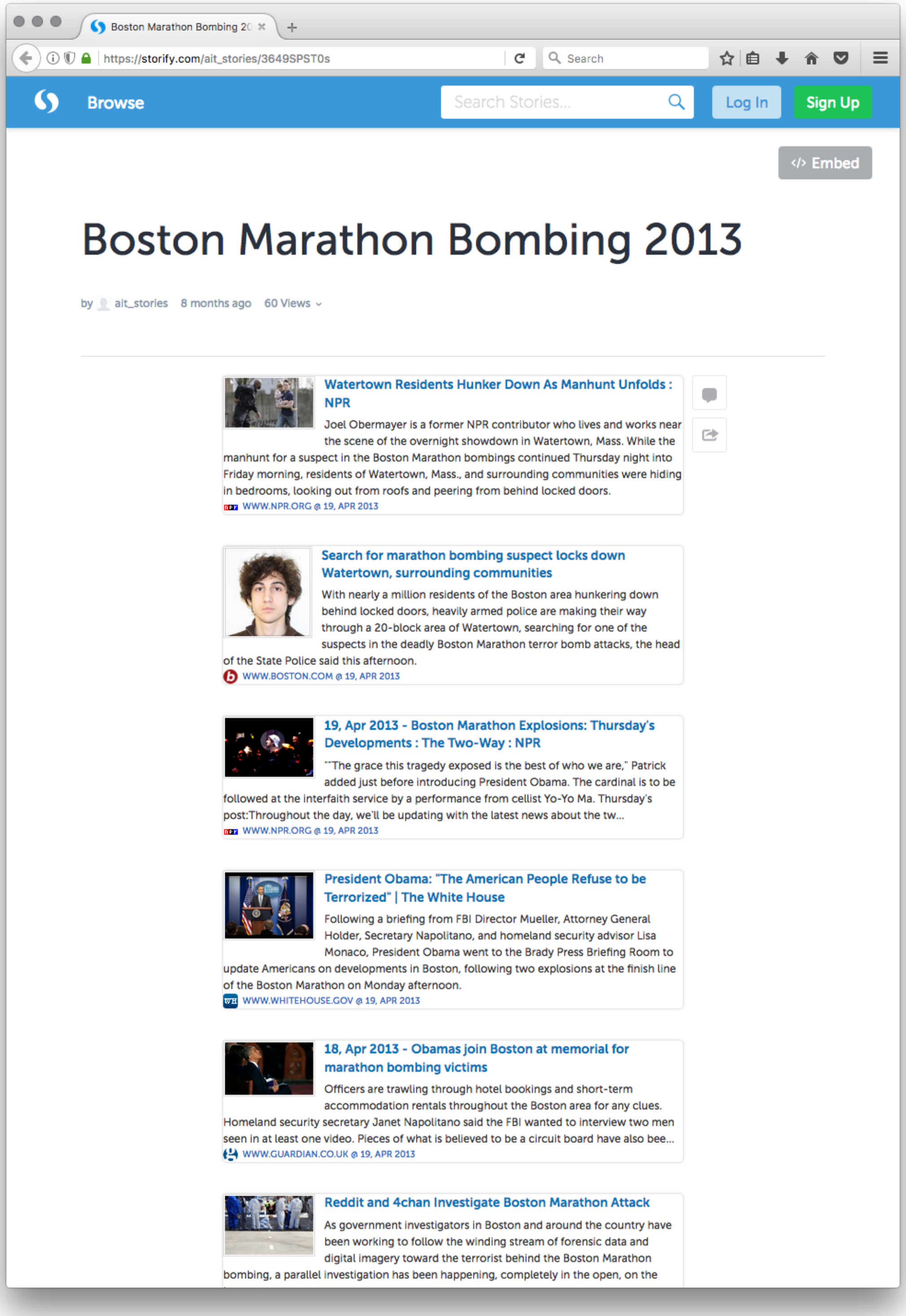


Fig. 6. The cumulative growth of Archive-It collections through 2020. This data was gathered in May 2021 by Hypercane, a tool to be covered in Chapter 4.5.


Boston Marathon Bombing 20
https://storify.com/ait_stories/3649SPST0s
Search
Browse
Search Stories...
Log In
Sign Up
</> Embed

Boston Marathon Bombing 2013

by ait_stories 8 months ago 60 Views

Watertown Residents Hunker Down As Manhunt Unfolds : NPR

Joel Obermayer is a former NPR contributor who lives and works near the scene of the overnight showdown in Watertown, Mass. While the manhunt for a suspect in the Boston Marathon bombings continued Thursday night into Friday morning, residents of Watertown, Mass., and surrounding communities were hiding in bedrooms, looking out from roofs and peering from behind locked doors.

WWW.NPR.ORG @ 19, APR 2013

Search for marathon bombing suspect locks down Watertown, surrounding communities

With nearly a million residents of the Boston area hunkering down behind locked doors, heavily armed police are making their way through a 20-block area of Watertown, searching for one of the suspects in the deadly Boston Marathon terror bomb attacks, the head of the State Police said this afternoon.

WWW.BOSTON.COM @ 19, APR 2013

19, Apr 2013 - Boston Marathon Explosions: Thursday's Developments : The Two-Way : NPR

""The grace this tragedy exposed is the best of who we are," Patrick added just before introducing President Obama. The cardinal is to be followed at the interfaith service by a performance from cellist Yo-Yo Ma. Thursday's post:Throughout the day, we'll be updating with the latest news about the tw...

WWW.NPR.ORG @ 19, APR 2013

President Obama: "The American People Refuse to be Terrorized" | The White House

Following a briefing from FBI Director Mueller, Attorney General Holder, Secretary Napolitano, and homeland security advisor Lisa Monaco, President Obama went to the Brady Press Briefing Room to update Americans on developments in Boston, following two explosions at the finish line of the Boston Marathon on Monday afternoon.

WWW.WHITEHOUSE.GOV @ 19, APR 2013

18, Apr 2013 - Obamas join Boston at memorial for marathon bombing victims

Officers are trawling through hotel bookings and short-term accommodation rentals throughout the Boston area for any clues. Homeland security secretary Janet Napolitano said the FBI wanted to interview two men seen in at least one video. Pieces of what is believed to be a circuit board have also bee...

WWW.GUARDIAN.CO.UK @ 19, APR 2013

Reddit and 4chan Investigate Boston Marathon Attack

As government investigators in Boston and around the country have been working to follow the winding stream of forensic data and digital imagery toward the terrorist behind the Boston Marathon bombing, a parallel investigation has been happening, completely in the open, on the

Fig. 7. Archive-It Collection 3649, *Boston Marathon Bombing 2013*, visualized using the now defunct storytelling service Storify as part AlNoamany's work [11]. (Screenshot taken in 2017.)

Elbert wants to help people like Natasha find and understand his collections so that they can make these decisions. Ling wants to understand the colossal collection that she has inherited. None can rely upon existing metadata because it is inconsistently applied, and human review of the mementos within a collection is an expensive proposition.

To assist web archive collection users, we desire to automatically generate summaries of collections, similar to Luhn's pioneering work on the automatic generation of abstracts [206]. In our case, our summaries do not attempt to provide a text abstract but a visualization of the collection. We aim to surface mementos that represent the collection, **exemplar mementos**, rather than just sentences, as elements for our summarization. To borrow concepts from information foraging theory [266], we aspire to maximize the value of the knowledge gained from our collection summary while minimizing the cost of interacting with the collection. Users rely upon textual and visual clues to determine if a resource fits their needs. A resource with textual and visual clues indicating that it meets a user's needs is said to have good **information scent** [56]. Thus, if the collection meets a particular user's needs, we want to ensure that the exemplar mementos chosen for our summarization and their visualizations have good information scent.

Once we have the best mementos representing a collection, we apply **social media storytelling** to provide a summary of the collection, as pioneered with web archives by AlNoamany [11]. Social media storytelling is a method developed by the now-defunct service Storify [289]. It provides an interface that is familiar to most users of the web. With this form of storytelling, the platform renders individual web pages as **surrogates**. Each surrogate provides a summary of a given page. Groups of surrogates summarize a topic. We would represent a collection by a single **story** that summarizes that collection using these surrogates, creating a summary of summaries. For example, Figure 7 shows how a user reduced Archive-It Collection 3649 about the Boston Marathon Bombing to a set of exemplar mementos which are themselves visualized as a set of surrogates, providing a story from this collection. Researchers should be able to glance at these stories and understand the collection that they represent.

Figure 8 displays our model of five processes that one must address to conduct storytelling with corpora. According to AlNoamany et al. [9] popular stories contain a median of 28 documents. Thus, for a collection of more than 28 mementos, an automated process or a human must **select exemplars** from the larger corpus. Additionally, to increase their information scent, authors often augment their stories with metadata, so we **generate story metadata** for the story. Story metadata can include general metadata from the overall collection (e.g., title, collection creator, creation date). It can also refer to metadata generated from the exemplars as a group (e.g., overall best story image, number of exemplars, most frequent terms). Storytelling relies heavily

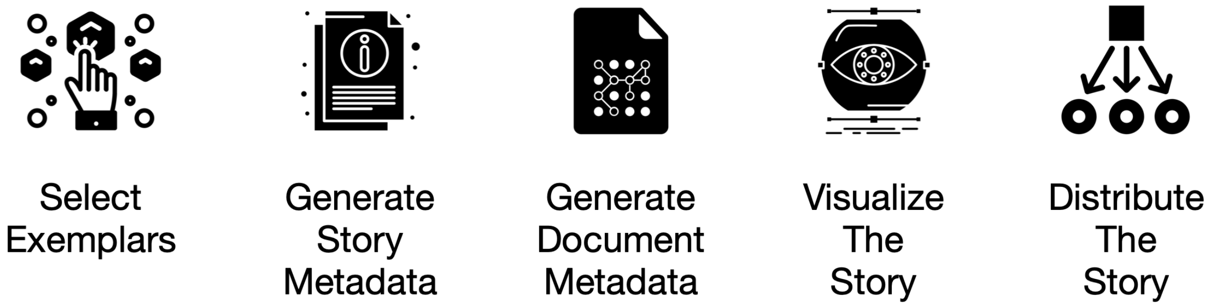
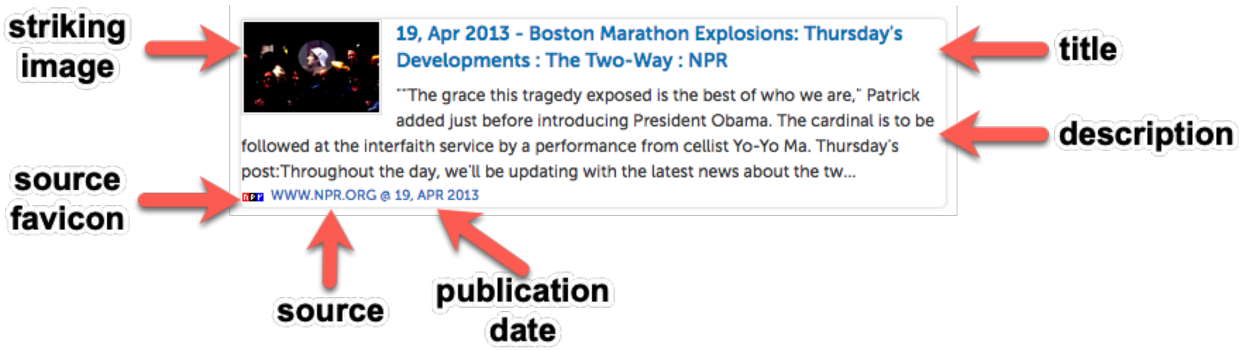


Fig. 8. Our five processes for storytelling with a corpus.


19, Apr 2013 - Boston Marathon Explosions: Thursday's Developments : The Two-Way : NPR

"'The grace this tragedy exposed is the best of who we are," Patrick added just before introducing President Obama. The cardinal is to be followed at the interfaith service by a performance from cellist Yo-Yo Ma. Thursday's post:Throughout the day, we'll be updating with the latest news about the tw...

WWW.NPR.ORG @ 19, APR 2013

Fig. 9. A social card from Figure 7, annotated to indicate the metadata concepts applied to create it.

upon surrogates to convey the topic of the story, so we must **generate document metadata** (e.g., title, description, publisher) for each exemplar. Humans can supply this document metadata, or an application might automatically generate it. For individual documents, we can visualize the document metadata as a surrogate, like the **social card** shown in Figure 9 that contains a title, description, publication date, source, source favicon, and striking image. With the story metadata and individual exemplars' metadata, we can then **visualize the story**. For an overall story, we must visualize not just individual surrogates, but story metadata as well. For example, in Figure 7 we see that the author opted to visualize the story by placing the story metadata (i.e., title) at the top of the page. The story author then formatted the exemplars' metadata as a series of social cards rendering the metadata of titles, descriptions, striking images, sources, and dates. Once visualized, a service or individual must **distribute the story** so that others can see the visualization and get the gist of what is in the collection.

Web archive collections have multiple dimensions with which we can tell stories: pages and times. These dimensions affect how we select our exemplars. For example, someone interested in summarizing the collection as a whole is interested in a story with **sliding pages and sliding times (SPST)**. With SPST stories, any page from any time covered by the collection is a candidate exemplar. We can also fix a particular aspect of the collection. For example, if we are only interested in a given page, then our process for selecting exemplars is reduced to that one page but may change in time. From here, we can tell a story with a **fixed page, but sliding times (FPST)**. Likewise, we can reduce our candidate exemplars to a specific time, but allow any page, telling a story with a **sliding page, but fixed time (SPFT)**. Table 1 lists our personas, their information needs, their roles, their collection understanding needs, and the story type that will meet their needs. Let us see how each of our storytelling processes and these storytelling types help the personas we have introduced.

Natasha is trying to choose between collections. For each collection that she is considering, someone must have created a story by following these processes. Exemplars save her from having to review thousands of documents unnecessarily. Her exemplars come from the entire set of pages and times; thus, an SPST story will meet her goals. Story metadata helps her better understand the collection and the exemplars as a whole. Document metadata helps her understand each exemplar. She has many collections to review, and a visualization of these components with good information scent helps her quickly understand each collection with a glance, saving her time.

Rustam wants to understand how a specific page changes over time. By selecting his page, he has reduced his exemplars. Now they can only slide in time, meaning an FPST story will satisfy his needs. Even though Rustam knows which page he is interested in, he needs to select exemplars

TABLE 1

The personas in this dissertation, their information needs, their roles, their collection understanding needs, and the story types that will meet their needs.

Persona	**Natasha**	**Rustam**	**Olayinka**	**Elbert**	**Ling**
Information need	Needs to quickly compare collections	Wants to follow a source over time	Wants to understand a given time from different sources	Wants to promote collections and help visitors understand them	Wants to understand a collection that they inherited
Role	visitor	visitor	visitor	archivist	archivist
Understanding Needs	overall collection	aspect of collection	aspect of collection	overall collection	overall collection
Story type	sliding page, sliding time	fixed page, sliding time	sliding page, fixed time	sliding page, sliding time	sliding page, sliding time

because he does not know which mementos exist. He needs story metadata so he can refer back to the collection from which the exemplars originated. Rustam needs a visualization of this story to understand all of this metadata. Finally, he wants to distribute his story in a manner that helps other researchers review and repeat his work.

Olayinka wants to understand the topics reflected in a set of news articles during a specific time. Her exemplars only come from a specific time period, but any page is acceptable, allowing an SPFT story to meet her goals. She is only interested in the zeitgeist for that particular moment in time. Even though she has chosen a time period, she still needs to select exemplars because there are still many pages from that time. Story metadata helps her summarize this timespan, and document metadata helps her understand the individual articles. Without a visualization, all of this metadata is just text, and Olayinka needs to share this visualization with others to support her work.

Elbert selects exemplars so he can promote his collection. He is not interested in a particular aspect but wants to promote the collection as a whole; therefore, an SPST story will work best. Elbert generates story and document metadata so that visitors will understand which collection these exemplars represent. He wants a visualization that has good information scent because it will help visitors like Natasha, and this metadata will help him get there. Finally, he needs to distribute this story far, so visitors know that the collection exists.

Ling is also an archivist, but she wants to understand the collection she inherited; thus, an SPST story will work best for her. She needs to select exemplars to have a small number of documents to review to understand the collection. Ling needs story metadata to get an overview of the collection from many angles. Ling requires document metadata so she can understand her exemplars. To make sense of all of this, she also needs a visualization. Where Elbert and Rustam are interested in sharing their stories with visitors, Ling is interested in sharing hers with other archivists, so she also needs to be able to distribute it.

To help people like Natasha, Rustam, Olayinka, Elbert, and Ling, we evaluated how to automate and improve each of the processes outlined in Figure 8. This work is the focus on the Dark and Stormy Archives (DSA) Project [148], an effort to combine storytelling with web archive collections. In addition to the research cited by this dissertation, we developed the tools of Hypercane [144, 145, 146, 149, 157, 166], MementoEmbed [156], and Raintale [156] not only to aid our research but also for the community. Here we discuss four research questions, how they related to our characters, and these tools. Figure 10 details how each process maps to a research question and a tool from the Dark and Stormy Archives Project, leading us to the research questions of this dissertation:

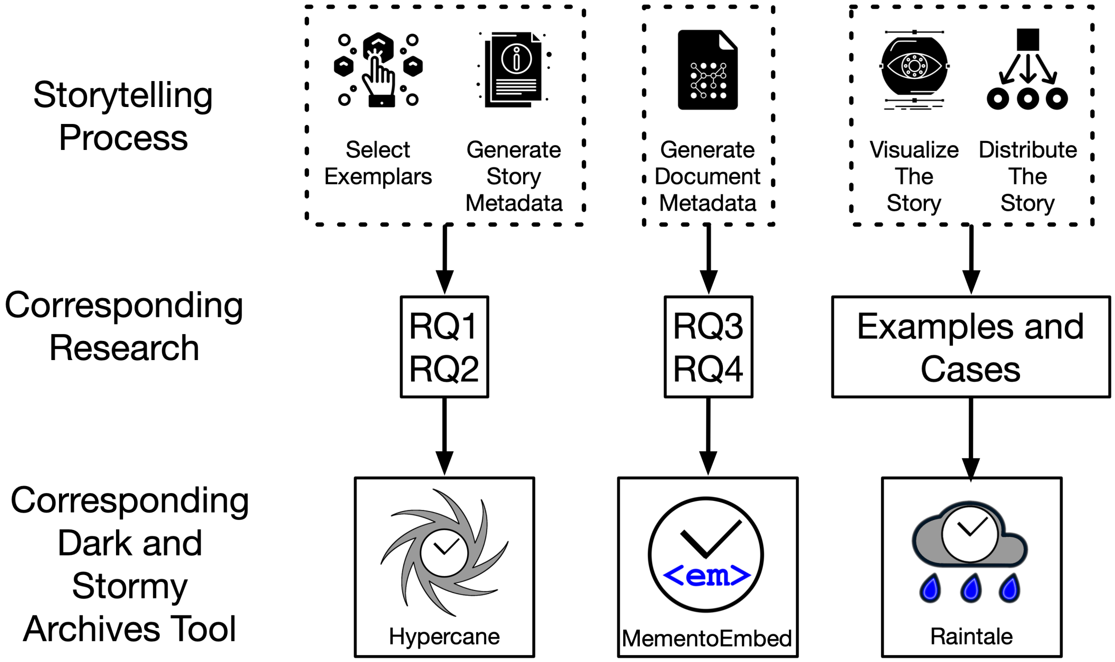


Fig. 10. The five processes for storytelling with a corpus mapped to the research in this dissertation.

RQ1: What types of web archive collections exist and what structural features do they have?

We address this question in Chapter 4.1. Before analyzing any document text, we analyze the different structural features that describe Archive-It collections. We can understand the collections through these structural features in terms of curatorial involvement, crawling behavior, and seed analysis. Structural features help us understand how to select exemplars and generate story metadata.

In Chapter 4.1 we conduct a review of the same 3,382 Archive-It collections and divide them into four different semantic categories. In addition, we discover that we can apply the structural features to predict these categories via Random Forest. By understanding structural features, we understand the nature of each collection both in terms of page and time, helping us select exemplars. Furthermore, understanding the type of collection can help us create useful story metadata. Prior work by AlNoamany et al. [11] only evaluated one type of collection stored in Archive-It.

RQ2: Which approaches work best for selecting exemplars from web archive collections?

We have many possible methods for selecting exemplars. As defined by Lu [205], **intelligent**

sampling aims “to reduce sampling cost without loss of accuracy by smartly designing sample sizes.” Inspired by this definition, we apply the term intelligent sampling in this dissertation to refer to sampling mementos from a collection to select exemplars. As part of intelligent sampling, we filter the collection. Web archive collections can contain mementos that are not helpful to our eventual summaries. Because of scale and automated crawling, web archive collections often contain mementos that are not representative of the collection and are off-topic. In Chapter 4.2 we analyze different text similarity methods such as Simhash [54] and Sørensen-Dice [69, 290] and determine that differences in word count performs best for identifying off-topic mementos in a collection. Filtering off-topic mementos from consideration helps us reduce our set of exemplars for further analysis.

Existing corpora summarization techniques [12, 195, 286, 338] group documents by their commonalities and then select the highest scoring document from each group. How does one group documents? Structural features, like the number of captures of a given page over time, help here. Other concepts, like topic modeling, help us understand the nature of a collection during a specific time period. Scoring is key to selecting exemplars from these groups. In Chapter 4.3 we develop a general model of summarizing web archive collections inspired by this research. We develop four algorithms and determine their effectiveness by comparing them against search engine results in Chapter 4.4. We demonstrate that intelligent sampling surfaces mementos from the collection that a user might otherwise miss if they do not examine thousands of search engine results.

We further realize this general model in the DSA tool Hypercane [166], introduced in Chapter 4.5. Both RQ1 and RQ2 feature work supporting the left column from Figure 10.

RQ3: What surrogates work best for understanding groups of mementos?

Search engines and social media platforms often visualize document metadata as a surrogate. By analyzing surrogates, we indirectly determine which document metadata is best for helping users to understand documents. Several options for surrogates and metadata exist. Are social cards the best surrogate type? **Browser thumbnails** provide a screenshot of the web page in a browser. Web archives have applied them to visualize pages in the past. Are they better than social cards? In this chapter, we have seen screenshots of different Archive-It collections, each applying the Archive-It surrogate type. Which surrogate type best assists understanding?

In Chapter 3.3 we discuss past studies of different surrogate types and their effectiveness for evaluating search results. In our case, we are trying to solve a slightly different problem. Where search result surrogates are trying to answer the question of “Does this meet my information need?”, our story surrogates are trying to answer “What does the underlying page contain?” We augment this work in Chapter 5.1 with a discussion of the surrogates produced by different web platforms and how they are not suitable for mementos. We present a user study in Chapter 5.2 where we show that 120 Mechanical Turkers perform best with social cards.

RQ4: What methods that automate the creation of surrogates produce results that best match humans' behavior?

Once we established that social cards were the best surrogate for understanding, we analyzed their creation. Platforms that support social cards require that the author insert metadata in their page to control the rendering of those social cards. In Chapter 5.3, we analyze the metadata of 277,724 mementos and determine that 40% cannot generate a full card because they lack this metadata. Facebook released the first set of these metadata standards in 2010, and many mementos were captured before that year. The Internet Archive has roughly 150 billion pages captured before 2010. However, we find that once social card metadata existed, its adoption rate grew every year. With this motivation in mind, we analyze existing web page author behavior to determine how best to automatically generate social cards. Chapter 5.3 briefly discusses web page authors' behavior with respect to the text in their cards before focusing on the more complex problem of how to automatically select the striking image representing a document for its card. From this work, we address the process of generating document metadata.

We apply the knowledge from RQ3 and RQ4 in the DSA tool MementoEmbed [151, 156], introduced in Chapter 5.4, which produces social cards and other types of surrogates for mementos. Addressing RQ3 and RQ4 closes out our work on the middle column of Figure 10.

Chapter 6 covers the right column of Figure 10. Here we introduce the tool Raintale [143, 156] which allows a user to visualize a story by rendering its story and document metadata through templates. With Raintale introduced, we demonstrate the stories generated by the DSA tools, but not just for mementos from Archive-It. We cover collections summarized by Archive-It, the National Library of Australia, and groups of mementos from the Internet Archive. We provide example stories that help Natasha, Rustam, Olayinka, Elbert, and Ling.

Our main contribution is the exploration of the automated summarization of collections by first selecting exemplars and then rendering the result as a social media story. Chapter 2 introduces the background concepts necessary to further understand the topics in this dissertation. In Chapter 3 we will show how AlNoamany's proof-of-concept implementation satsifies the five processes from Figure 8. Chapter 3 talks about the work similar to our research questions and how our work differs. Chapter 4 focuses on the processes of selecting exemplars and generating collection metadata. Chapter 5 focuses on generating document metadata. Chapter 6 ties all of these processes together in a set of examples and use cases. With Chapter 7 we discuss potential future applications of this work and Chapter 8 helps summarize this journey.

CHAPTER 2

BACKGROUND

Web archive collections are complex entities. We provide a brief overview of some base concepts before discussing the problems of collection understanding in web archive collections. We will first cover the web and web archives to explain how preserving the web is different from other forms of digital preservation. Because it contains more than 14,000 collections, we will focus on Archive-It as the target of our research. In Chapter 6, however, we will show that our storytelling processes can work not just for Archive-It, but other web archives as well.

We will cover web page surrogates and highlight why they are essential to our storytelling visualization. From there, we will discuss other forms of storytelling and fit ours into that spectrum. The other part of our solution involves conveying aboutness for a collection. We will demonstrate both human-led and machine methods of conveying aboutness for documents and corpora.

Finally, we will introduce some models of understanding. These models will be relevant to the development of our solution and will also provide us some ideas for its evaluation. We finish by discussing user studies and how they map our work in later chapters.

2.1 THE WEB, HTTP, AND HTML

It would be irresponsible to discuss web archiving without first introducing the concepts of the World Wide Web, hereafter referred to as just **the web**. As part of the 1945 article “As We May Think” Vannevar Bush [47] expressed many ideas later incorporated into the web. He mentioned how people might read documents on screens and then easily reach documents on connected topics. The concept of the interconnected documents has been brought to life numerous times during the 20th Century as HyperCard [76, 169, 253], Intermedia [101], NLS [82, 329], FRESS [28], and a variety of other systems [26, 27, 31, 38, 103, 113, 127, 215, 220, 252, 272, 278, 330]. The current web [33] created in 1989 by Sir Tim Berners Lee is the latest, and most widely used, realization of Bush’s concept.

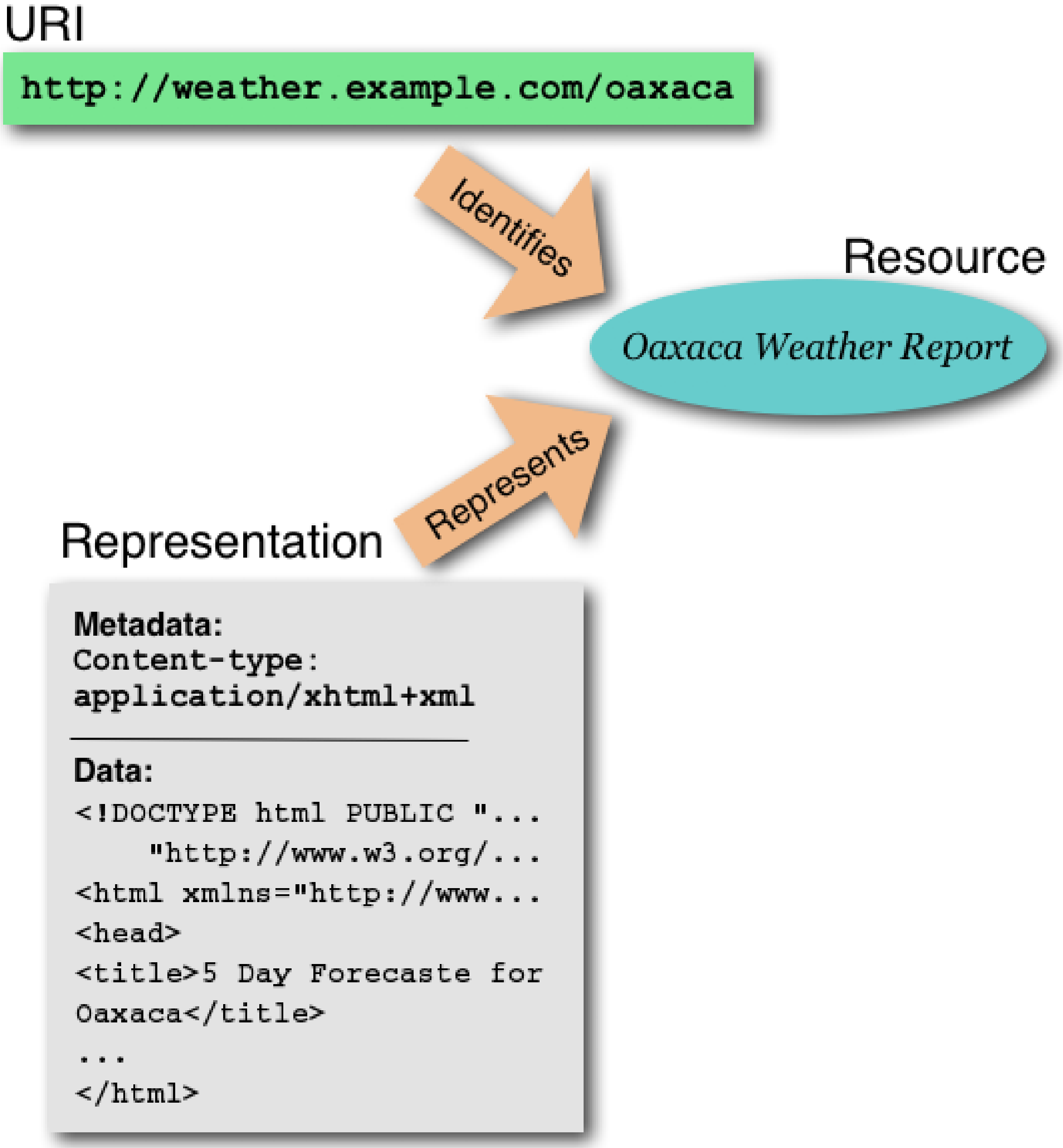
Interlinked collections of documents often existed on a single machine or within an organization’s network. The web is different because it exists on the much larger **Internet**. Colloquially, many users use the term Internet to mean the web, but we will keep them separate in this work for clarity. The Internet consists of a set of **protocols** established so that different systems can interact effectively. A protocol is analogous to a standardized language that allows one system to request information from a second system and the second system to respond to the first. Systems on the web use the **Hypertext Transfer Protocol** [90, 91, 92, 93, 94, 95], or **HTTP**, to exchange information. **HTTPS** [275], a more secure form of HTTP, is often used to protect information. Curators build most web archive collections by acquiring documents using HTTP or HTTPS.

On the Internet, a numeric Internet Protocol address [248], or **IP address** identifies each connected system. One system can serve many different protocols. The system does so by listening on different channels, or **ports**, for each protocol. Thus a system with an IP address of 10.168.1.220 can listen for connections on port 80 for HTTP and port 21 for the file transfer protocol (FTP) [270] at the same time. Because IP addresses are difficult to remember, most systems have a registered **hostname** (e.g., `xenon.cs.odu.edu`) that users can employ to connect to a given system. Each hostname contains a **domain name** corresponding to the organization that owns that system (e.g., cs.odu.edu corresponds to the Old Dominion University Computer Science Department). The Domain Name System (DNS) [224, 225, 226] uses specific systems to map hostnames to IP addresses to provide this convenience. We use hostnames, protocols, and sometimes ports to identify items on the web.

Items on the web do not need to be static documents. Consider a weather report. A physical paper weather report cannot change. Users of a paper weather report only have access to the information from the time of publication. If a storm changes direction, users of an out-of-date paper weather report will not know to update their plans and may get rained on. In the electronic world, a resource can change based on updated information. A web weather report can provide updated and current information. A web weather report's users can find shelter, knowing that the storm is coming. Thus, the web consists of interconnected documents and interconnected resources, where some are static, but all have the potential to be dynamic.

Figure 11 shows the relationship between the main concepts that allow the web to function. Each resource on the web has a Uniform Resource Identifier [32, 33, 34], or **URI**. Each URI identifies a single resource, but a single resource can be identified by multiple URIs. If one user wants to help her friend have current weather information, she can share the URI for the weather report. Her friend can then access that report using the URI. A subset of URIs, Uniform Resource Locators, or URLs, is commonly used to share web resources. We use the more general term, URI, in this work.

URIs can take different forms. A URI begins with a **scheme** which corresponds to the protocol used to retrieve the resource. A colon and two slashes follow the scheme. Following the scheme is the hostname of the system servicing the request. The path to the specific resource follows the hostname. If no path component exists in the URI, it is assumed to be the character '/,' referring to



Metadata:

```
Content-type:
application/xhtml+xml
```

Data:

```
<!DOCTYPE html PUBLIC "...
    "http://www.w3.org/...
<html xmlns="http://www...
<head>
<title>5 Day Forecaste for
Oaxaca</title>
...
</html>
```

Fig. 11. The Oaxaca Weather Report example borrowed from Berners-Lee, et al. [33] showing the relationship between a URI, the resource it identifies, and a representation of that resource.

TABLE 2

Some example HTTP methods.

Method	Description
GET	Request a resource
POST	Update a resource with given information
PUT	Create a new resource with the given information
DELETE	Remove a resource
HEAD	Request a response identical to GET, but without the content, just the metadata
OPTIONS	Request the options for communicating with the resource
TRACE	Test the connection

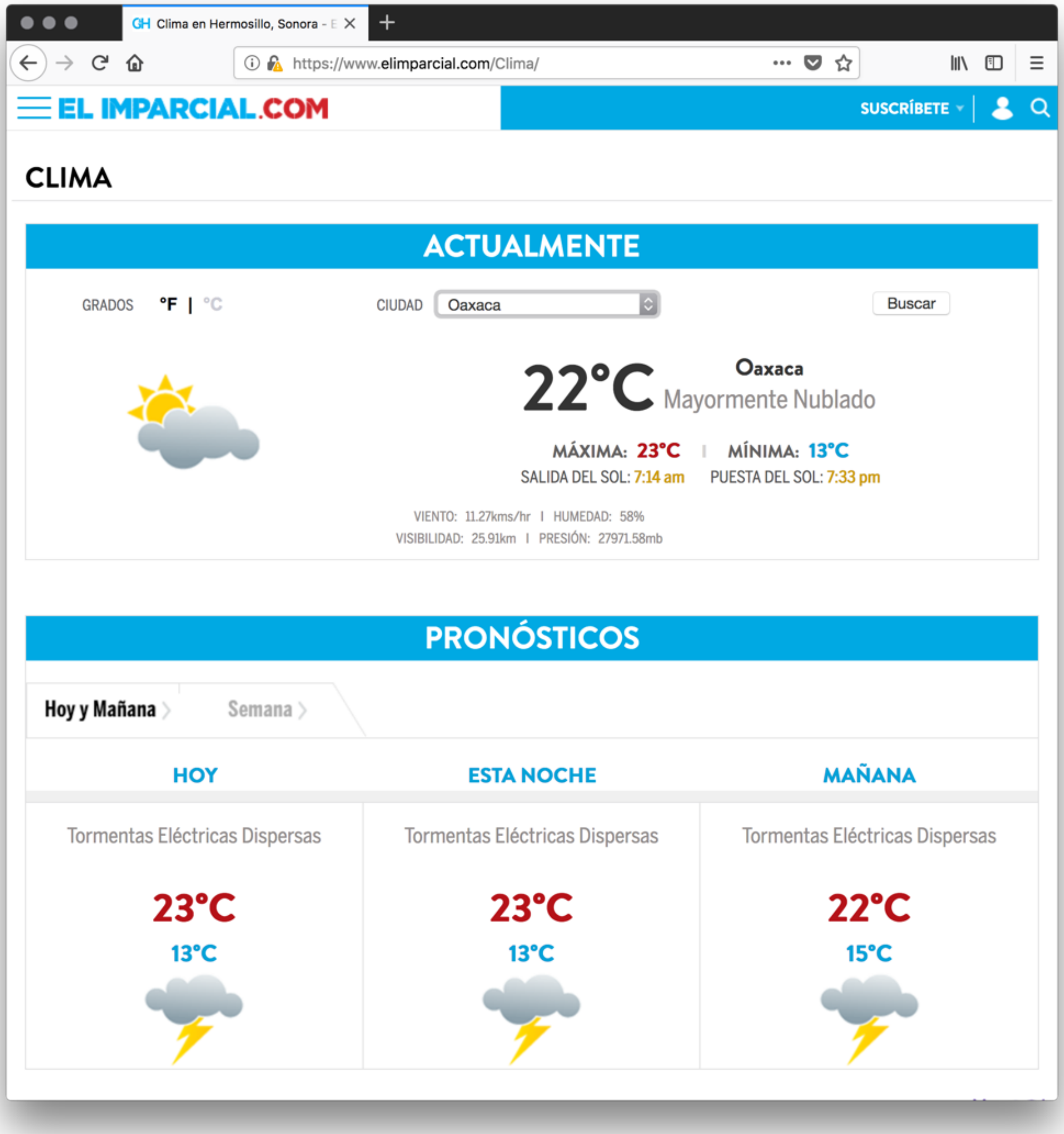
the default document available on the server. In the URI from Figure 11, the scheme is `http`, the domain name is `weather.example.com`, and the path is `/oaxaca`. By employing this URI, a user is asking a system with the hostname `weather.example.com` for the resource named `/oaxaca` using protocol HTTP.

Each resource can have one or more **representations**. For example, a weather report for Oaxaca, Mexico, will likely be published in Spanish. Non-Spanish speakers may not understand words like *lluvia* and *viento*, indicating wind and rain. Thus, the storm may catch these users by surprise. With multiple representations, the same resource can service multiple languages, so there can be an English version of the same weather report at the same URI. Each URI is **dereferenced** to acquire a **representation** of the resource at the current time and in the preferred language, file format, and compression.

Users typically experience web resources in a software application known as a **user-agent** – colloquially referred to as a “browser” because that is what most humans use for interactive sessions – as shown in Figure 12. It is important to note that though other software programs also explore the web, the browser is the most frequently employed class of software for experiencing the web. To understand better how the web delivers representations to both users and machines, we must provide an overview of HTTP.

2.1.1 HTTP AND CONTENT NEGOTIATION

The HyperText Transfer Protocol (HTTP) is a stateless protocol [90, 91, 92, 93, 94, 95]. Two basic roles exist in HTTP. The **client** initiates a **request** as a standard message to a **server**, who


GH Clima en Hermosillo, Sonora - E
https://www.elimparcial.com/Clima/
EL IMPARCIAL.COM
SUSCRÍBETE
Buscar

CLIMA

ACTUALMENTE

GRADOS °F | °C

CIUDAD Oaxaca

22°C Oaxaca

Mayormente Nublado

MÁXIMA: 23°C | MÍNIMA: 13°C

SALIDA DEL SOL: 7:14 am PUESTA DEL SOL: 7:33 pm

VIENTO: 11.27kms/hr I HUMEDAD: 58%

VISIBILIDAD: 25.91km I PRESIÓN: 27971.58mb

PRONÓSTICOS

Hoy y Mañana › Semana ›

HOY	ESTA NOCHE	MAÑANA
Tormentas Eléctricas Dispersas	Tormentas Eléctricas Dispersas	Tormentas Eléctricas Dispersas
23°C	23°C	22°C
13°C	13°C	15°C

Fig. 12. A screenshot of the Oaxaca Weather Report published by *El Impartial*, as rendered by the Firefox Browser. (Screenshot was taken in 2019.)

```
GET /publications/ HTTP/1.1
Host: www.shawnmjones.org
User-Agent: curl/7.54.0
Accept: text/html
Accept-Language: en-US
```

Fig. 13. An example HTTP request for `https://www.shawnmjones.org/publications/`.

TABLE 3

Some example HTTP request headers.

Header	**Description**
Accept	Allows the client to request certain formats
Accept-Language	Allows the client to request a representation in a particular language
Content-Length	Allows the client to indicate the length of the content sent in the case of POST or PUT requests
Content-Type	Allows the client to specify the format of the content sent in the case of POST or PUT requests
Cookie	Allows the client to inform the server of past state information
Host	Allows the client to specify the domain name part of the URI, may also include a port number
Referer	Allows the client to inform the server of the URI of the previous page that linked to the URI in the request
User-Agent	Allows the client to identify its software to the server, sometimes used to influence the representation returned

```
HTTP/1.1 200 OK
Date: Mon, 10 Sep 2018 15:47:54 GMT
Server: Apache/2.4.6 (CentOS) OpenSSL/1.0.2k-fips PHP/5.4.16
Content-Length: 15210
Content-Type: text/html; charset=utf-8
Vary: Accept-Encoding
Last-Modified: Thu, 06 Sep 2018 04:47:25 GMT
ETag: "3b6a-5752c965bd5c8"

... 15210 bytes of content follow and are omitted for brevity...
```

Fig. 14. An example HTTP response.

then provides another standard message back the the client as a **response**. The request indicates the action to take on a resource. The response contains information about how the server handled the request.

An HTTP request for `https://www.shawnmjones.org/publications/` is shown in Figure 13. The **status line** is the first line in each HTTP message. In this request, the status line consists of the HTTP method `GET`, the URI path `/publications/`, and the HTTP version that the client would like to use for this exchange, HTTP/1.1. Essentially, the client is asking the server to return the resource at the URI path `/publications/` using HTTP version 1.1. Common HTTP methods are shown in Table 2.

After the status line, the request includes headers. Some headers are mandatory, like `Host`, which the server needs to locate the appropriate resource. Additional headers are informational, like `User-Agent`, which tells the server which client software issued the request. Moreover, still others, like `Accept`, are used by the client to influence the server to return a specific representation of a resource. In this case, `Accept` indicates that the client desires a representation in the HTML file format. The `Accept-Language` header indicates that the client would like an English (`en-US`) representation. Some HTTP request headers are shown in Table 3.

The server provides an HTTP response like the one shown in Figure 14. The status line of the HTTP response starts with the version of HTTP used in the response, followed by a status code. The value of 200 indicates a successful response to the client's request. Some HTTP status codes are shown in Table 4.

TABLE 4

Some example HTTP status codes.

Status Code	Description
200	The request was successfully fulfilled
301	The resource has permanently moved and can be found at the URI in the Location response header
307	The resource is temporarily located at another URI listed in the Location response header
400	The client generated an improper request, either syntactically or logically, and the server cannot fulfill it due to the client's error
401	The client needs to authenticate to access this resource
403	The client is not permitted to access this resource
404	The requested resource cannot be found on this server
405	The method used in the request is not supported for this resource
408	The server timed out while waiting for the client to finish the request
429	The server has received too many requests from this client and the client needs to wait before making another request
500	The server experienced an error while fulfilling the client's request
501	The server does not support the method in the request
503	The server is not ready to handle the request
505	The server does not support the version of HTTP requested by the client

TABLE 5

Some example HTTP response headers.

Response Header	**Description**
Cache-Control	Used by the server to send caching information to the client
Content-Encoding	The encoding used on the data, often for compression
Content-Language	The language of the response body
Content-Length	The length, in bytes, of the content
Content-Type	The format of the content in the response
Date	The date and time of the response
ETag	An identifier for the version of the resource to be used to inform caches
Location	Used in 300-series responses to indicate the URI that the client should use for this redirect
Link	Informs the client of other resources and their relationships to this resource
Set-Cookie	Allows the server to give the client some stateful information for future requests
Server	Identifies the server's software
Vary	Used to indicate to the client which request headers can be used to request a different representation of this resource
WWW-Authenticate	In 401 responses, tells the client which authentication scheme to use

TABLE 6

HTTP request and response headers used for content negotiation.

Request Header	**Corresponding Response Header**	**Description**	**Standards**
Accept	Content-Type	Format of the representation	RFC 7231, RFC 2616 (historical)
Accept-Charset	Content-Type (via the charset attribute)	Character set of the representation	RFC 7231, RFC 2616 (historical)
Accept-Datetime	Memento-Datetime	Client can influence the time period of the representation	RFC 7089
Accept-Encoding	Content-Encoding	Medium, typically compression, that the entity has been processed with and also what will need to be done by the user agent to return the entity to its original form	RFC 7231, RFC 2616 (historical)
Accept-Language	Content-Language	Language of the representation	RFC 7231, RFC 2616 (historical)

Table 5 displays some examples of response headers that follow the status line. Always included in an HTTP response is the Date header, indicating the date and time of the response. In the response shown in Figure 14, the server provides an informational Server header indicating its software name and version. A final blank line indicates the end of the headers. If present in the response, the actual content of the web resource follows this final blank line.

When the content is present, response headers instruct the client on how to handle the content returned. The Content-Length header tells the client how many bytes to read after the blank line. The Content-Type header indicates the format of those bytes so that clients can process the resulting content effectively. In Figure 14, the format is HTML (text/html) and it is encoded with the UTF-8 character set. The URI and these headers are used to produce this representation.

Browsers can avoid resending unnecessary requests through caching and instead display the last version of a resource that they received. Different headers exist to assist with caching decisions. The Last-Modified header contains the date and time that the resource was last changed, and the ETag header provides an identifier for the specific version of this resource. The Vary header informs the client which headers combined with a URI match a given representation. A client can use these headers for its own caching algorithms to make decisions to save time and network bandwidth.

Clients can influence the representation on behalf of the user through a process known as **content negotiation** [118]. A user can request a representation in a particular language, format, character set, encoding, and time period through content negotiation. Web archive collections consist of captures (mementos) of resources as they existed at some point in the past; thus, the ability to negotiate in time is key to accessing these mementos. In Figure 13, a user requested an English representation using the Accept-Language header. Content negotiation is the unsung hero of HTTP interactions, often working so well that many do not realize that it is present. In Figure 14, the server respondes with a Vary: Accept-Encoding, indicating that it can negotiate in that dimension. Most servers support content negotiation with Accept-Encoding and Content-Encoding, but since the browsers hand this automatically, users do not notice. Table 6 shows the dimensions of content negotiation supported by HTTP.

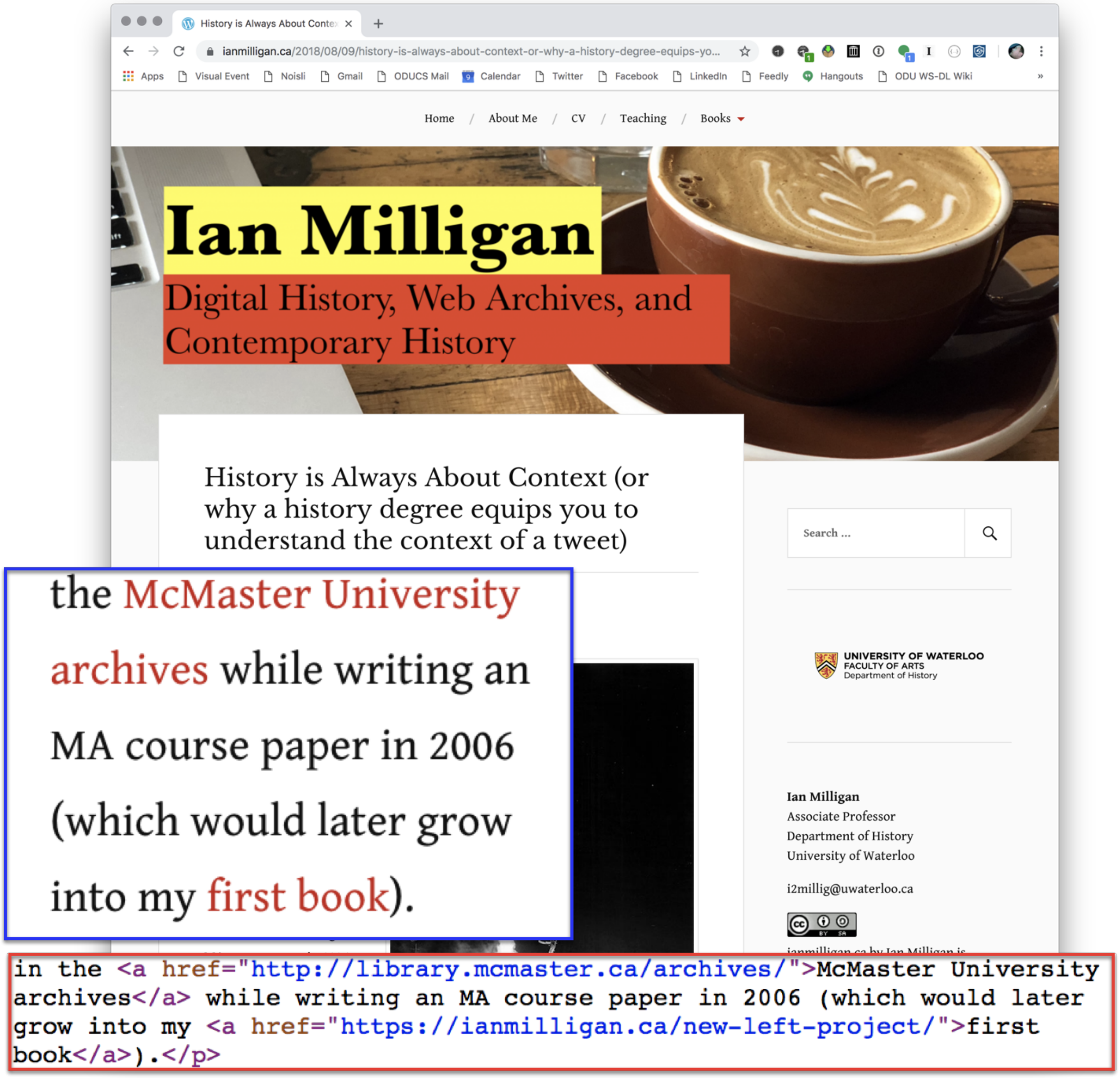
```
<!DOCTYPE html>
<html>
  <head>
    <title>Simple document</title>
  </head>
  <body>
    <p>This is some text in a paragraph that will be seen
       by viewers.</p>
  </body>
</html>
```

Fig. 15. A simple HTML document showing the use of opening and closing tags. Example from Wikibooks [319].

TABLE 7

HTML tags and their attributes used to linking to external resources

HTML tag citing external resource	Attribute used to specify URI	Description
`img`	`src`	Instructs the browser to download the image at the URI specified by src and render it at this place in the HTML page.
`link`	`href`	Used primarily for stylesheets, but may also inform machine clients of relationships to other documents.
`video`	`src`	Instructs the browser to download a video at the URI specified by src and render it at this place in the HTML page.
`script`	`src`	Instructs the URI to download a script specified by src and then execute it for use with this web page.



Ian Milligan

Digital History, Web Archives, and Contemporary History

History is Always About Context (or why a history degree equips you to understand the context of a tweet)

the McMaster University archives while writing an MA course paper in 2006 (which would later grow into my first book).

Ian Milligan
Associate Professor
Department of History
University of Waterloo

i2millig@uwaterloo.ca

```
in the <a href="http://library.mcmaster.ca/archives/">McMaster University
archives</a> while writing an MA course paper in 2006 (which would later
grow into my <a href="https://ianmilligan.ca/new-left-project/">first
book</a>).</p>
```

Fig. 16. A screenshot of a blog post by Ian Milligan with a section magnified (outlined in blue) showing the corresponding HTML (outlined in red) emphasizing the syntax for creating links using the a (for anchor) tag. Screenshot is from Milligan [222] taken on September 9, 2018.

2.1.2 HTML AND CRAWLERS

Most of the content in web archive collections consists of representations in **Hypertext Markup Language (HTML)** [89]. This file format consists of **elements**, or **tags**, that specify the structure and appearance of the representation, often referred to as a web page. Figure 15 shows a simple HTML document consisting of `html`, `head`, `title`, `body`, and `p` (for paragraph) tags. HTML documents support the concept of a **Document Object Model** [313] whereby text is included within tags, and some tags can include other tags (e.g., `title` can only exist in `head`).

The anchor tag, `a`, is key to giving the web its interconnected nature. With anchor tags, web page authors can specify that a section of the page links to another resource on the web. Consider the example in Figure 16, where historian Ian Milligan wishes to cite McMaster University archives in an article and supplies the following HTML to do so:

```
<a href="http://library.mcmaster.ca/archives/">
McMaster University archives</a>
```

A user can link to a URI by specifying it in the `href` attribute of the anchor tag. The web browser then renders the text *McMaster University archives* differently from the surrounding text to indicate that it is a link.

Many other tags exist, and we will not repeat the HTML standard here, but it is important to note that other forms of linking exist in HTML documents, listed in Table 7. Styles controlling the appearance of a webpage can be imported from other sources using the `link` tag. Likewise, by using the `script` tag, we can import external scripts that provide interactive elements for web pages. Multimedia can be included using the `img` and `video` tags. These additional resources mean that each web page is constructed of not one but many resources. The existence of these additional resources has implications for web archiving because the software that captures a web page needs to capture all of these resources as well.

A **crawler** [52] finds and captures these resources. The crawler starts with a single web page, finds its linked resources, downloads them, and continues the cycle for each linked resource. As shown in Figure 17, the crawler downloads web pages via a multi-threaded downloader that extracts URIs from the links in these pages. The crawler then submits these URIs to a queue. Once in the queue, the crawler schedules the URIs for further submission to the downloader, and the cycle starts again. The system stores the text and metadata about the capture of each page. Crawlers are what search engines use to discover pages. Search engines build indices on the latest version of the page that they have crawled, whereas web archives store all previous versions of

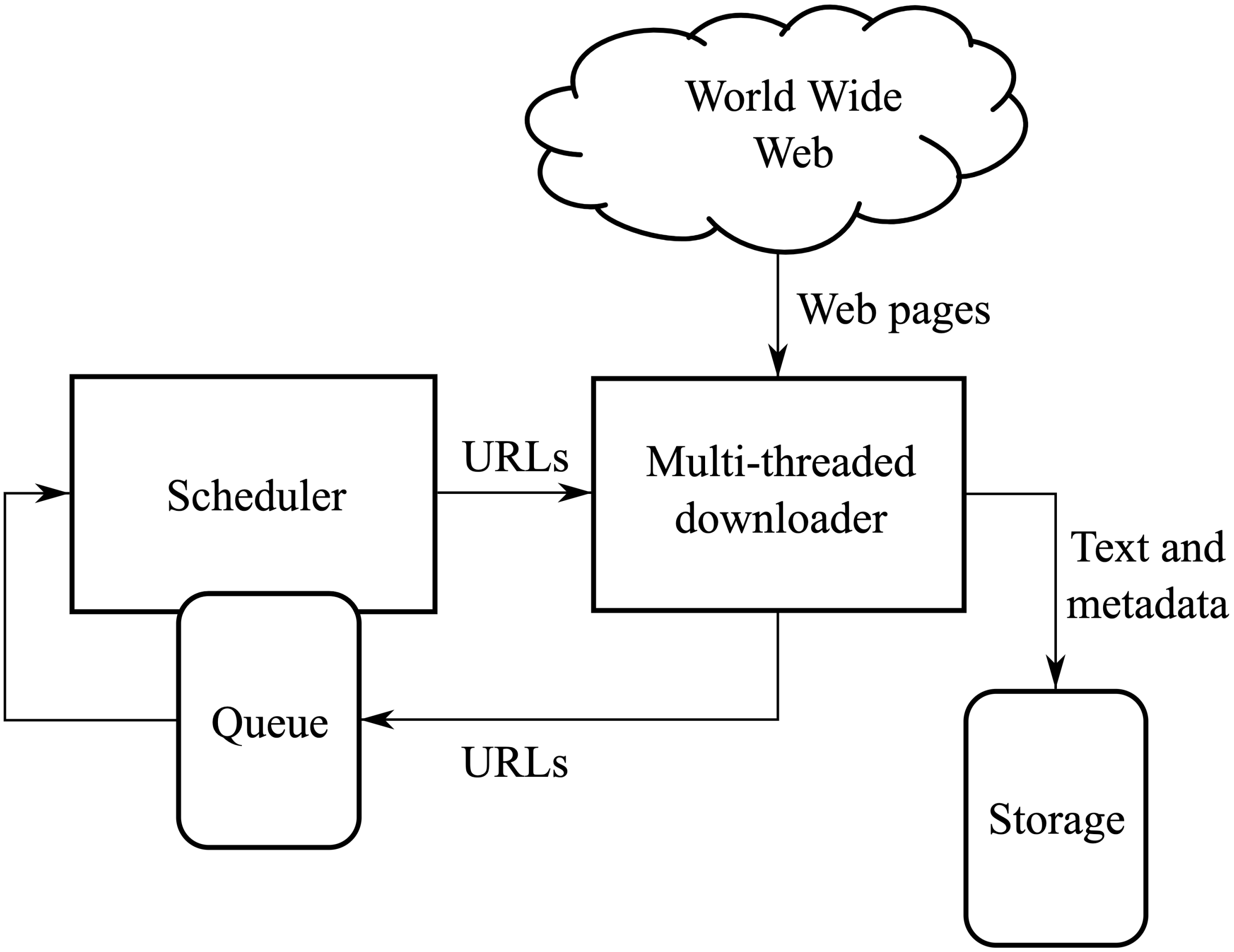


Fig. 17. The architecture of a simple web crawler. Image in the public domain, courtesy of Castillo [52] and Wikimedia Commons [61].

a web that they have crawled. Crawlers make search engines and web archives possible because they can capture many web pages by merely following the linked resources, just like humans do.

2.2 WEB ARCHIVES AND MEMENTO

We have covered the basics of HTML and how HTTP enables its links to interconnect. We also discussed crawlers, which search engines primarily use to index the web. Web pages link to other pages, but sometimes the linked page disappears. Web users often refer to this phenomena as **broken links**, or **link rot** [183]. Because web archives allow users to revisit web pages that have disappeared, addressing link rot is one of the use cases for web archives. Here we will describe how to preserve web pages, who preserves them, and how we access these archived web pages,

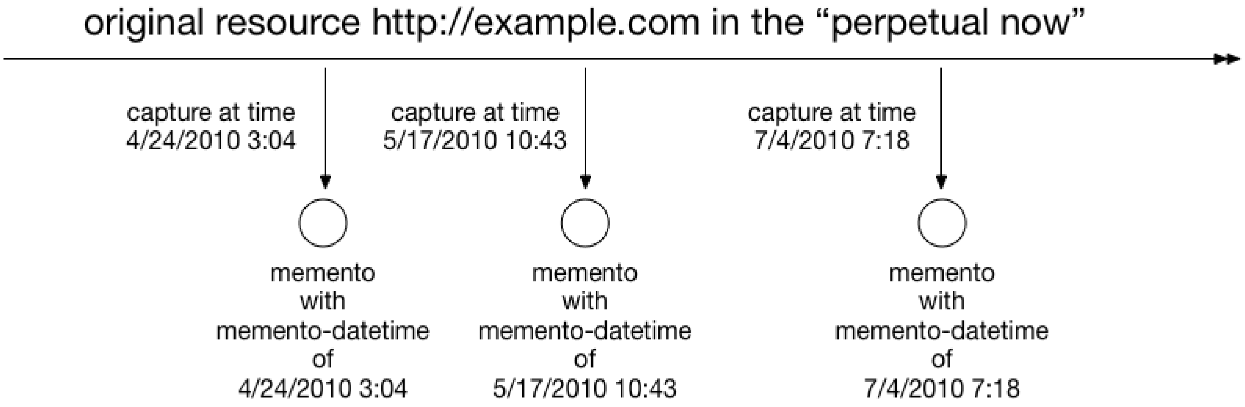

original resource http://example.com in the "perpetual now"
capture at time 4/24/2010 3:04
capture at time 5/17/2010 10:43
capture at time 7/4/2010 7:18
memento with memento-datetime of 4/24/2010 3:04
memento with memento-datetime of 5/17/2010 10:43
memento with memento-datetime of 7/4/2010 7:18

Fig. 18. Mementos are created by captures from the original resource at specific times, which become their memento-datetimes.

also known as **mementos**.

The web exists in the "perpetual now." Every resource returns a representation of its state at this moment. A news reporter can update a story tomorrow, meaning that its content will not be the same as it is today. A meteorologist can update a weather report with current information within the next few minutes. These changes are referred to as **content drift**, and though a normal function of the web, they also create problems for people trying to visit a resource from a specific point in time [154, 162]. Mementos are observations of web resources at specific points in time [159]. Because URIs point to resources and mementos are resources themselves, we formally refer to the resource to be archived as an **original resource**. Figure 18 displays the relationship between the original resource in the perpetual now and the captures made at some point in the past, which become the observations we view in web archives as mementos. Their time of capture is their **memento-datetime**.

Archiving an original resource typically involves capturing a web page via an HTTP request, storing the HTML, scanning the HTML for links to embedded content, and then issuing requests for this content so the archive can save it as well. Where search engines wish to index the web by using the content of web pages, web archives seek to preserve what they find. Thus, web archives record this data in a format named **WARC** (for Web ARChive) [128].

Web archives engage in two primary activities: **capture** (also called **recording**) and **playback** (also called **replay**). Capture involves crawling web pages and saving the crawl sessions as WARC files. Playback allows users to view these captures using a web browser as one would for any other web page. Remember that mementos have a temporal component not present in other web resources. For example, the web archive captured http://www.example.com at the three datetimes shown in Figure 18. There are now three records in the WARC file for http://www.example.com, requiring the playback system to ask the user which date and time they wish to view for http://www.example.com. The playback system must take these two pieces of information, the original resource (http://www.example.com) and the memento-datetime (May 17, 2010, at 10:43:48 GMT), and search the WARC file to find the appropriate record to present to the user in a browser. To quickly find this information, playback engines require a **CDX** file [123] which is an index of one or more WARC files. Given an original resource and a datetime, web archives use CDX files to locate the WARC file and the position in that WARC containing the desired memento.

Many different web archives employ these techniques [63, 88, 106, 213, 299]. The Internet Archive[1] is the largest web archive. They use the Heritrix [193] and Brozzler [126] crawlers to capture mementos and save them to WARCs. The Wayback playback system then makes these mementos available to users. Other web archives include archive.is[2] (also known as Archive.today), Web Cite[3], the UK Web Archive[4], the Portuguese Web Archive[5], the Icelandic Web Archive[6], the Web Archive Austria[7], and Perma.cc[8]. There are also tools for users and organizations to create their own web personal archives, such as WARCreate[9], WAIL[10], and Conifer[11] (formerly Webrecorder.io).

The existence of multiple web archives necessitated the development of standards. Most web archives use WARC and CDX files, but these are not accessible via HTTP. Their playback representation is accessible via HTTP, but we needed a standard method of access to facilitate the development of tools that enable the discovery of mementos and the aggregation of resources. The Memento Protocol [155, 312] provides this access.

Which memento for http://example.com is closest to April 25, 2010? What mementos exist for http://example.com? The Memento Protocol allows one to make HTTP requests in order to answer these questions. Without it, web archives would need to share WARC and CDX files among them. The Memento Protocol formalizes several constructs from web archiving and

[1] https://archive.org/
[2] http://archive.is/
[3] http://webcitation.org/
[4] https://www.webarchive.org.uk/
[5] https://arquivo.pt/
[6] https://vefsafn.is/
[7] https://webarchiv.onb.ac.at/
[8] https://perma.cc/
[9] https://warcreate.com/
[10] https://machawk1.github.io/wail/
[11] https://conifer.rhizome.org/

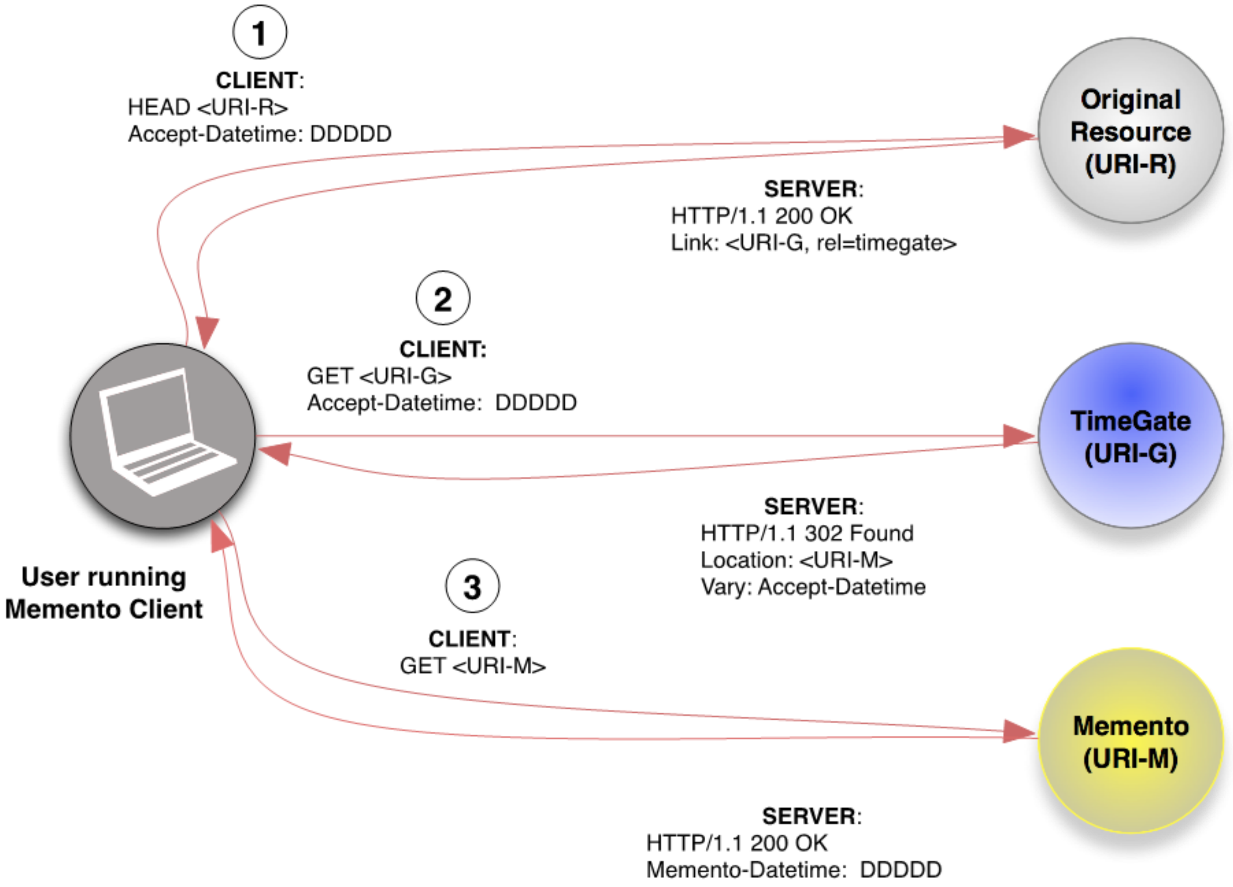


Fig. 19. This diagram by Jones [139] describes content negotiation in the dimension of datetime with a Memento TimeGate.

introduces new resources to make this kind of access possible.

TimeGates are Memento Protocol resources that allow users to find the memento closest to a given datetime. A TimeGate resource exists for each original resource in a web archive. Figure 19 shows an exchange of HTTP messages using a TimeGate. A client would contact the original resource (**URI-R**), which would then respond with the URI of the TimeGate (**URI-G**) using the `Link` response header. To acquire the memento closest to a given datetime, the client submits their desired datetime as part of an HTTP request to the URI-G using the `Accept-Datetime` header. The TimeGate then responds with a 302 redirect containing the appropriate memento URI (**URI-M**) in the `Location` header of the response. The client then issues a last request to this URI-M and processes it as any other web page. These three steps are called **datetime negotiation**, content negotiation in the dimension of datetime.

Ideally, a response from an original resource would contain the URI-G in the `Link` header of

URI-R: `https://ianmilligan.ca/`
URI-G: `https://web.archive.org/web/https://ianmilligan.ca/`
URI-T: `https://web.archive.org/web/timemap/link/https://ianmilligan.ca/`

Fig. 20. If a client cannot discover a URI-G or URI-T in the response headers for a URI-R, then it can typically generate these values for a given web archive by prepending a known prefix to the URI-R. These examples apply to the Internet Archive (archive.org).

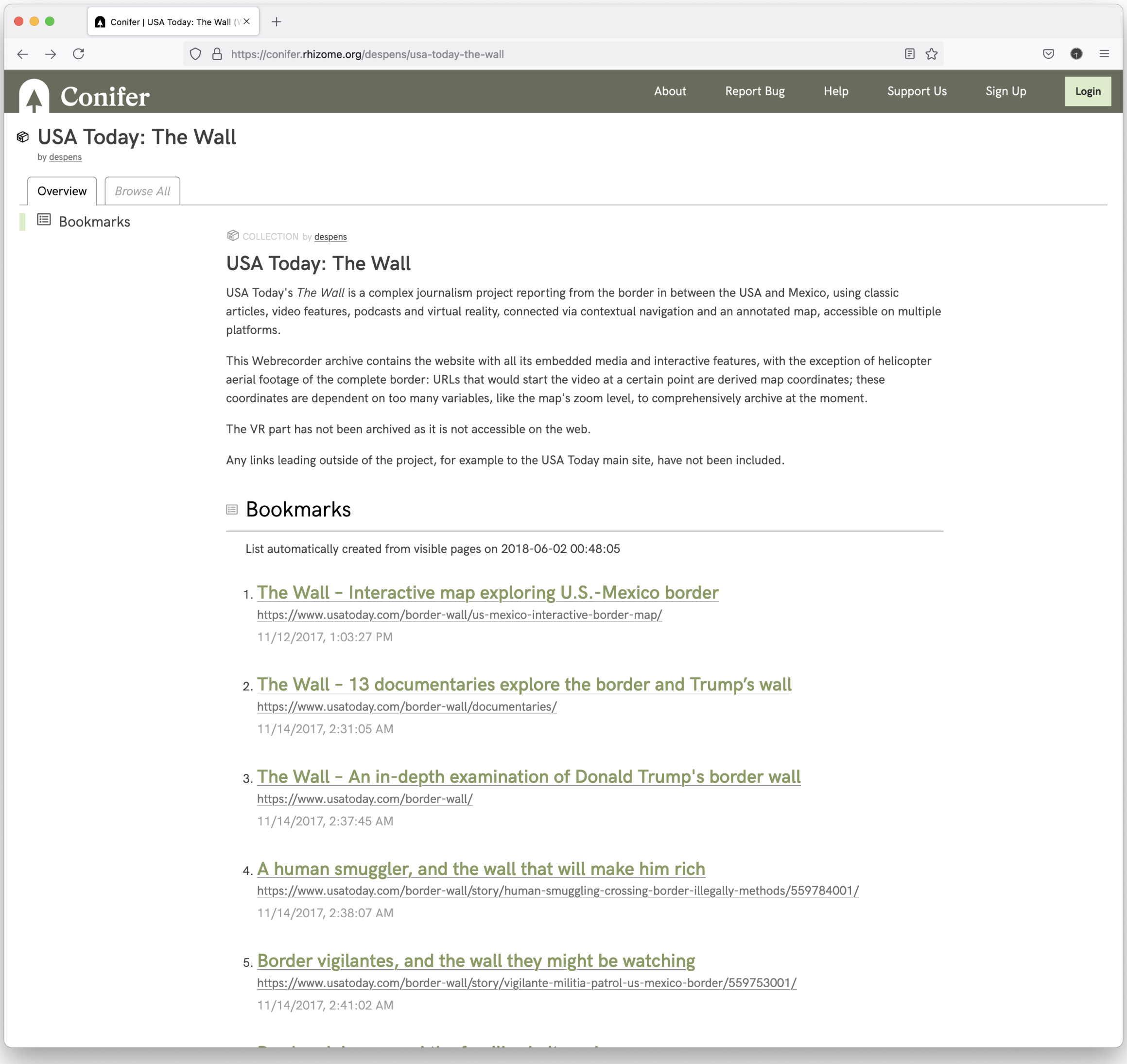
its response. In reality, not all original resources support this feature, and thus clients must know the URI-G beforehand. Most web archives allow a user to construct a URI-G by prepending a web archive's URI-G prefix to the URI-R; thus, clients can discover a URI-G this way. See Figure 20 for an example.

TimeMaps are web resources that contain a list of all mementos for a given original resource. A client can discover the URI of a TimeMap (**URI-T**) in the Link header of an original resource's HTTP response. If the client cannot find the URI-T this way, they can construct it by prepending a web archive's URI-T prefix to the URI-R and discover it that way. See Figure 20 for an example.

This way, the Memento protocol allows users to employ TimeGates to browse the web as if it were a given datetime in the past, allowing them to visit documents about specific historical events or permitting them to avoid spoilers in TV shows [159]. The Memento Protocol also allows researchers to get a list of mementos for a given resource using TimeMaps. TimeMaps can be used to study the changes in a resource over time [162] or the availability of current resources [183].

2.3 ARCHIVE-IT

We now know how the web works, how authors construct web pages, how they are archived, and how we can access those archived pages. Many web archives are collections of pages saved as part of a massive general crawl or saved one at a time by users. As noted in Chapter 1, users create themed web archive collections. Unlike general web crawls, these themed collections consist of original resources (and hence mementos) that are all semantically linked. Collection capabilities exist at different web archives, such as Conifer (Figure 21), the Croatian Web Archive (Figure 22), the Library of Congress (Figure 23), and Trove (Figure 24). We have designed our solution to work with any web archive collection, and are are currently piloting our work with collections from the National Library of Australia [234]. Archive-It is the most commonly used web archive collection creation tool used by organizations, and thus it is the focus of this dissertation. Archive-It provides a user-friendly interface allowing a curator to select original resources as seeds (Figure


Conifer | USA Today: The Wall
https://conifer.rhizome.org/despens/usa-today-the-wall
Conifer
About
Report Bug
Help
Support Us
Sign Up
Login
Overview
Browse All
Bookmarks

USA Today: The Wall

by despens

COLLECTION by despens

USA Today: The Wall

USA Today's *The Wall* is a complex journalism project reporting from the border in between the USA and Mexico, using classic articles, video features, podcasts and virtual reality, connected via contextual navigation and an annotated map, accessible on multiple platforms.

This Webrecorder archive contains the website with all its embedded media and interactive features, with the exception of helicopter aerial footage of the complete border: URLs that would start the video at a certain point are derived map coordinates; these coordinates are dependent on too many variables, like the map's zoom level, to comprehensively archive at the moment.

The VR part has not been archived as it is not accessible on the web.

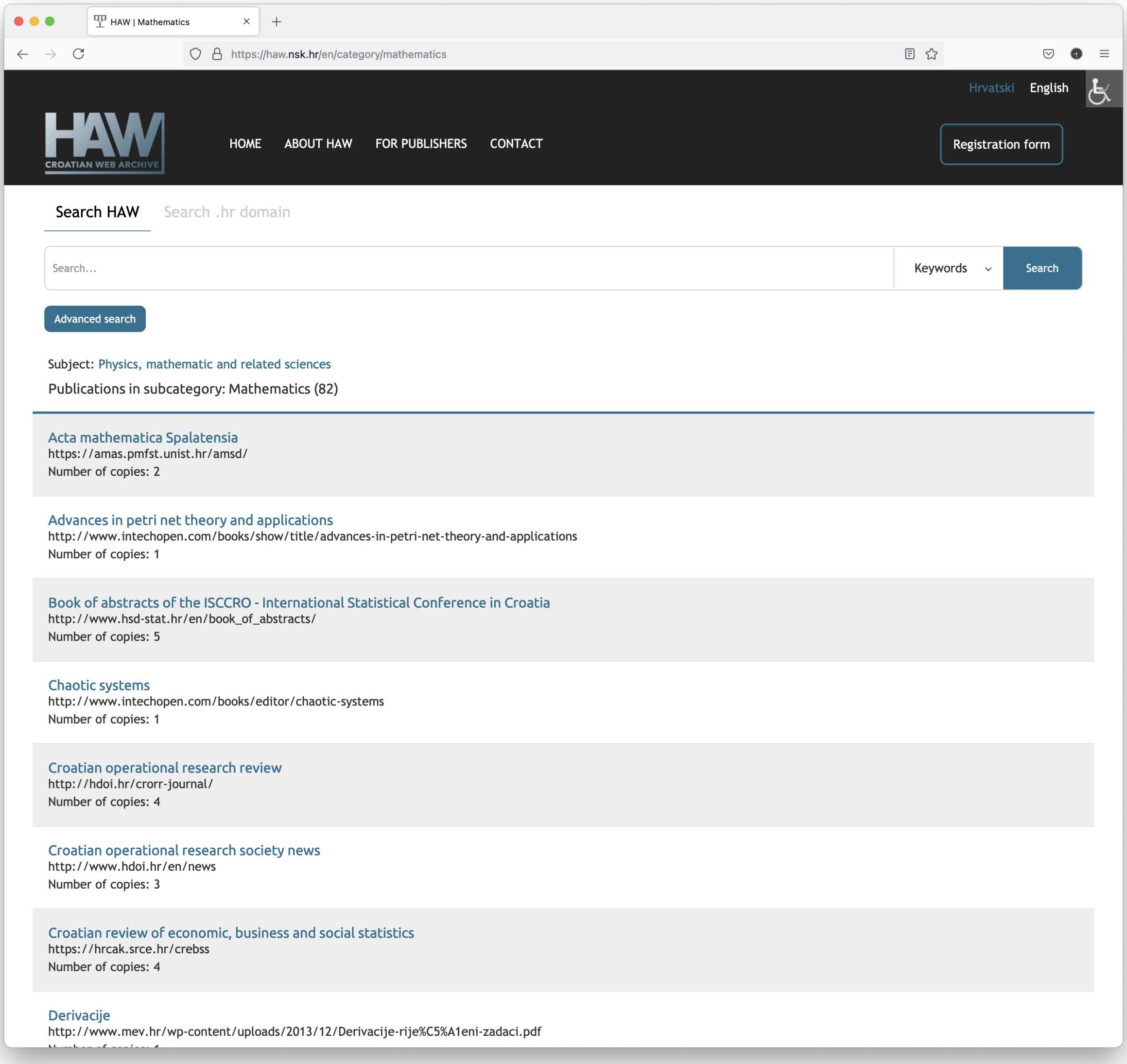
Any links leading outside of the project, for example to the USA Today main site, have not been included.

Bookmarks

List automatically created from visible pages on 2018-06-02 00:48:05

1. The Wall – Interactive map exploring U.S.-Mexico border
   https://www.usatoday.com/border-wall/us-mexico-interactive-border-map/
   11/12/2017, 1:03:27 PM
2. The Wall – 13 documentaries explore the border and Trump's wall
   https://www.usatoday.com/border-wall/documentaries/
   11/14/2017, 2:31:05 AM
3. The Wall – An in-depth examination of Donald Trump's border wall
   https://www.usatoday.com/border-wall/
   11/14/2017, 2:37:45 AM
4. A human smuggler, and the wall that will make him rich
   https://www.usatoday.com/border-wall/story/human-smuggling-crossing-border-illegally-methods/559784001/
   11/14/2017, 2:38:07 AM
5. Border vigilantes, and the wall they might be watching
   https://www.usatoday.com/border-wall/story/vigilante-militia-patrol-us-mexico-border/559753001/
   11/14/2017, 2:41:02 AM

Fig. 21. Examples of different web archive collection capabilities:
Conifer collection *USA Today: The Wall*.
URI: `https://conifer.rhizome.org/despens/usa-today-the-wall`



Subject: Physics, mathematic and related sciences

Publications in subcategory: Mathematics (82)

Acta mathematica Spalatensia
https://amas.pmfst.unist.hr/amsd/
Number of copies: 2

Advances in petri net theory and applications
http://www.intechopen.com/books/show/title/advances-in-petri-net-theory-and-applications
Number of copies: 1

Book of abstracts of the ISCCRO - International Statistical Conference in Croatia
http://www.hsd-stat.hr/en/book_of_abstracts/
Number of copies: 5

Chaotic systems
http://www.intechopen.com/books/editor/chaotic-systems
Number of copies: 1

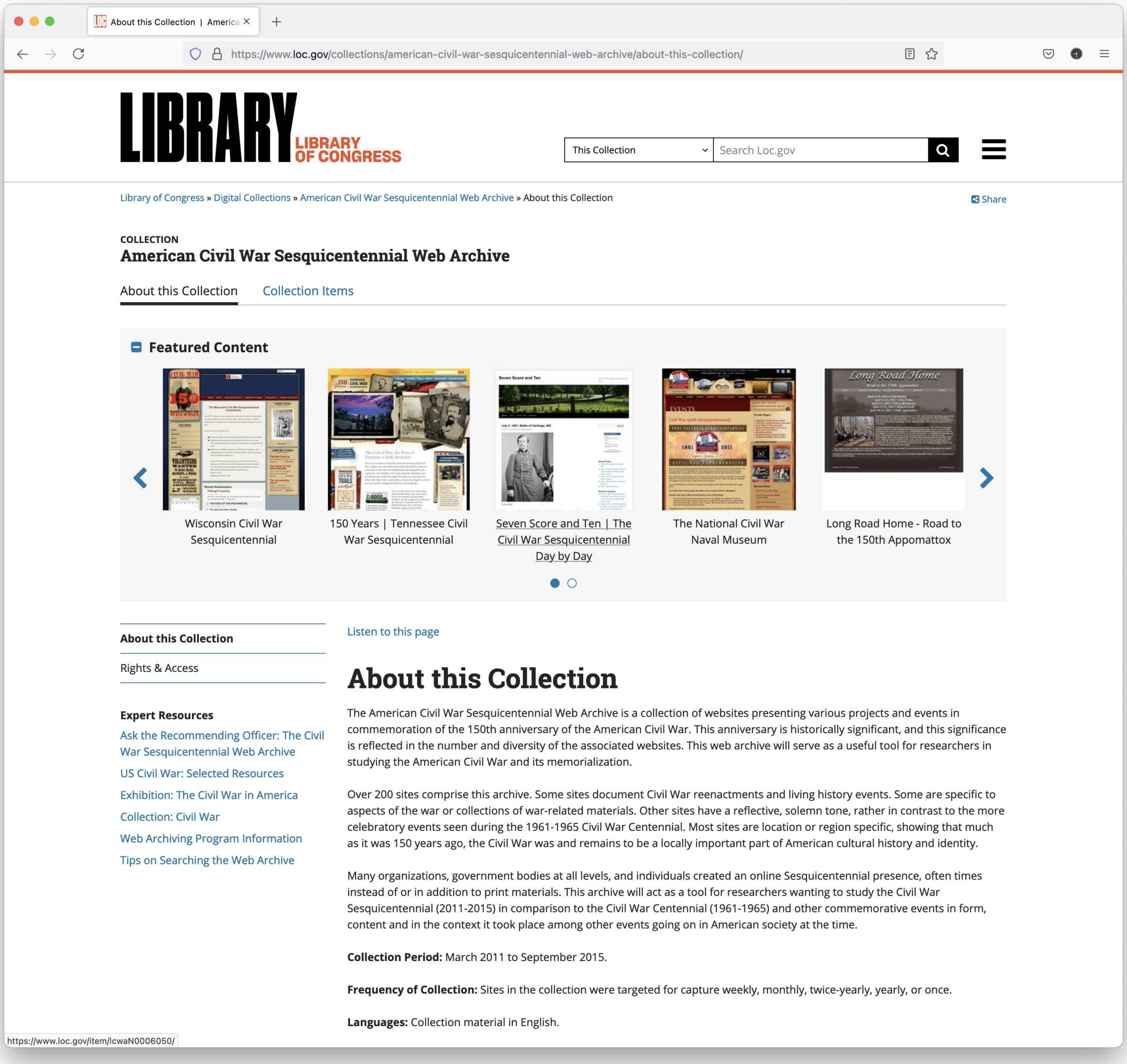
Croatian operational research review
http://hdoi.hr/crorr-journal/
Number of copies: 4

Croatian operational research society news
http://www.hdoi.hr/en/news
Number of copies: 3

Croatian review of economic, business and social statistics
https://hrcak.srce.hr/crebss
Number of copies: 4

Derivacije
http://www.mev.hr/wp-content/uploads/2013/12/Derivacije-rije%C5%A1eni-zadaci.pdf
Number of copies: 1

Fig. 22. Examples of different web archive collection capabilities: Croatian Web Archive (HAW) collection *Mathematics*. URI: `https://haw.nsk.hr/en/category/mathematics`


About this Collection | America
https://www.loc.gov/collections/american-civil-war-sesquicentennial-web-archive/about-this-collection/
LIBRARY
LIBRARY OF CONGRESS
This Collection
Search Loc.gov
Share
Listen to this page
https://www.loc.gov/item/lcwaN0006050/

Library of Congress » Digital Collections » American Civil War Sesquicentennial Web Archive » About this Collection

COLLECTION

American Civil War Sesquicentennial Web Archive

About this Collection | Collection Items

Featured Content

Wisconsin Civil War Sesquicentennial

150 Years | Tennessee Civil War Sesquicentennial

Seven Score and Ten | The Civil War Sesquicentennial Day by Day

The National Civil War Naval Museum

Long Road Home - Road to the 150th Appomattox

About this Collection

Rights & Access

Expert Resources

Ask the Recommending Officer: The Civil War Sesquicentennial Web Archive

US Civil War: Selected Resources

Exhibition: The Civil War in America

Collection: Civil War

Web Archiving Program Information

Tips on Searching the Web Archive

About this Collection

The American Civil War Sesquicentennial Web Archive is a collection of websites presenting various projects and events in commemoration of the 150th anniversary of the American Civil War. This anniversary is historically significant, and this significance is reflected in the number and diversity of the associated websites. This web archive will serve as a useful tool for researchers in studying the American Civil War and its memorialization.

Over 200 sites comprise this archive. Some sites document Civil War reenactments and living history events. Some are specific to aspects of the war or collections of war-related materials. Other sites have a reflective, solemn tone, rather in contrast to the more celebratory events seen during the 1961-1965 Civil War Centennial. Most sites are location or region specific, showing that much as it was 150 years ago, the Civil War was and remains to be a locally important part of American cultural history and identity.

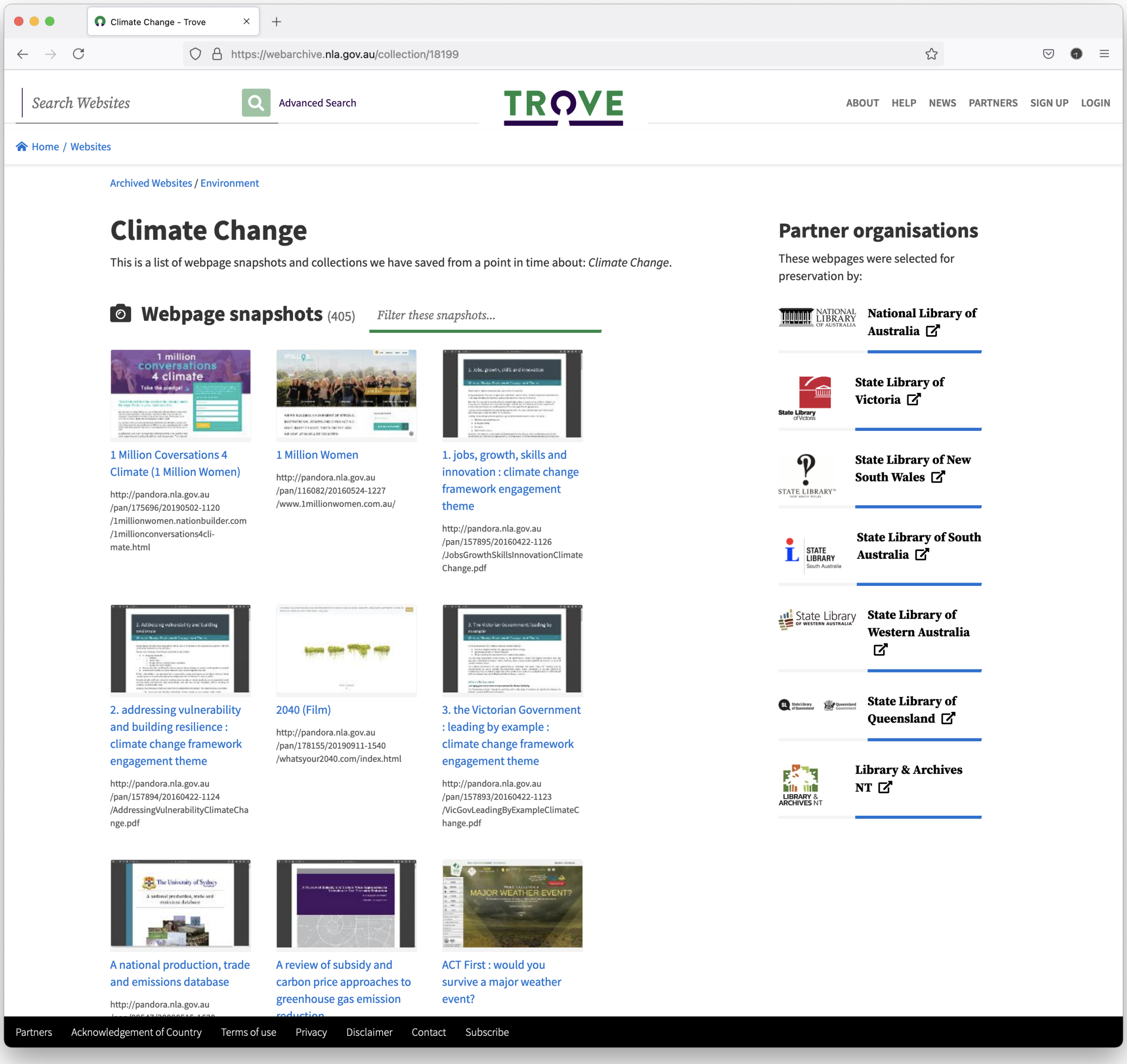
Many organizations, government bodies at all levels, and individuals created an online Sesquicentennial presence, often times instead of or in addition to print materials. This archive will act as a tool for researchers wanting to study the Civil War Sesquicentennial (2011-2015) in comparison to the Civil War Centennial (1961-1965) and other commemorative events in form, content and in the context it took place among other events going on in American society at the time.

Collection Period: March 2011 to September 2015.

Frequency of Collection: Sites in the collection were targeted for capture weekly, monthly, twice-yearly, yearly, or once.

Languages: Collection material in English.

Fig. 23. Examples of different web archive collection capabilities: Library of Congress collection *American Civil War Sesquicentennial Web Archive*. URI: `https://www.loc.gov/collections/american-civil-war-sesquicentennial-web-archive/about-this-collection/`



Climate Change - Trove

https://webarchive.nla.gov.au/collection/18199

Search Websites | Advanced Search | TROVE | ABOUT HELP NEWS PARTNERS SIGN UP LOGIN

Home / Websites

Archived Websites / Environment

Climate Change

This is a list of webpage snapshots and collections we have saved from a point in time about: *Climate Change*.

Webpage snapshots (405)

Filter these snapshots...

1 Million Coversations 4 Climate (1 Million Women)
http://pandora.nla.gov.au/pan/175696/20190502-1120/1millionwomen.nationbuilder.com/1millionconversations4climate.html

1 Million Women
http://pandora.nla.gov.au/pan/116082/20160524-1227/www.1millionwomen.com.au/

1. jobs, growth, skills and innovation : climate change framework engagement theme
http://pandora.nla.gov.au/pan/157895/20160422-1126/JobsGrowthSkillsInnovationClimateChange.pdf

2. addressing vulnerability and building resilience : climate change framework engagement theme
http://pandora.nla.gov.au/pan/157894/20160422-1124/AddressingVulnerabilityClimateChange.pdf

2040 (Film)
http://pandora.nla.gov.au/pan/178155/20190911-1540/whatsyour2040.com/index.html

3. the Victorian Government : leading by example : climate change framework engagement theme
http://pandora.nla.gov.au/pan/157893/20160422-1123/VicGovLeadingByExampleClimateChange.pdf

A national production, trade and emissions database
http://pandora.nla.gov.au

A review of subsidy and carbon price approaches to greenhouse gas emission reduction

ACT First : would you survive a major weather event?

Partner organisations

These webpages were selected for preservation by:

- National Library of Australia
- State Library of Victoria
- State Library of New South Wales
- State Library of South Australia
- State Library of Western Australia
- State Library of Queensland
- Library & Archives NT

Partners | Acknowledgement of Country | Terms of use | Privacy | Disclaimer | Contact | Subscribe

Fig. 24. Examples of different web archive collection capabilities: Trove collection *Climate Change*.
URI: `https://webarchive.nla.gov.au/collection/18199`

25a). Using these seeds, a curator can then schedule crawls to record observations of these seeds as mementos (Figure 25b). Curators can schedule crawls per seed to occur once, twice daily, daily, weekly, monthly, bimonthly, quarterly, semiannually, or annually. Once captured, mementos are then available for browsing as part of the collection.

Archive-It supports private collections, an example of which is seen in Figure 26. In our work, we focus on public collections for study because their data is available. It is possible for a collection to start private and then become publicly available or vice versa.

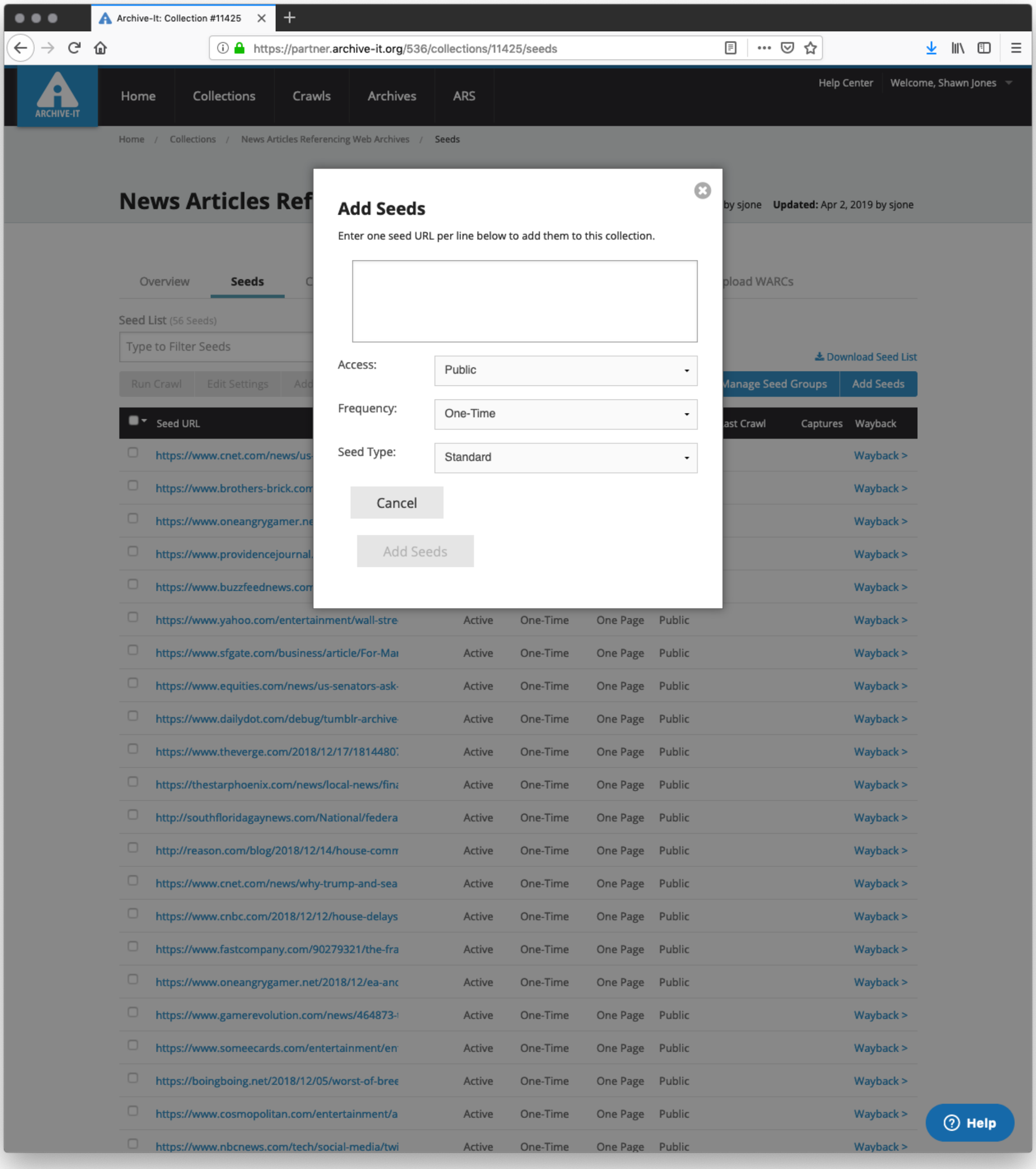
Archive-It allows the user to describe a collection via metadata, shown in Figure 27, using fields from the Dublin Core [19] standard. The system also allows users to create their metadata fields. Unlike the metadata applied to standard library records for published work, there is no standard entity establishing or enforcing content standards or rules of interpretation for Archive-It metadata. This lack of enforcement allows curators to grow collections quickly, but as mentioned in Chapter 1, this lack of enforcement makes it difficult to use the metadata to compare two collections or tell one collection apart from another.

The Archive-It user interface allows a user to search in order to match seed metadata (Figure 28a), or memento content (Figure 28b). For this to be effective, a user needs to know what they are looking for *a priori* and there are more than 14,000 Archive-It collections to view. In Chapter 4.4, we show how our intelligent sampling algorithms surface mementos that the search engine does not return to the user.

We can access Archive-It collections by appending the collection number to the URI `https://archive-it.org/collections/`, so collection 366 would be `https://archive-it.org/collections/366`. Throughout this work, we identify these collections by numbers rather than repeating the same URI prefix.

Figure 29 displays a hierarchical view of an Archive-It collection. The curator provides seeds and seed metadata. Crawls produce mementos from seeds. The seeds at the start of each crawl are **seed mementos**. There may be multiple crawls for each collection, and hence multiple seed mementos. The curator specifies crawling rules and can change them at any time. As noted in Section 2.2, crawling often follows links from the crawled pages, leading to **deep mementos**. In our work, we rely upon the semantic meaning imparted to the collection by the curator; thus, we know that the curator intended to add metadata, seeds, and seed mementos to the collection. Due to the nature of crawling rules, we do not know which deep mementos the curator intended to be part of the collection; hence we do not consider them in our analysis. The term mementos in the rest of this work will refer to seed mementos.

In Chapter 4.1 we will describe the structural features present in an Archive-It collection. In



Archive-It: Collection #11425
https://partner.archive-it.org/536/collections/11425/seeds
ARCHIVE-IT
Home
Collections
Crawls
Archives
ARS
Help Center
Welcome, Shawn Jones
Home / Collections / News Articles Referencing Web Archives / Seeds

News Articles Ref

by sjone **Updated:** Apr 2, 2019 by sjone

Seed List (56 Seeds)

Add Seeds

Enter one seed URL per line below to add them to this collection.

Access: Public

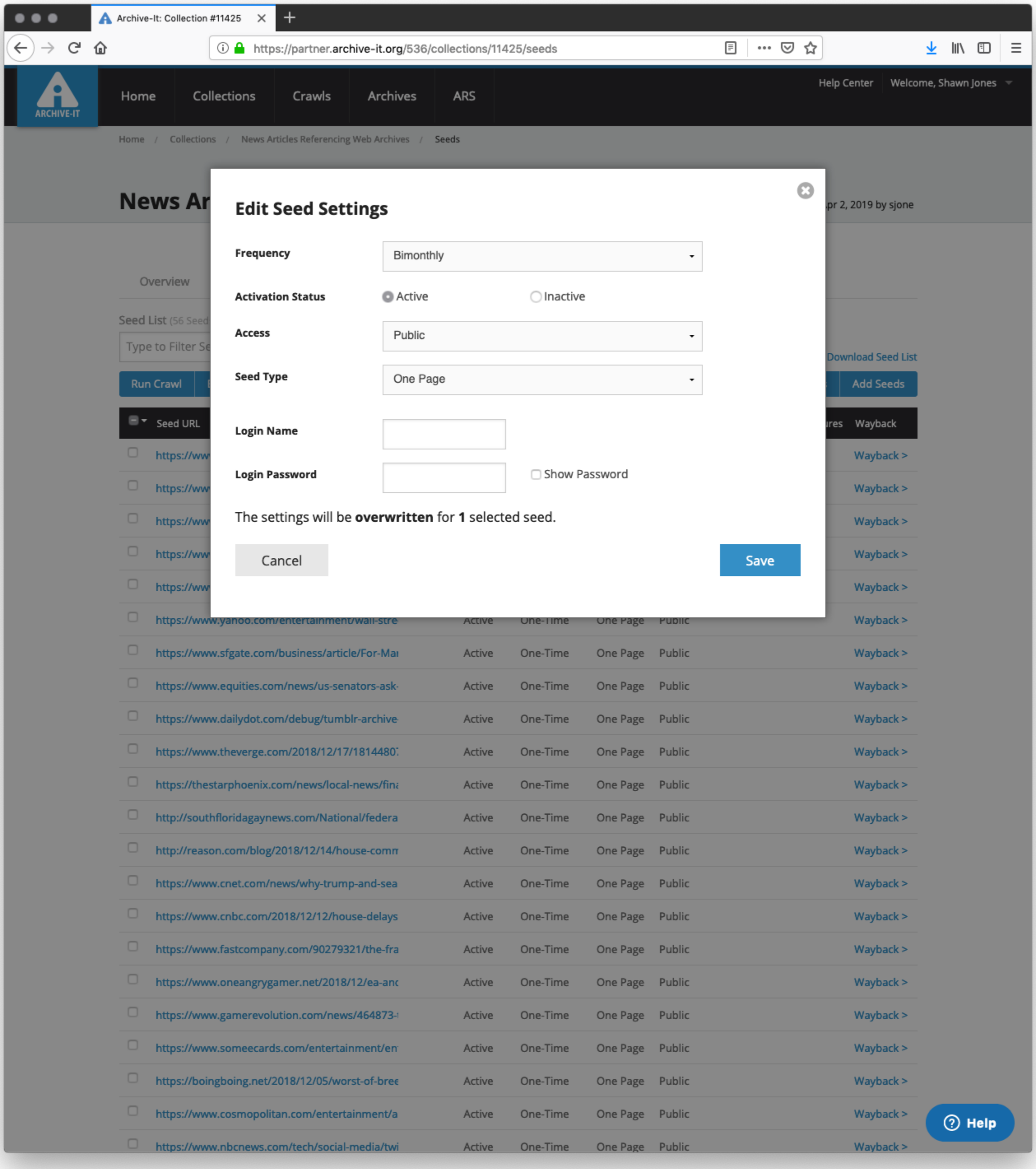
Frequency: One-Time

Seed Type: Standard

☐	Seed URL					ast Crawl	Captures	Wayback
☐	https://www.cnet.com/news/us							Wayback >
☐	https://www.brothers-brick.com							Wayback >
☐	https://www.oneangrygamer.ne							Wayback >
☐	https://www.providencejournal.							Wayback >
☐	https://www.buzzfeednews.com							Wayback >
☐	https://www.yahoo.com/entertainment/wall-stre	Active	One-Time	One Page	Public			Wayback >
☐	https://www.sfgate.com/business/article/For-Mai	Active	One-Time	One Page	Public			Wayback >
☐	https://www.equities.com/news/us-senators-ask-	Active	One-Time	One Page	Public			Wayback >
☐	https://www.dailydot.com/debug/tumblr-archive-	Active	One-Time	One Page	Public			Wayback >
☐	https://www.theverge.com/2018/12/17/1814480	Active	One-Time	One Page	Public			Wayback >
☐	https://thestarphoenix.com/news/local-news/fin	Active	One-Time	One Page	Public			Wayback >
☐	http://southfloridagaynews.com/National/federa	Active	One-Time	One Page	Public			Wayback >
☐	http://reason.com/blog/2018/12/14/house-comn	Active	One-Time	One Page	Public			Wayback >
☐	https://www.cnet.com/news/why-trump-and-sea	Active	One-Time	One Page	Public			Wayback >
☐	https://www.cnbc.com/2018/12/12/house-delays	Active	One-Time	One Page	Public			Wayback >
☐	https://www.fastcompany.com/90279321/the-fra	Active	One-Time	One Page	Public			Wayback >
☐	https://www.oneangrygamer.net/2018/12/ea-anc	Active	One-Time	One Page	Public			Wayback >
☐	https://www.gamerevolution.com/news/464873-	Active	One-Time	One Page	Public			Wayback >
☐	https://www.someecards.com/entertainment/en	Active	One-Time	One Page	Public			Wayback >
☐	https://boingboing.net/2018/12/05/worst-of-bree	Active	One-Time	One Page	Public			Wayback >
☐	https://www.cosmopolitan.com/entertainment/a	Active	One-Time	One Page	Public			Wayback >
☐	https://www.nbcnews.com/tech/social-media/twi	Active	One-Time	One Page	Public			Wayback >

(a) adding seeds

Fig. 25. Archive-It provides an easy interface for users to perform operations. (Screenshots taken in 2019.)



Edit Seed Settings

Frequency Bimonthly

Activation Status ● Active ○ Inactive

Access Public

Seed Type One Page

Login Name

Login Password □ Show Password

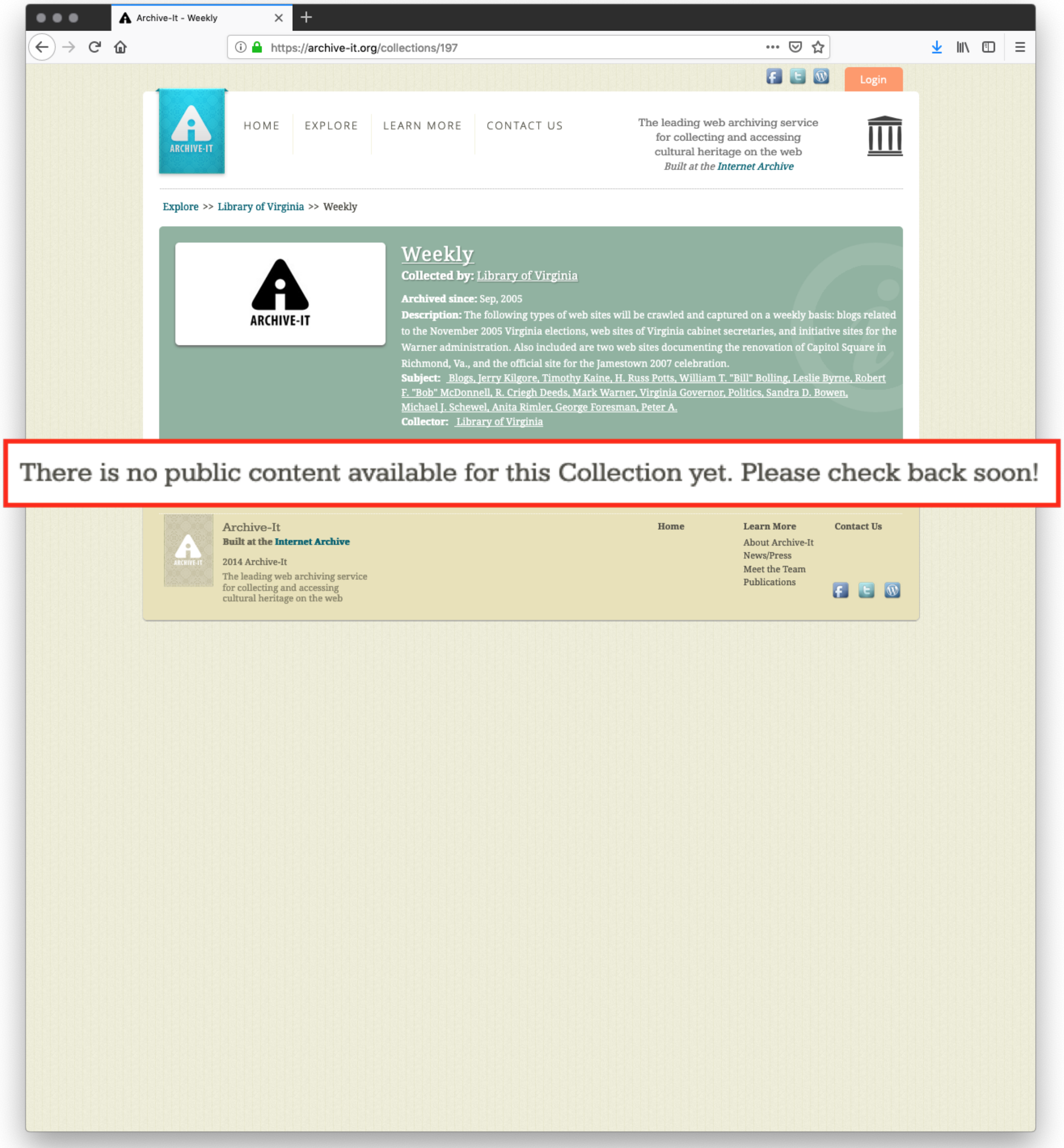
The settings will be **overwritten** for **1** selected seed.

Cancel Save

Seed URL					Wayback
https://ww					Wayback >
https://ww					Wayback >
https://ww					Wayback >
https://ww					Wayback >
https://ww					Wayback >
https://www.yahoo.com/entertainment/wall-stre	Active	One-Time	One Page	Public	Wayback >
https://www.sfgate.com/business/article/For-Ma	Active	One-Time	One Page	Public	Wayback >
https://www.equities.com/news/us-senators-ask-	Active	One-Time	One Page	Public	Wayback >
https://www.dailydot.com/debug/tumblr-archive	Active	One-Time	One Page	Public	Wayback >
https://www.theverge.com/2018/12/17/1814480	Active	One-Time	One Page	Public	Wayback >
https://thestarphoenix.com/news/local-news/fin	Active	One-Time	One Page	Public	Wayback >
http://southfloridagaynews.com/National/federa	Active	One-Time	One Page	Public	Wayback >
http://reason.com/blog/2018/12/14/house-comn	Active	One-Time	One Page	Public	Wayback >
https://www.cnet.com/news/why-trump-and-sea	Active	One-Time	One Page	Public	Wayback >
https://www.cnbc.com/2018/12/12/house-delays	Active	One-Time	One Page	Public	Wayback >
https://www.fastcompany.com/90279321/the-fra	Active	One-Time	One Page	Public	Wayback >
https://www.oneangrygamer.net/2018/12/ea-anc	Active	One-Time	One Page	Public	Wayback >
https://www.gamerevolution.com/news/464873-	Active	One-Time	One Page	Public	Wayback >
https://www.someecards.com/entertainment/en	Active	One-Time	One Page	Public	Wayback >
https://boingboing.net/2018/12/05/worst-of-bree	Active	One-Time	One Page	Public	Wayback >
https://www.cosmopolitan.com/entertainment/a	Active	One-Time	One Page	Public	Wayback >
https://www.nbcnews.com/tech/social-media/twi	Active	One-Time	One Page	Public	Wayback >

(b) specifying crawling settings

Fig. 25. (Continued) Archive-It provides an easy interface for users to perform operations. (Screenshots taken in 2019.)



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

Explore >> Library of Virginia >> Weekly

Weekly

Collected by: Library of Virginia

Archived since: Sep, 2005

Description: The following types of web sites will be crawled and captured on a weekly basis: blogs related to the November 2005 Virginia elections, web sites of Virginia cabinet secretaries, and initiative sites for the Warner administration. Also included are two web sites documenting the renovation of Capitol Square in Richmond, Va., and the official site for the Jamestown 2007 celebration.

Subject: Blogs, Jerry Kilgore, Timothy Kaine, H. Russ Potts, William T. "Bill" Bolling, Leslie Byrne, Robert F. "Bob" McDonnell, R. Criegh Deeds, Mark Warner, Virginia Governor, Politics, Sandra D. Bowen, Michael J. Schewel, Anita Rimler, George Foresman, Peter A.

Collector: Library of Virginia

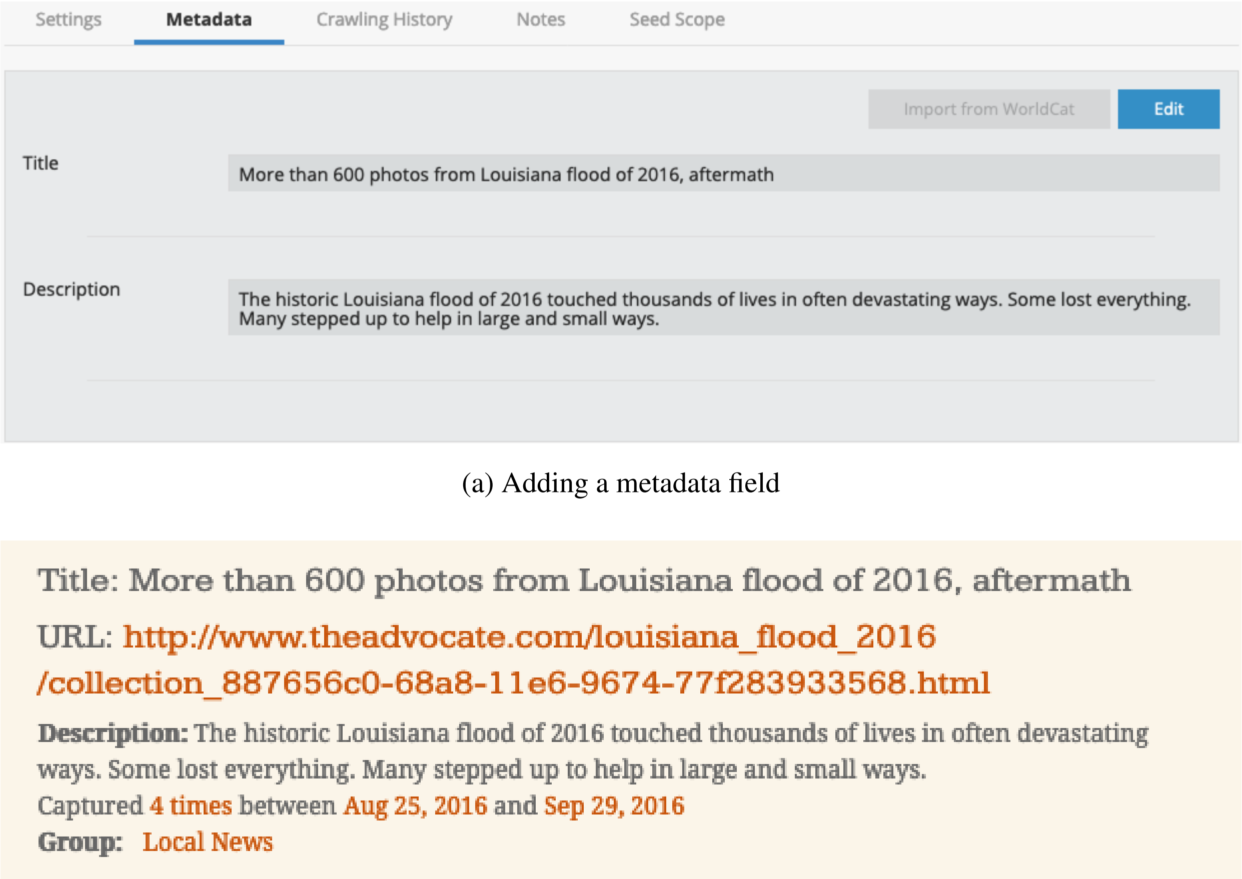
There is no public content available for this Collection yet. Please check back soon!

Archive-It
Built at the Internet Archive

2014 Archive-It

The leading web archiving service
for collecting and accessing
cultural heritage on the web

Fig. 26. Not all Archive-It collections are public. This is private collection 197: *Weekly* by the Library of Virginia. (Screenshot taken in 2019.)



Title: More than 600 photos from Louisiana flood of 2016, aftermath

Description: The historic Louisiana flood of 2016 touched thousands of lives in often devastating ways. Some lost everything. Many stepped up to help in large and small ways.

(a) Adding a metadata field

Title: More than 600 photos from Louisiana flood of 2016, aftermath

URL: http://www.theadvocate.com/louisiana_flood_2016/collection_887656c0-68a8-11e6-9674-77f283933568.html

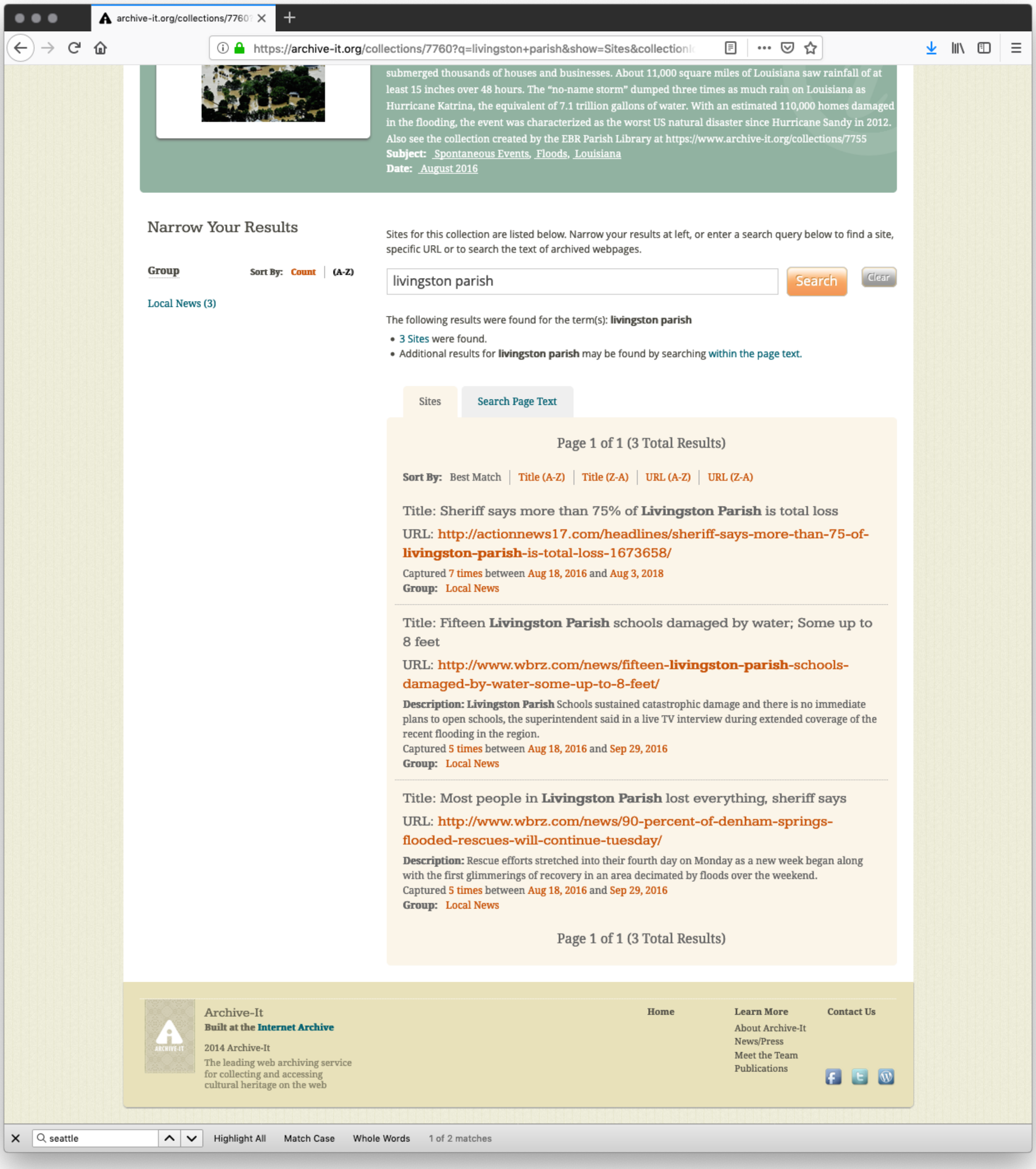
Description: The historic Louisiana flood of 2016 touched thousands of lives in often devastating ways. Some lost everything. Many stepped up to help in large and small ways.

Captured 4 times between Aug 25, 2016 and Sep 29, 2016

Group: Local News

(b) Viewing added metadata field

Fig. 27. Archive-It allows a user to add metadata fields to seeds. In this case, the metadata fields added were “Title” and “Description.” (Screenshots taken in 2019.)


archive-it.org/collections/7760
https://archive-it.org/collections/7760?q=livingston+parish&show=Sites&collection
seattle
Highlight All
Match Case
Whole Words
1 of 2 matches

submerged thousands of houses and businesses. About 11,000 square miles of Louisiana saw rainfall of at least 15 inches over 48 hours. The "no-name storm" dumped three times as much rain on Louisiana as Hurricane Katrina, the equivalent of 7.1 trillion gallons of water. With an estimated 110,000 homes damaged in the flooding, the event was characterized as the worst US natural disaster since Hurricane Sandy in 2012. Also see the collection created by the EBR Parish Library at https://www.archive-it.org/collections/7755
Subject: Spontaneous Events, Floods, Louisiana
Date: August 2016

Narrow Your Results

Group **Sort By:** **Count** | **(A-Z)**

Local News (3)

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.

livingston parish | Search | Clear

The following results were found for the term(s): **livingston parish**

- 3 Sites were found.
- Additional results for **livingston parish** may be found by searching within the page text.

Sites | Search Page Text

Page 1 of 1 (3 Total Results)

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

Title: Sheriff says more than 75% of **Livingston Parish** is total loss
URL: http://actionnews17.com/headlines/sheriff-says-more-than-75-of-**livingston-parish**-is-total-loss-1673658/
Captured 7 times between Aug 18, 2016 and Aug 3, 2018
Group: Local News

Title: Fifteen **Livingston Parish** schools damaged by water; Some up to 8 feet
URL: http://www.wbrz.com/news/fifteen-**livingston-parish**-schools-damaged-by-water-some-up-to-8-feet/
Description: Livingston Parish Schools sustained catastrophic damage and there is no immediate plans to open schools, the superintendent said in a live TV interview during extended coverage of the recent flooding in the region.
Captured 5 times between Aug 18, 2016 and Sep 29, 2016
Group: Local News

Title: Most people in **Livingston Parish** lost everything, sheriff says
URL: http://www.wbrz.com/news/90-percent-of-denham-springs-flooded-rescues-will-continue-tuesday/
Description: Rescue efforts stretched into their fourth day on Monday as a new week began along with the first glimmerings of recovery in an area decimated by floods over the weekend.
Captured 5 times between Aug 18, 2016 and Sep 29, 2016
Group: Local News

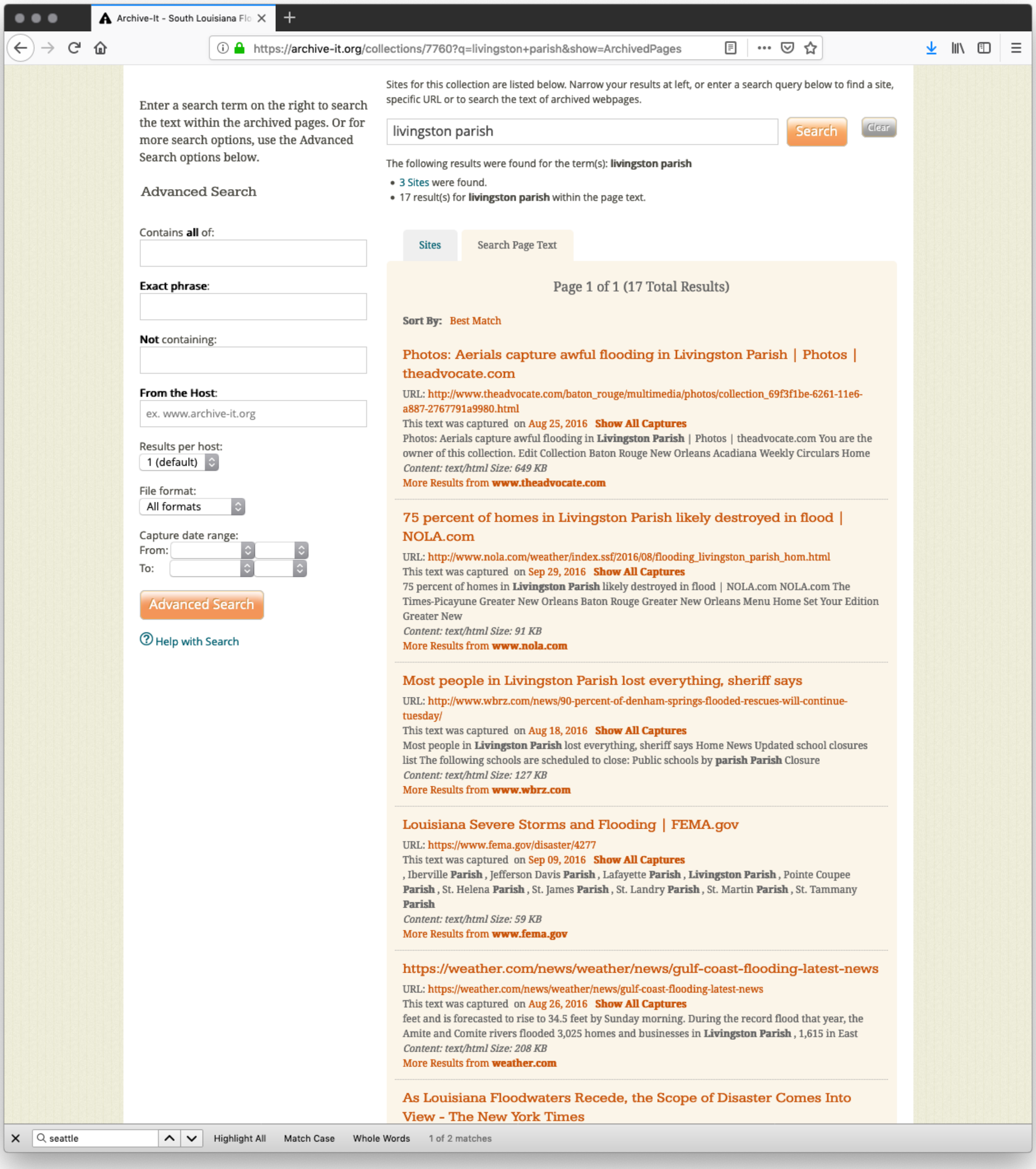
Page 1 of 1 (3 Total Results)

Archive-It
Built at the Internet Archive
2014 Archive-It
The leading web archiving service for collecting and accessing cultural heritage on the web

Home **Learn More** **Contact Us**
About Archive-It
News/Press
Meet the Team
Publications

(a) matching queries to seed metadata

Fig. 28. Archive-It's search engine provides different search targets. (Screenshots taken in 2019.)


Archive-It - South Louisiana Flo
https://archive-it.org/collections/7760?q=livingston+parish&show=ArchivedPages

Enter a search term on the right to search the text within the archived pages. Or for more search options, use the Advanced Search options below.

Advanced Search

Contains **all** of:

Exact phrase:

Not containing:

From the Host: ex. www.archive-it.org

Results per host: 1 (default)

File format: All formats

Capture date range:
From:
To:

Advanced Search

Help with Search

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.

livingston parish

Search Clear

The following results were found for the term(s): **livingston parish**

- 3 Sites were found.
- 17 result(s) for **livingston parish** within the page text.

Sites | Search Page Text

Page 1 of 1 (17 Total Results)

Sort By: Best Match

Photos: Aerials capture awful flooding in Livingston Parish | Photos | theadvocate.com
URL: http://www.theadvocate.com/baton_rouge/multimedia/photos/collection_69f3f1be-6261-11e6-a887-2767791a9980.html
This text was captured on Aug 25, 2016 **Show All Captures**
Photos: Aerials capture awful flooding in **Livingston Parish** | Photos | theadvocate.com You are the owner of this collection. Edit Collection Baton Rouge New Orleans Acadiana Weekly Circulars Home
Content: text/html Size: 649 KB
More Results from **www.theadvocate.com**

75 percent of homes in Livingston Parish likely destroyed in flood | NOLA.com
URL: http://www.nola.com/weather/index.ssf/2016/08/flooding_livingston_parish_hom.html
This text was captured on Sep 29, 2016 **Show All Captures**
75 percent of homes in **Livingston Parish** likely destroyed in flood | NOLA.com NOLA.com The Times-Picayune Greater New Orleans Baton Rouge Greater New Orleans Menu Home Set Your Edition Greater New
Content: text/html Size: 91 KB
More Results from **www.nola.com**

Most people in Livingston Parish lost everything, sheriff says
URL: http://www.wbrz.com/news/90-percent-of-denham-springs-flooded-rescues-will-continue-tuesday/
This text was captured on Aug 18, 2016 **Show All Captures**
Most people in **Livingston Parish** lost everything, sheriff says Home News Updated school closures list The following schools are scheduled to close: Public schools by **parish Parish** Closure
Content: text/html Size: 127 KB
More Results from **www.wbrz.com**

Louisiana Severe Storms and Flooding | FEMA.gov
URL: https://www.fema.gov/disaster/4277
This text was captured on Sep 09, 2016 **Show All Captures**
, Iberville **Parish** , Jefferson Davis **Parish** , Lafayette **Parish** , **Livingston Parish** , Pointe Coupee **Parish** , St. Helena **Parish** , St. James **Parish** , St. Landry **Parish** , St. Martin **Parish** , St. Tammany **Parish**
Content: text/html Size: 59 KB
More Results from **www.fema.gov**

https://weather.com/news/weather/news/gulf-coast-flooding-latest-news
URL: https://weather.com/news/weather/news/gulf-coast-flooding-latest-news
This text was captured on Aug 26, 2016 **Show All Captures**
feet and is forecasted to rise to 34.5 feet by Sunday morning. During the record flood that year, the Amite and Comite rivers flooded 3,025 homes and businesses in **Livingston Parish** , 1,615 in East
Content: text/html Size: 208 KB
More Results from **weather.com**

As Louisiana Floodwaters Recede, the Scope of Disaster Comes Into View - The New York Times

(b) matching queries to memento content

Fig. 28. (Continued) Archive-It's search engine provides different search targets. (Screenshots taken in 2019.)

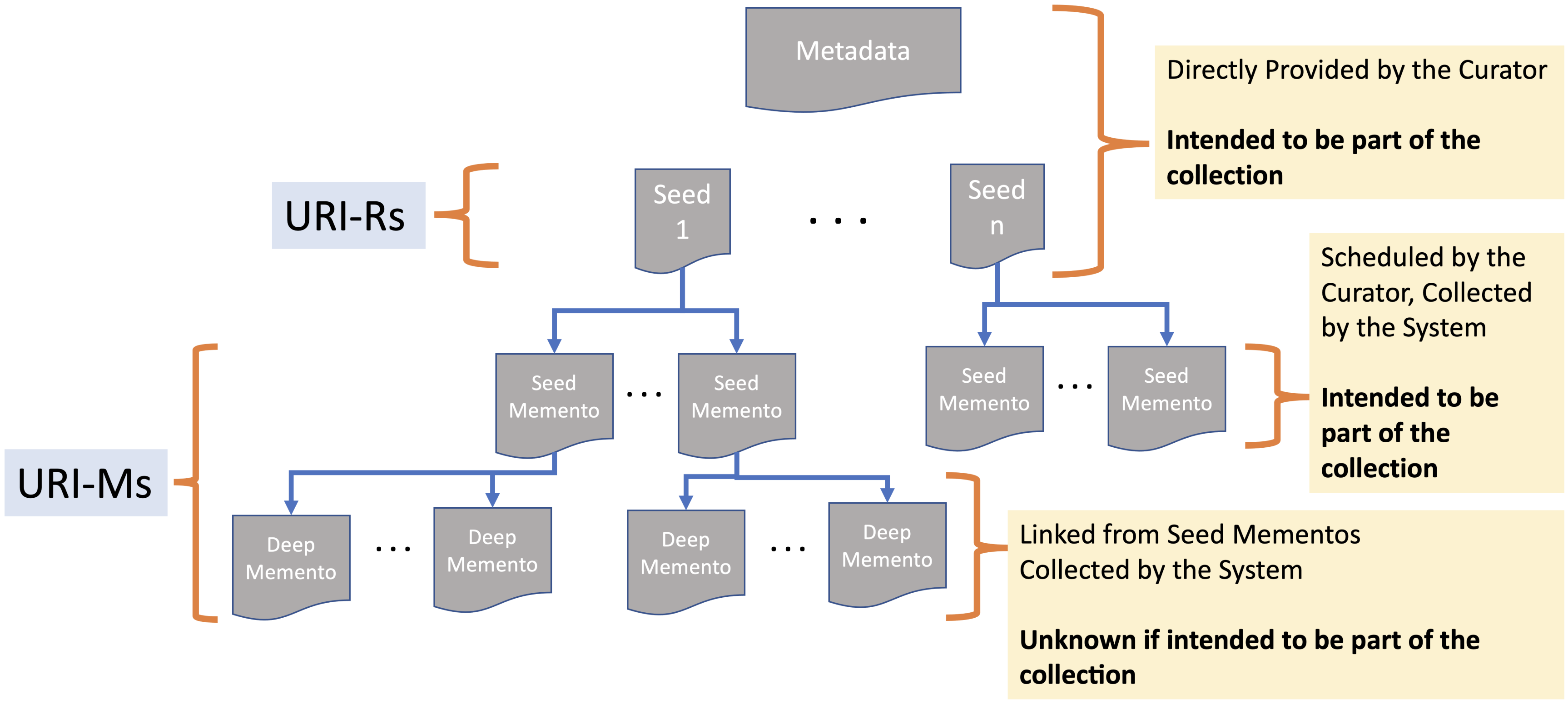

Metadata
URI-Rs
Seed 1
. . .
Seed n
Directly Provided by the Curator
Intended to be part of the collection
Seed Memento
. . .
Seed Memento
Seed Memento
. . .
Seed Memento
Scheduled by the Curator, Collected by the System
Intended to be part of the collection
URI-Ms
Deep Memento
. . .
Deep Memento
Deep Memento
. . .
Deep Memento
Linked from Seed Mementos Collected by the System
Unknown if intended to be part of the collection

Fig. 29. A view of an Archive-It collection from the outside. The top two layers are what we see from the Archive-It collection page.

Chapter 4.2 we will discuss the problem of seeds changing their content over time, resulting in seed mementos that are off-topic. Detecting these off-topic mementos is a necessary first step to selecting exemplars.

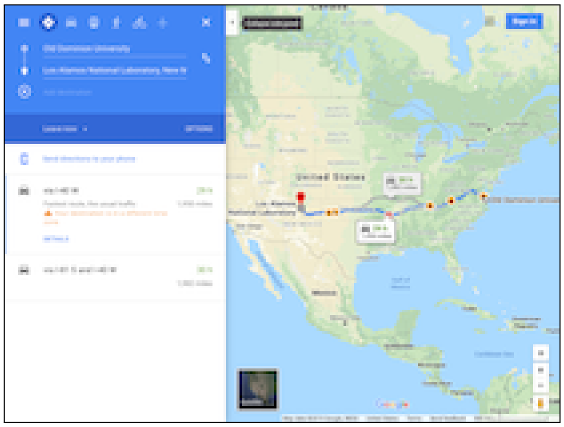
2.4 WEB PAGE SURROGATES

The web standard intends for URIs to be opaque [32]: the URI does not need to contain text from the underlying web resource. For example, the URI of a CNN news article about Tropical Storm Barry is `https://www.cnn.com/us/live-news/tropical-storm-barry-saturday-2019-intl/index.html` and the URI to a YouTube video from CNN on the same topic is `https://www.youtube.com/watch?v=x8EPX35dOjQ`. While the first URI contains the terms "tropical", "storm", and "barry", the second does not indicate the content behind the URI. Opaqueness frees the generation of URIs from their underlying content, allowing the content to change while the URI continues to link to the same resource. Opaqueness also presents an issue with viewing URIs themselves, as we do not know what exists in the referenced resource. Surrogates give the user some idea of the content of the underlying resource. Surrogates fill this role by providing a summary of the resource. For information retrieval (IR), surrogates individually help the user answer the question of "Will this link meet my information need?" For our storytelling purposes, surrogates as a group help answer the question of "Will this collection meet my information need?"

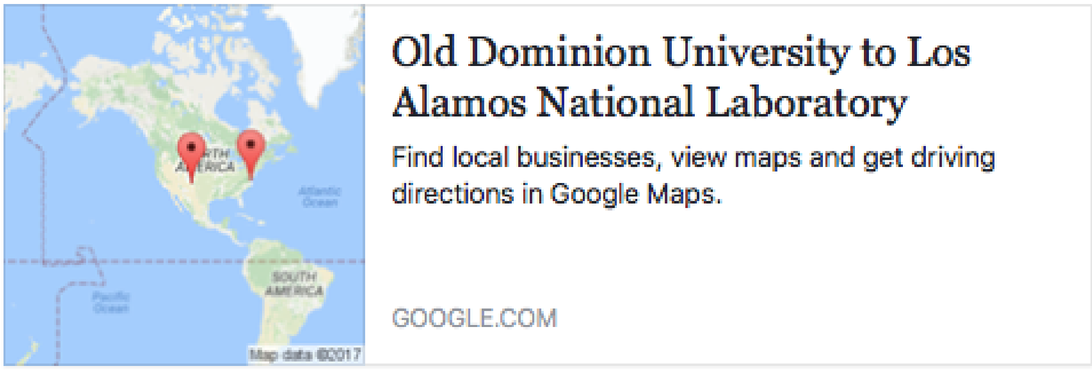
Consider the URI in Figure 30a. Even though the ability to copy and paste mitigates many issues with having to type this URI, it is still difficult to comprehend. There is also very little information in the URI indicating to what document it will lead the end-user. Different types of surrogates help users understand what is behind the URI. Figure 30b shows a browser thumbnail produced by taking a screenshot of the page with this URI. Figure 30c shows a social card generated by Facebook for that same URI. Each type of surrogate contains different information and thus provides the user with a different perspective on the URI.

As shown in Figure 31, full social cards consist of a page title, striking image, domain name, and description. To allow authors to supply their own values for these units, both Facebook and Twitter have developed the respective Open Graph Protocol (OGP) [87] and Twitter card [306] standards. To supply a value for a social card unit, a web page author inserts the key-value pairs into the attributes of an HTML `META` element (e.g., `<META property="og:image" content="http://example/image.png">`). Table 8 lists the standard fields to be used with each social card unit. Most fields are optional, and cards will omit missing data, as seen in Figure 32. From our experience, Twitter requires that `twitter:card` and `twitter:title` or `https://www.google.com/maps/dir/Old+Dominion+University,+Norfolk,+VA/Los+Alamos+National+Laboratory,+New+Mexico/@35.3644614,-109.356967,4z/data=!3m1!4b1!4m13!4m12!1m5!1m1!1s0x89ba99ad24ba3945:0xcd2bdc432c4e4bac!2m2!1d-76.3067676!2d36.8855515!1m5!1m1!1s0x87181246af22e765:0x7f5a90170c5df1b4!2m2!1d-106.287162!2d35.8440582`

(a) A long URI



(b) The same URI rendered as a browser thumbnail surrogate.



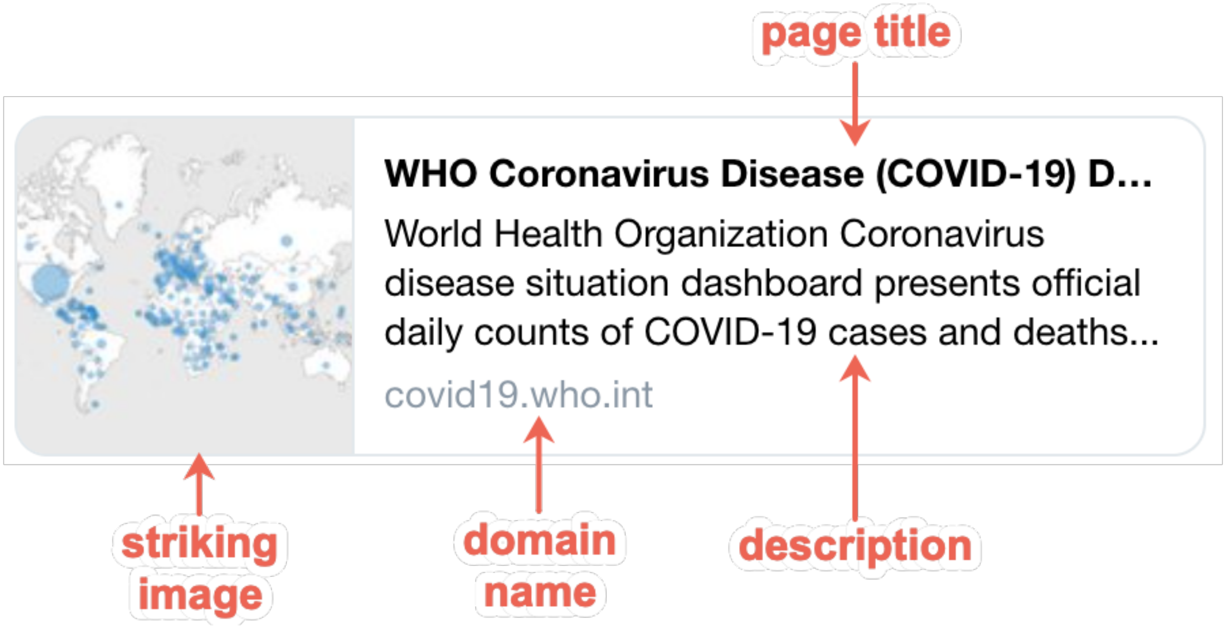
Old Dominion University to Los Alamos National Laboratory

Find local businesses, view maps and get driving directions in Google Maps.

GOOGLE.COM

(c) The same URI rendered as a social card surrogate.

Fig. 30. Web page surrogates give a view into the content behind the URI. (Screenshots taken in May 2018.)



WHO Coronavirus Disease (COVID-19) D...

World Health Organization Coronavirus disease situation dashboard presents official daily counts of COVID-19 cases and deaths...

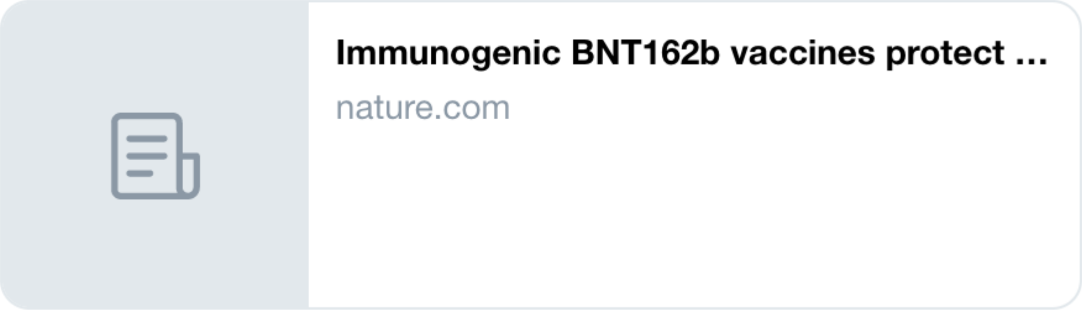
covid19.who.int

Fig. 31. An annotated social card from Twitter for the URI `https://covid19.who.int/` identifying its social card units. (Screenshot taken in January 2021.)

TABLE 8

Social card units and their associated cards standards field keys.

Social Card Unit	OGP (Facebook)	Twitter Card
title	`og:title`	`twitter:title`
description	`og:description`	`twitter:description`
striking image	`og:image`	`twitter:image`
identify the resource	`og:url`	N/A
specify the type of resource	`og:type`	`twitter:card`



Immunogenic BNT162b vaccines protect ...

nature.com

Fig. 32. A social card from Twitter for the *Nature* article URI `https://www.nature.com/articles/s41586-021-03275-y`, showing a lack of description or image.
(Screenshot taken in January 2021.)

`og:title` exist or it will not create a card. Facebook is more forgiving, generating a card consisting only of the content of a page's HTML `title` element and domain name if no metadata fields are specified. With the exception of `twitter:card`, if any other `twitter:*` fields are missing, Twitter will use the values of a corresponding `og:*` field if present [306]. Deprecated or undocumented equivalents, such as `twitter:image:url`, may also be interpreted by the corresponding social media service. If present, Facebook favors the value of `og:title` over the content of the page's `title` element. The OGP specification states that Facebook should generate a card containing a description if `og:description` is present, but based on our experience, Facebook will provide a description in its card only if both `og:description` and `og:image` are present.

When visualized as a group, social cards work for social media storytelling because they have the same content elements in the same locations on each card. Each social card is a visualization. Because they contain the same metadata in the same location, a group of social cards is an instance of **small multiples**, as defined by Tufte [303]. Figure 33 displays an example of a small multiple for line graphs.

Social cards and browser thumbnails are just two types of surrogates. In Chapter 5.1 we will summarize the results of an evaluation of 55 platforms that perform one or more of our storytelling processes with more detail available in Jones et al. [140, 156]. In Chapter 5.2, we conduct a user study to determine which surrogate is best for collection understanding. We will discuss how to automatically create surrogates when we have no document metadata in Chapter 5.3. In Chapter 5.4 we will introduce our own surrogate service, MementoEmbed, and in Chapter 6.1, its

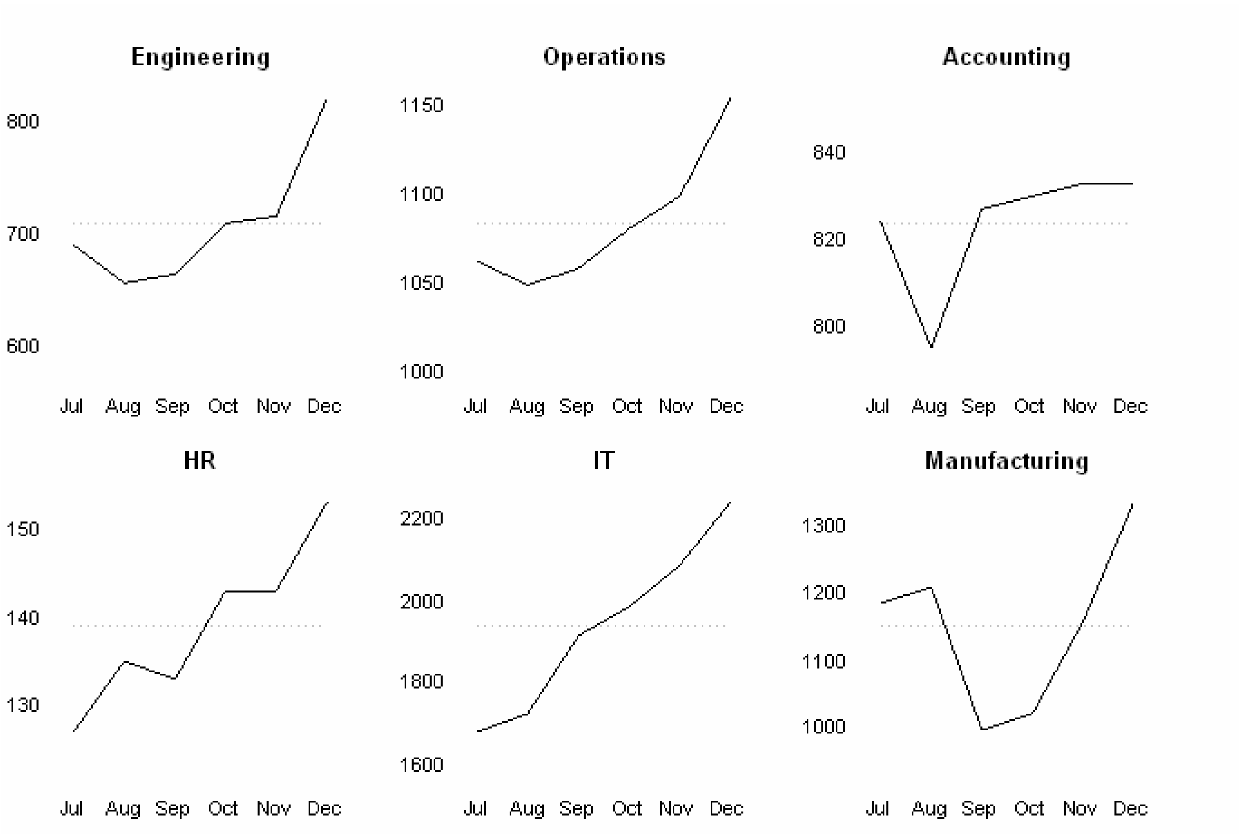


Fig. 33. An example of small multiples using line charts. Image courtesy of Wikipedia [326].

companion storytelling application, Raintale.

2.5 FORMS OF STORYTELLING

"Storytelling" is a word applied to many different concepts. We focus on the concept of social media storytelling as conveyed by services such as Wakelet[12] and the now-defunct Storify. These social media stories are micro-collections using text, images, and web page surrogates. In all cases, the images and text provide a way for the author to clarify the narrative. The example in Figure 34a shows how cards and text are used with Wakelet to describe the crisis of the Rohingya minority in India. Figure 34b shows a story built with Storify as preserved by the Internet Archive. The web page surrogates provide links to supporting evidence, but they also give the reader a summary of that evidence. The surrogates allow the end-user to consume the story without clicking on the links until they are ready.

According to Wikipedia [327], stories "involve plot, characters, and a narrative point of view." Social media stories contain characters as named entities displayed in surrogates. The ordering and choice of resources define the plot. The narrative point of view comes as a result of the resources that the storyteller chooses.

Librarians use interviews to elicit stories from people who have lived through a historical event or have esoteric knowledge. In this case, the listener is asking the participant to be the storyteller. From the storyteller's information, the listener can then synthesize a micro-collection of resources such as interview responses (text), photographs (images), and summaries of personal writings (surrogates). Storied[13] and StoryKit[14] are tools that help libraries with this form of storytelling.

Businesses and other organizations use stories to help the organization retain knowledge and move employees to future projects. After a team completes a project, it is common for team members to meet and discuss what worked and what did not. These retrospectives or debriefings allow the business to consolidate metrics, personal accounts, and other information into a single (typically internal) narrative about the project. The business will document this single narrative, and the employees will carry this narrative to future projects. Agile Cockpit[15], Retrium[16], RetroDesk[17], RetroTime[18], Scatterspoke[19], and Sensei[20] are all tools that facilitate information gathering and

[12] https://wakelet.com/

[13] https://www.getstoried.com

[14] https://californialistens.org/projects/story-tool-kit-and-storyfaire/

[15] https://www.agilecockpit.com/

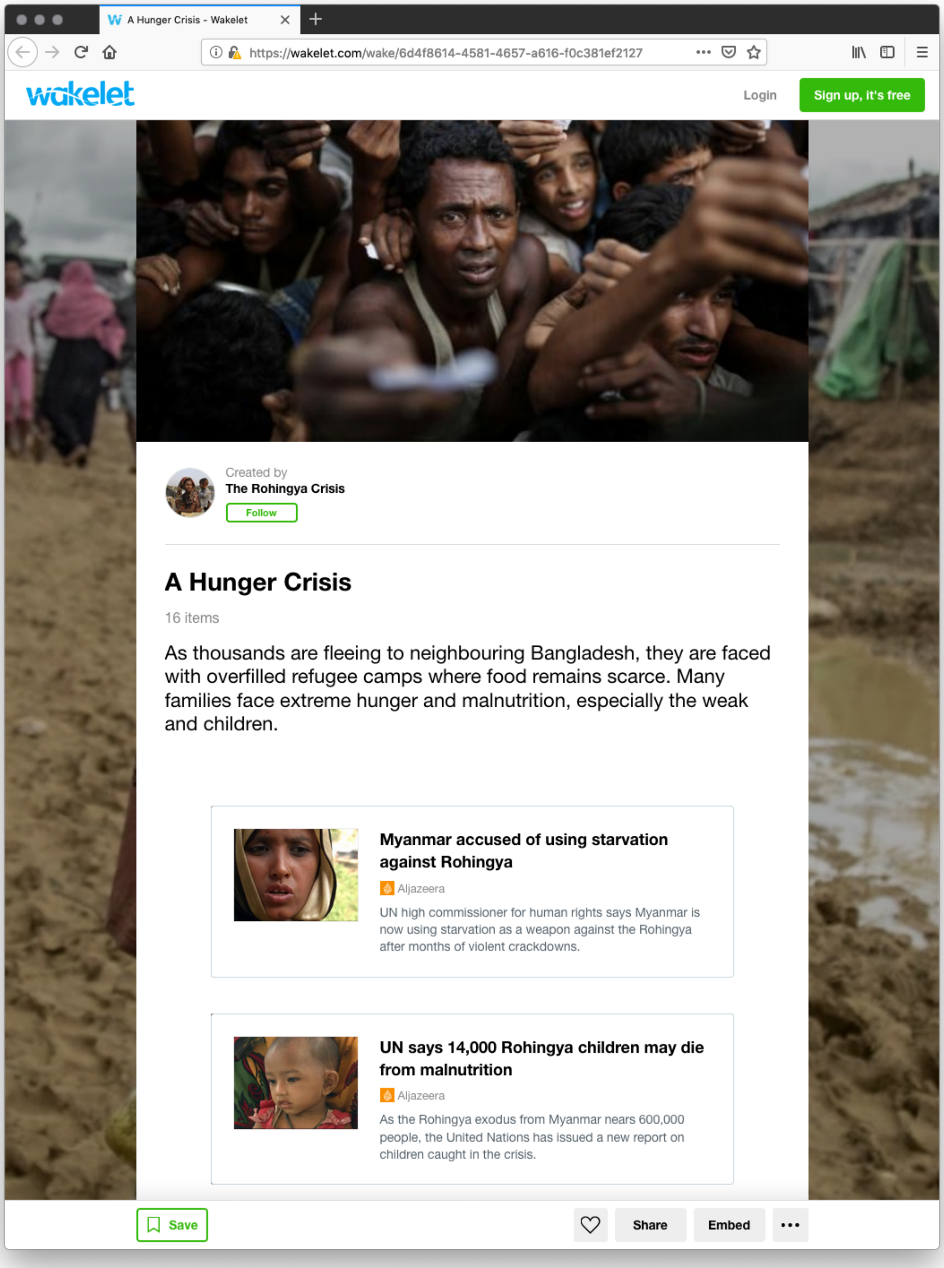
[16] https://www.retrium.com/

[17] https://www.scrumdesk.com/retrodesk/

[18] https://retroti.me/

[19] https://www.scatterspoke.com/

[20] https://www.senseitool.com/


A Hunger Crisis - Wakelet
https://wakelet.com/wake/6d4f8614-4581-4657-a616-f0c381ef2127
wakelet
Login
Sign up, it's free
Follow
Save
Share
Embed

Created by
The Rohingya Crisis

A Hunger Crisis

16 items

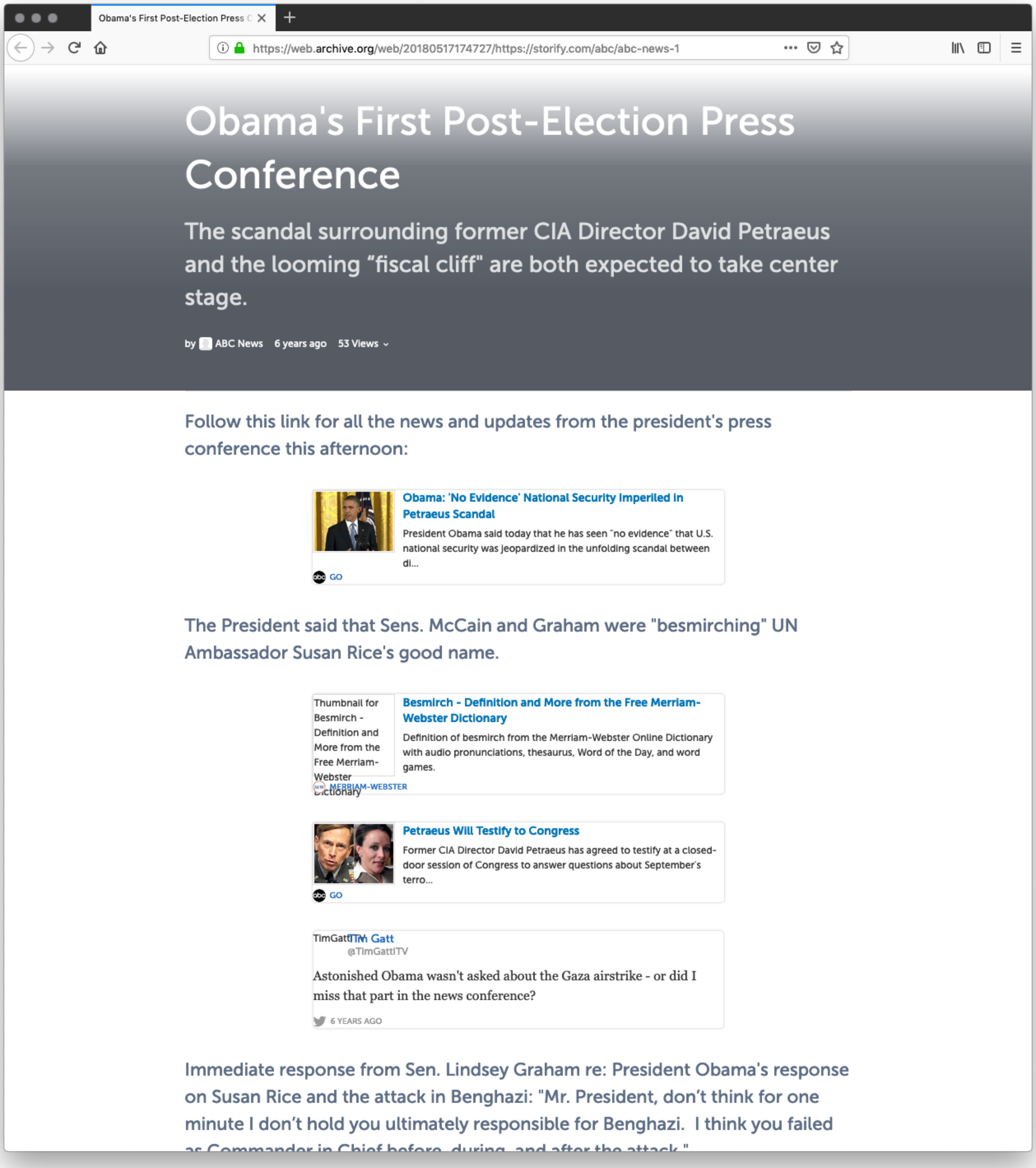
As thousands are fleeing to neighbouring Bangladesh, they are faced with overfilled refugee camps where food remains scarce. Many families face extreme hunger and malnutrition, especially the weak and children.

Myanmar accused of using starvation against Rohingya
Aljazeera
UN high commissioner for human rights says Myanmar is now using starvation as a weapon against the Rohingya after months of violent crackdowns.

UN says 14,000 Rohingya children may die from malnutrition
Aljazeera
As the Rohingya exodus from Myanmar nears 600,000 people, the United Nations has issued a new report on children caught in the crisis.

(a) Wakelet Story
URI: `https://wakelet.com/wake/6d4f8614-4581-4657-a616-f0c381ef2127`

Fig. 34. Stories created by humans using the format we intend to emulate.


Obama's First Post-Election Press C
https://web.archive.org/web/20180517174727/https://storify.com/abc/abc-news-1

Obama's First Post-Election Press Conference

The scandal surrounding former CIA Director David Petraeus and the looming "fiscal cliff" are both expected to take center stage.

by ABC News 6 years ago 53 Views

Follow this link for all the news and updates from the president's press conference this afternoon:

Obama: 'No Evidence' National Security Imperiled in Petraeus Scandal

President Obama said today that he has seen "no evidence" that U.S. national security was jeopardized in the unfolding scandal between di...

GO

The President said that Sens. McCain and Graham were "besmirching" UN Ambassador Susan Rice's good name.

Thumbnail for Besmirch - Definition and More from the Free Merriam-Webster Dictionary

Besmirch - Definition and More from the Free Merriam-Webster Dictionary

Definition of besmirch from the Merriam-Webster Online Dictionary with audio pronunciations, thesaurus, Word of the Day, and word games.

MERRIAM-WEBSTER

Petraeus Will Testify to Congress

Former CIA Director David Petraeus has agreed to testify at a closed-door session of Congress to answer questions about September's terro...

GO

TimGattTV Tim Gatt
@TimGattlTV

Astonished Obama wasn't asked about the Gaza airstrike - or did I miss that part in the news conference?

6 YEARS AGO

Immediate response from Sen. Lindsey Graham re: President Obama's response on Susan Rice and the attack in Benghazi: "Mr. President, don't think for one minute I don't hold you ultimately responsible for Benghazi. I think you failed as Commander in Chief before, during, and after the attack."

(b) Storify Story (from Internet Archive)
URI: `https://web.archive.org/web/20180517174727/https://storify.com/abc/abc-news-1`

Fig. 34. (Continued) Stories created by humans using the format we intend to emulate.

help an organization build the resulting narrative.

Marketers use storytelling to make brands seem favorable to customers. The company telling the story will synthesize metrics, corporate history, competitor history, and more into a single public narrative. This narrative typically conveys how well their brand serves the community, is better than the competition, and why it should be a part of the listener's life.

In all of these cases, information is gathered and then summarized into a smaller account of what occurred. Our social media stories are no different. We have the mementos present in a web archive collection rather than using diverse information sources like the actors mentioned above. Our algorithms will interrogate the collection and select exemplars that convey as much aboutness as possible. We then generate document metadata and visualize these mementos as surrogates just like users once did with Storify.

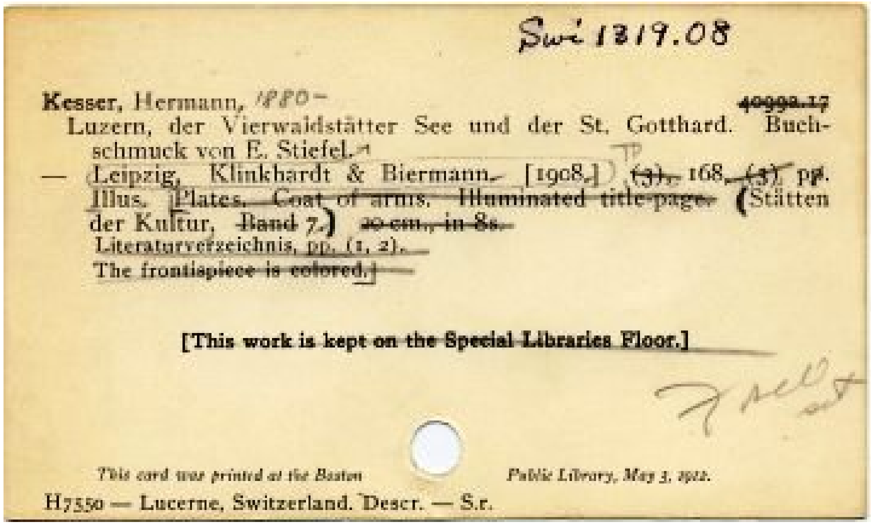
2.6 ABOUTNESS

To select exemplars and generate story metadata for a collection, we must cover the vague and inexact concept of **aboutness**. Aboutness conveys the topics, ideas, concepts, people, places, organizations, time periods, and other information about a document or an entire collection. By applying different aspects of aboutness to a resource, a system can help a user locate the item for which they are searching, or it can summarize an item.

2.6.1 TRADITIONAL METHODS OF CONVEYING ABOUTNESS

Understanding the difference between two items, whether they be collections or research papers, is usually handled by a record containing metadata about the item. Information scientists and librarians apply metadata to individual works within a collection, through a process known as **cataloging**. This metadata exists to answer specific questions about a given work, such as its author, the topics the work covers, and its publication date. For each item, its metadata is used to generate a surrogate representing that item. Historically surrogates took the form of cards sorted in drawers as part of a card catalog cabinet, as shown in Figure 35. Cataloging standards, such as the AACR2 [136], exist to ensure that metadata is gathered consistently between individuals and organizations. Once the metadata is collected, it is shared between systems using a digital format such as MAchine-Readable Cataloguing (MARC) [17]. A user can compare two surrogates to determine which item to retrieve.

The physical cards in a library card catalog are analogous to the surrogates used in search engines and social media to represent a web resource. With the physical card, a user examines the card to determine if it meets their information needs. There is often an identifier on the card



Swi 1319.08

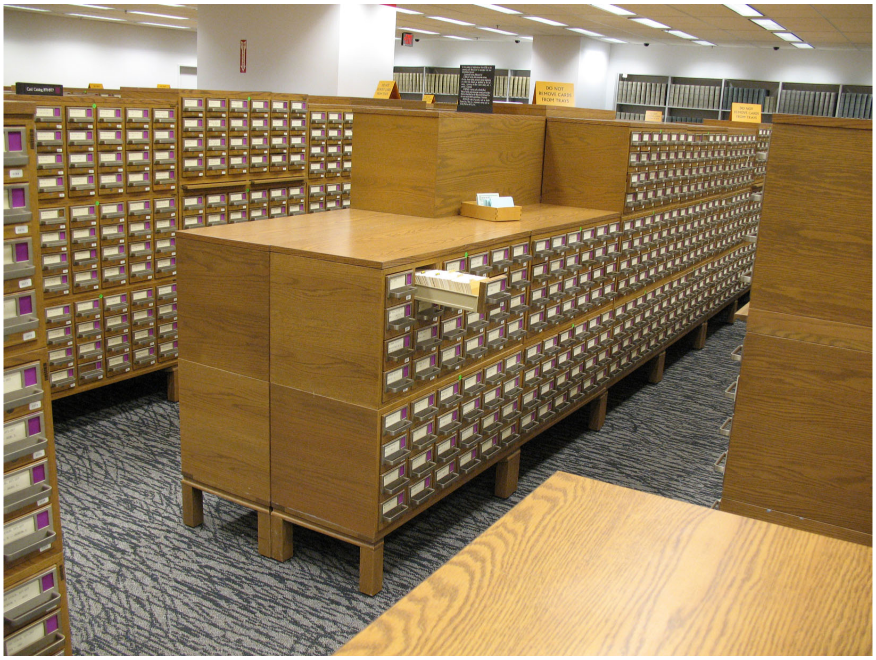
Kesser, Hermann, 1880- ~~40999.17~~
Luzern, der Vierwaldstätter See und der St. Gotthard. Buchschmuck von E. Stiefel.
— Leipzig, Klinkhardt & Biermann. [1908.] ~~(3),~~ 168, ~~(3)~~ pp. Illus. ~~Plates. Coat of arms. Illuminated title-page.~~ (Stätten der Kultur, ~~Band 7.~~) ~~20 cm., in 8s.~~
Literaturverzeichnis, pp. (1, 2).
The ~~frontispiece is colored.~~

[This work is kept on the Special Libraries Floor.]

This card was printed at the Boston Public Library, May 3, 1912.

H7550 — Lucerne, Switzerland. Descr. — S.r.

(a) A physical paper card as a surrogate for a book by Hermann Kesser, from the Harvard College Library. Image courtesy of Wikipedia [320].



(b) A card catalog cabinet at the Library of Congress. Image courtesy of Wikipedia [60].

Fig. 35. Photographs of physical library cards as surrogates representing resources in a library.

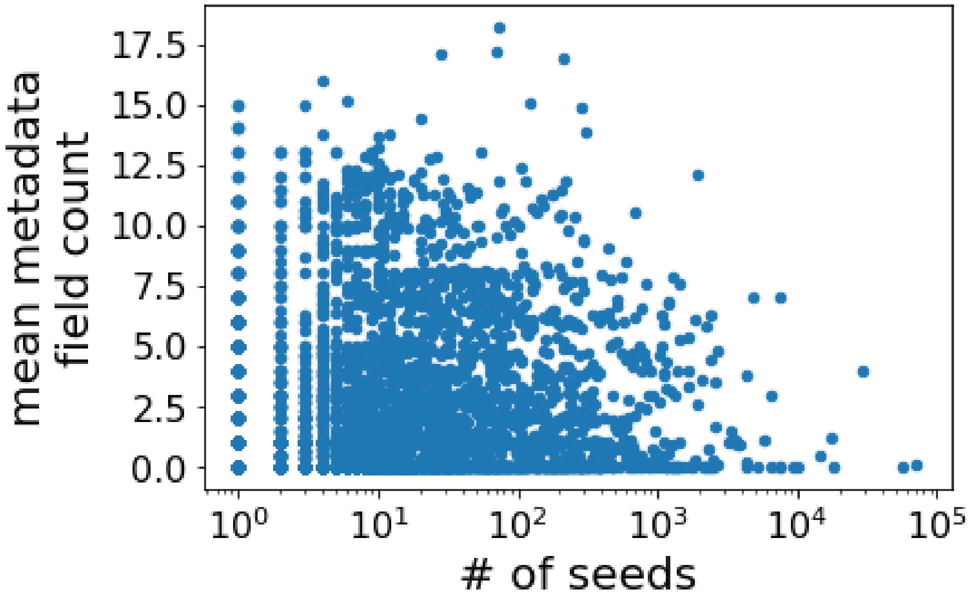


Fig. 36. A visualization of metadata field usage across all Archive-It seeds [165]. Generally, as the number of seeds in a collection goes up, the amount of metadata available goes down. Thus, the more metadata a reader needs to understand a collection, the less they have available.

(e.g., a call number) that corresponds to a physical item that the user can retrieve from a library location. Likewise, the web surrogates we introduced earlier present the user with the information needed to determine if the resource represented by the card meets their information needs. If the user wants to view the item, they no longer need to record a call number and walk to a location to retrieve it; instead, the card itself is a link, and they click on it to ask HTTP to retrieve the document for them.

As we cover in Chapter 3.2, efforts exist to provide metadata for web archive collections and provide this same level of aboutness, but all of these methods require manual application and review. For example, Archive-It provides metadata fields based on Dublin Core [19]. Unfortunately, curators do not consistently use these fields. In addition, many Archive-It collections contain hundreds of thousands of mementos for which curators would need to create metadata. Human generation of this metadata is quite a costly proposition in terms of both time and personnel. We analyzed all of the seeds from all Archive-It collections in 2019 [165] and determined that as the number of seeds increases, the amount of metadata per seed decreases, as shown in Figure 36. Thus, the more a reader needs metadata from an Archive-It collection, the less metadata there is.

2.6.2 MACHINE METHODS OF DISCOVERING ABOUTNESS

Humans can process content semantically to discover aboutness but are limited in how fast they can process the content. Machines are far faster, but their inability to understand content limits them. The sheer size of the web has necessitated the development of machine methods of processing documents. This section details some different problems and machine solutions for evaluating documents.

Humans perceive characters and words; machines only see bytes and hence are limited to processing the values of these bytes. In order to facilitate this processing, semantic models have given way to other concepts, such as the **bag of words** [65] model. A system splits the text into individual words. The text is typically **preprocessed** with processes like **stemming** and **stop word** removal. Stemming removes suffixes from words to reduce them to their root form (e.g., stemming would reduce `bake`, `baking`, and `baked` to the same word). Stemming improves the results of document and query comparison. Stop words (e.g., the, a, of) are common connecting words in sentences but create noise in the bag of words, so algorithms typically remove them. By using the bag of words model, a machine can leverage different techniques to deliver information about a document or corpus.

As a first step to converting an HTML document into a bag of words, a system must first remove the tags and excess content, a process referred to as **boilerplate removal**. Boilerplate removal is handled by many different software packages, [244, 264] with *justext* [269] as one example.

Measuring the Similarity of Two Documents

The Jaccard Index (sometimes called Jaccard Coefficient) compares two sets [129]. Jaccard originally developed this similarity measure to evaluate the differences between the features of plant species. The **Jaccard distance** can provide a comparable distance measurement to compare the two sets of words making up the documents. Jaccard distance calculates the percentage of overlap between the words in both documents, as shown in Equation 1,

$$d_J(a,b) = \frac{|t(a) \cup t(b)| - |t(a) \cap t(b)|}{|t(a) \cup t(b)|} \tag{1}$$

where $t(a)$ is the set of terms from document a and $t(b)$ is the set of terms from document b.

The Sørensen-Dice Coefficient is another method of comparing two sets [69, 290]. Like the Jaccard coefficient, it was also developed to compare the differences between species. The

Sørensen-Dice distance, shown in Equation 2 can provide a score measuring the difference between the two documents:

$$d_S(a,b) = 1 - \frac{2|t(a) \cap t(b)|}{|t(a)| + |t(b)|} \tag{2}$$

where $t(a)$ is the set of terms from document a and $t(b)$ is the set of terms from document b.

Both distance measures have scores ranging from 0.0, meaning that the documents are the same, to 1.0, meaning that the documents are entirely dissimilar.

Further exploiting the bag of words model, we can construct a data structure of a document consisting of each word and how often it occurs in the text, its **term frequency**. Simhash is a piecewise hashing algorithm that can generate an identifier for a document based on a mathematical hash function [54]. A simplified version is shown in Figure 37. Simhash works by computing the term frequencies of each word in the document and then hashing each word. A vector is then formed by summing the term frequencies. The Simash algorithm then computes a document fingerprint based on these pieces. The number of bits different between two document fingerprints is the measurement of how different the documents are from one another. Simhash has been evaluated with web pages to enhance crawling, and indexing [209].

Building on the data structure of terms and term frequencies, we can produce a vector representing a document. Consider the two documents in Figure 38. Using these vectors, we can measure the cosine between their two lines in multidimensional space to measure how different two documents are from each other. The score from this **cosine similarity** [314] can be used to rank how similar a document is to a query.

The equation for cosine similarity is:

$$s(q,d) = \frac{\sum\limits_{i=1}^{n} q_i d_i}{\sqrt{\sum\limits_{i=1}^{n} q_i^2}\sqrt{\sum\limits_{i=1}^{n} d_i^2}} \tag{3}$$

where q_i and d_i are components of the document vectors of query q and document d.

To ensure that unique terms are scored more highly in such measures, we can normalize term frequency scores with Inverse Document Frequency (IDF) [137], shown in Equation 4, where the IDF for term t in document D is the logarithm of the total number of documents N divided by d, number of documents containing term t:

$$idf(t,D) = \log\frac{N}{d} \tag{4}$$

```
original text = "Tropical fish include fish found in tropical
   environments around the world, including both freshwater and
   salt water species."

Converting to words with weights:

tropical      2  fish          2
include       1  found         1
environments  1  around        1
world         1  including     1
both          1  freshwater    1
salt          1  water         1
species       1

8-bit hash values

tropical    01100001  fish         10101011  include 11100110
found       00011110  environments 00101101  around  10001011
world       00101010  including    11000000  both    10101110
freshwater  00111111  salt         10110101  water   00100101
species     11101110

Vector formed by summing weights:

1 -5 9 -9 3 1 3 3

8-bit fingerprint formed from vector

1 0 1 0 1 1 1 1
```

Fig. 37. An example of Simhash generation.

```
document 1 = "Web archives are archives of web resources."
document 2 = "Web archive collections are archives of some web
   resources."

Converting the documents to terms and term frequencies:

web          2 2
archives     2 1
are          1 1
archive      0 1
of           1 1
resources    1 1
collections  0 1
some         0 1

produces the following vectors:

document 1 vector = [ 2, 2, 1, 0, 1, 1, 0, 0 ]
document 2 vector = [ 2, 1, 1, 1, 1, 1, 1, 1 ]
```

Fig. 38. Conversion of two simple documents into document vectors.

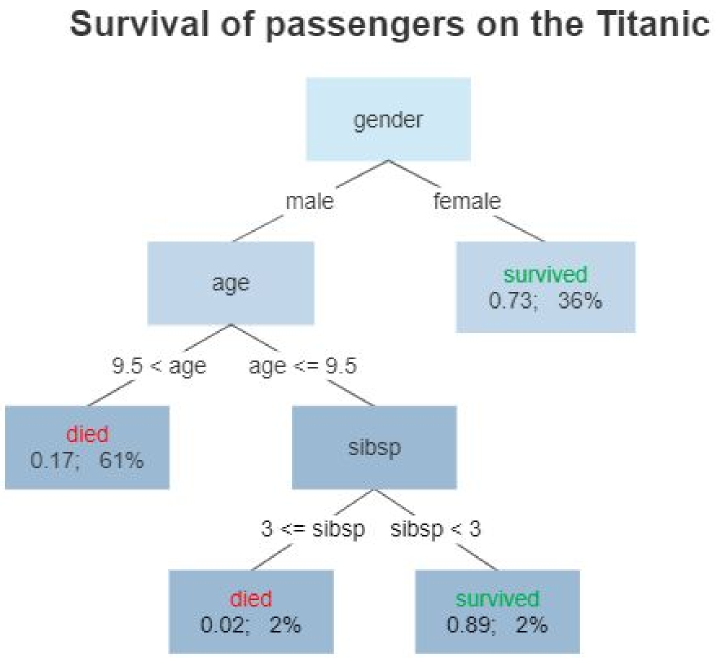


Fig. 39. An image of a decision tree courtesy of Wikipedia [324].

The resulting normalization is known as TF-IDF, where a high term frequency coupled with a low inverse document frequency weights a given word more heavily. TF-IDF can give us more precise cosine similarity measurements.

Machine Learning

Machine learning is an umbrella term for a set of techniques for discovering meaning in data. Chollet [58] points out that classical computing involves humans submitting rules and data to a computer to arrive at answers. Machine learning accepts answers and data and produces rules.

One form of machine learning is the **decision tree**. Assume that we also have the passenger manifest from the *HMS Titanic* and a listing of who did survive that disaster. We could use the results from the Titanic dataset to estimate the survivors in the disaster. If we submit the Titanic data to a decision tree algorithm, like CART [203], then it will produce a tree like the one shown in Figure 39. In this example, each data record refers to a passenger on the *Titanic*. The answer corresponding to each data record falls into either the **class** *died* or *survived*. Each branch of the tree reflects a conditional statement about the input data, and following several conditions helps us arrive at a prediction. For example, based on the tree in Figure 39, if an input record represented a male younger than 9.5 years, then there was a 61% chance that they died. Information like the age and sex of the passenger are the **features** of the record that the decision tree uses to make its prediction. We can apply this decision tree to a different passenger manifest from a ship disaster from the same time period to predict who may have died.

Ho [171], and later Breiman [41] extended the concepts of decision trees into a new algorithm named Random Forest. Random Forest averages results over multiple decision trees by building many possible trees and then selecting the class predicted by the most trees. It includes additional capabilities for selecting features (e.g., bagging) that we will not cover here. The Random Forest implementation provided by scikit-learn [261] provides a probability based on this "voting" process, and we can apply the probability scores to rank results, as we do with striking image candidates in Chapter 5.3.

These are not the only machine learning algorithms. Gaussian Naïve Bayes and Linear Discriminant Analysis employ statistics to predict to what class an input record belongs. AdaBoost optimizes the feature set to determine which features will work best for prediction and passes these features to other classifiers before aggregating their predictions. Multilayer Perceptrons attempt to model the interactions of biological neurons. These machine learning algorithms all implement **supervised learning** because they require the answers and data so they can generate rules. We evaluate different supervised learning methods in Chapter 4.1 and Chapter 5.3. Because these specific supervised learning methods predict the class to which a record belongs, we refer to these machine learning algorithms as **classifiers**.

We have two metrics for selecting the machine learning method that performs best for our problem space. Our first metric for evaluating supervised learning is the **F-measure** or F_1 **score** as shown in Equation 5. Here TP is the number of **true positive** responses – those records that the classifier correctly predicted to belong to a given class. FP is the number of **false positive** responses – those records that the classifier incorrectly predicted to belong to a given class. TN represents the number of **true negative** responses – those records that the classifier correctly predicted as not belonging to a given class.

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{5}$$

Accuracy, shown in Equation 6 is an additional measure we can use to evaluate classifiers. In addition to the measures applied by the F_1 score, Accuracy also incorporates the number of **false negative** responses – those records that do belong to the class, but that the classifier incorrectly predicted as not belonging to that class. We apply Equations 5 and 6 in Chapters 4.1 and 4.2.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \tag{6}$$

Alternatively, methods exist to discover structure in data, referred to as **unsupervised learning**. The unsupervised learning technique we employ in this dissertation is **clustering**, with such algorithms as *k*-means clustering and DBSCAN. K-means clustering converts the features of each record into a set of coordinates and calculates the center coordinates (**centroid**) of each group of records. It assigns all records to *k* clusters depending on their distances from these centroids.

DBSCAN also converts each record into a set of coordinates. It establishes some records as **core points** and applies a minimum distance (ε) to consider other records as belonging to the same cluster as that point. DBSCAN considers points that are too distant from any core point as **noise points** and does not assign them to a cluster. Thus, DBSCAN could help identify outliers when clustering our data.

We apply clustering in the algorithms we detail in Chapter 4.3 and then compared those algorithms' performance to search engine results in Chapter 4.4.

Topic Modeling

Topic modeling produces clusters of documents from a corpus. Through topic modeling, users can find related documents in a corpus. The top-ranked words and their weights within each cluster typically identify the cluster. Here we introduce two topic modeling techniques, both based on the bag-of-words model.

Latent Semantic Analysis (LSA) [75], also known as **Latent Semantic Indexing (LSI)**, provides a method of clustering documents in a collection based on their topics. LSI takes the documents in a collection and converts them into a matrix of document vectors. It then uses singular value decomposition [105] on that matrix to generate three new matrices. The last of these matrices contains a set of documents with coordinates. Using these coordinates and some desired threshold, we can identify clusters of documents and thus establish topics of related words for the corpus. LSI assumes that each word corresponds to a single topic, which is often not the case.

Latent Dirichlet Allocation (LDA) [36] provides an alternative. LDA assumes that documents are distributions of topics. It also assumes that each topic is a distribution of words. LDA starts with a matrix of document vectors. It then generates two additional matrices, one matrix associating topics with words and a second matrix associating topics with documents. LDA tries to adjust the values in these matrices through a series of iterations. Unlike LSI, we do not get

The continued value of web archive collections comes when others like Natasha, Rustam, Olayinka, Elbert, and Ling can reuse these same collections for their work.

Fig. 40. An example sentence from Chapter 1 with the named entities colored blue.

separate, distinct clusters for each document set, but we do get a better match of documents to topics.

Both LSI and LDA require that the end-user supply the number of topics desired before starting. Different corpora may require different numbers of topics to be specified. As we show in Chapter 4.2, we have used LSI to find off-topic mementos in a collection, and we had to try with a range of values for the number of topics. Each topic is a cluster of documents. In Chapter 4.4 we apply LDA to cluster the documents in a collection. We apply the python library gensim [341, 342] for its implementations of LSI and LDA.

Named Entities

An important part of understanding an event or topic is knowing what people, places, and organizations are involved. Figure 40 contains a sentence from Chapter 1 with the named entities colored blue. **Named-entity Recognition** (NER) extracts named entities from documents. NER works by breaking a document into tokens. These tokens are then processed to discover specific concepts like people, places, organizations, datetimes, monetary amounts, and percentages. These tokens may consist of single words (**unigrams**) or multiple words (**n-grams**). Using the n-grams from these tokens, we can infer meaning about a document or a collection that is not possible by merely evaluating its bag of words. For example, Nanni et al. [232] have used Wikipedia to learn relationships between entities and predefined aspects in order to understand a topic.

We are not attempting to improve NER. It can be accomplished via work by hand-crafted rules [175], and machine learning methods [97]. In our work, we will rely upon existing NER systems such as spaCy[21].

Generating Queries for Search Engines

In our previous examples, we started with a document and derived aboutness from it so humans could understand some aspect of the document or determine if two documents were similar. We

[21] `https://github.com/explosion/spaCy`

can also use the document's aboutness to evaluate other systems, like search engines. If we generate a good enough query, not only does it convey some aboutness for the document, but we can apply it to evaluate how easy a document is to retrieve from information retrieval system, that document's **retrievability**. We measure retrievability, as defined by Azzopardi and Vinay [25], with equation the following equation:

$$r(d) = \sum_{q \in Q} o_q \cdot f(k_{dq}, c) \tag{7}$$

where $r(d)$ is the retrievability of document d, q is a query from the set of queries Q, o_q is a weight placed on q, k_{dq} is the rank of document d for query q, and $f(k_{dq}, c)$ is a cost function that returns 1 if $k_{dq} \leq c$ and 0 otherwise.

A document's retrievability tells us how easy a user can find it given a set of queries, but how do we generate queries? Efforts like doc2query-T5 [236] apply neural networking to approximate which queries a naive user would submit to a search engine to find a given document. This method generates queries like "What happened at Virginia Tech on April 16?" Alternatively, power users have a better understanding of how search engines respond, and we can approximate their queries through **lexical signatures** initially developed by Phelps and Wilensky [265]. If we remove stop words, then the words with the highest TF-IDF score become our lexical signature. Phelps and Wilensky noted that a lexical signature of five words was sufficient to find a known document with a search engine. Park et al. [258] later verified this result while examining several different methods of generating lexical signatures, with TF-IDF performing best. Klein et al. [182] explored lexical signatures of length seven and found that they worked slightly better with the now-defunct Yahoo BOSS search API[22]. In that same work, Klein et al. also discovered that titles work remarkably well as search terms, retrieving a given document as well or better than lexical signatures.

These three concepts simulate different types of users searching for a document. The doc2query-T5 method tries to simulate a naive user searching for information through questions. The lexical signatures can simulate a power user searching for information by keywords. Finally, titles as queries can simulate a user searching for a specific document. We apply these methods in an experiment in Chapter 4.4.

2.7 EVALUATING UNDERSTANDING

RQ3 centers on which visualization is best for understanding. In this section we cover different

[22] http://boss.yahoo.com/

models of understanding and discuss how user studies have been employed in the past to evaluate understanding.

2.7.1 MODELS OF UNDERSTANDING

News stories, legal investigations, and researchers try to answer various questions about events and concepts by using a basic set of questions. Originally developed by Aristotle [287] in 340 BCE, they have been known throughout the years as “elements of circumstances”, “the Circumstances”, *Septem Circumstantiae*, or “Five Ws”. This concept of understanding of an event or concept relies upon addressing the questions of

- *who*
- *what*
- *when*
- *where*
- *why*
- (sometimes) *how*

The Circumstances are of interest to us because the questions of *who*, *when*, *where*, and, to some degree, *what* correspond to named entities that we can extract using natural language processing. This model of conveying understanding for a concept gives us some ideas on how we might build our stories and how we might evaluate them. For automated evaluations of understanding, named entities are thus more important to our work than “normal” words because named entities can help us detect when documents address the Circumstances. We apply named entities in Chapters 4.3, 4.4, and 5.2.

More recently, Pirolli and Card [266] developed information foraging theory to describe how users seek information on the web. Information foraging draws parallels between how humans search for information and how animals search for food. Those looking for information are drawn to visual cues that allow them to choose one resource over another. If these visual cues help the user find the information they are looking for, the cues are said to have a good information scent. As part of this work, we are not providing the user with a search interface. Archive-It already has this. Information foraging provides us with exciting ideas on how users select the items they view, which helps us better select our exemplars and visualize them as surrogates.

Bloom et al. [37] developed a model for classifying objectives of learning. This model was updated in 2001 by Anderson et al. [18] in 2001. It was designed for educators to help them discuss and understand the processes by which students acquire knowledge. With these concepts in mind, an educator can develop ideas for evaluating the student. Kelly et al. [178] has successfully applied this taxonomy by Anderson et al. in evaluations of information retrieval systems. Because we are evaluating our systems for understanding, our user study in Chapter 5.2 applies these methods.

2.7.2 STUDYING USERS

Kelly [177] performed extensive work evaluating information systems with user studies. They focused on evaluating interactive information retrieval (IIR) systems, but her work has applications for other user studies. We must solve several problems when trying to conduct user studies. We must identify the variables to be measured. These variables are a method of providing us a proxy for some aspect of a user's understanding. In aggregate, the values of these measurements help us address a hypothesis. To extract our measurements, we will need a battery of targets against which to evaluate our participants. User tasks fill this role. In IR experiments, the researcher gives a participant a search query or a general search question as a task. The researcher then evaluates the participant's actions to determine the performance of the system under test.

User Tasks

Tasks have several definitions. According to Vakkari et al. [310], a task is an "activity to be performed in order to accomplish a goal". Byström and Hansen [48] define different aspects of tasks. They identify two views of tasks — the task description, which specifies the goal of the task, and the task process, which focuses on the steps necessary to complete an item of work. According to them, "a task is seen as a set of (physical, affective, or cognitive) actions in pursuit of a certain, but not unchangeable goal."

In Kelly's work [177], she applied Anderson and Krathwohl's taxonomy of learning objectives [18, 37] to the evaluation of IIR systems. This analysis derived five different cognitive processes, target outcomes, and mental activities that apply to those using search engines. We list them here in order from least to most cognitively challenging.

Remember - Remember tasks involve the user learning a specific fact. It involves "retrieving, recognizing, and recalling relevant knowledge from long-term memory". With a remember task, the participant learns a specific fact, which requires the participant to identify particular items during the review of their search results. With the remember task, the participant only needs to recognize the fact if they encounter it in the results. Remember tasks are considered to be the least cognitively challenging.

Understand - Understand tasks requires that the participant “provide an exhaustive list of items”. Understanding involves “constructing meaning” by “summarizing, inferring, comparing, and explaining”. In this case, the participant gathers several different factors relevant to their query together, generating a new list of items from multiple sources, if necessary.

Analyze - To complete these tasks, a participant must not only compile a list, as with understand tasks, but also describe how the elements of these lists are similar to, different from, or otherwise related to each other. The relationships between the facts gathered during their search become a new body of knowledge due to their analysis.

Evaluate - Remember tasks establish facts, understand tasks provide a list of items to convey knowledge about the query, and analyze tasks describe these items, their relationships, and their similarities and differences. In contrast, evaluate tasks build upon the analysis by having the participant form a recommendation based on the knowledge they gained.

Create - After establishing facts, generating a list of items to establish knowledge, understanding the relationship between these items, and forming recommendations, the user then creates a plan of how to tackle a particular problem. Create is the most cognitively challenging process on Kelly’s list.

In Chapter 5.2, we report the results of a user study that evaluates a user’s ability to learn specific facts (remember) and construct meaning (understand) with different surrogate types.

Recruiting Participants

Once we have our user tasks and our methods of measurement, we must recruit participants. Many studies have solved the problem of recruitment by using Amazon’s Mechanical Turk (MT)[23]. MT is a service provided by Amazon which allows participants to apply to complete tasks in exchange for money. It has become a popular online location for recruiting subjects for research studies. Its success has led to the development of additional crowdwork platforms like Figure-Eight (formerly Crowdflower)[24], and Prolific Academic[25] [262]. MT is by far the largest source of study participants, with Amazon reporting over 500,000 available participants.

MT is an automated system that facilitates the interaction of two actors: the **requester** and the **worker**. A worker signs up for an Amazon account and must wait a few days to be approved.

[23] https://www.mturk.com

[24] https://www.figure-eight.com

[25] https://prolific.ac

Once approved, MT displays the work available to be performed to the worker. Workers are the equivalents of subjects or participants found in research studies. A Human Interface Task (HIT) is an MT assignment. Workers perform HITs for anywhere from $0.01 up to $5.00 or more. They earn as much as $50 per week from these HITs. Unless discussing the internals of MT specifically, we will use the term **participant** to refer to those who evaluate our visualizations.

Requesters are the creators of HITs. When a worker completes a HIT, the requester decides whether or not to accept the HIT. If they accept the HIT, then the requestor pays the worker. Requesters use the MT interface to specify the amount to be paid for a HIT, how many unique workers per HIT, how much time to allot to workers, and when the HIT will no longer be available for work (expire). Also, requesters can specify that they only want workers with specific qualifications. The Master Qualification is assigned automatically by the MT system based on the behavior of the workers. Requesters can also specify that they only want workers with certain acceptance levels.

The HITs themselves are HTML forms entered into the MT system. Requesters have much freedom within the interface to design HITs to meet their needs, even including JavaScript for form validation, if necessary. Once the requester has entered the HTML into the system, they can preview the HIT to ensure that it looks and responds as expected. Once the requester is done creating the HIT, they can then save it for use. HITs may also contain variables for links to visualizations or other external information.

When the requester is ready to publish a HIT for workers to perform, they can submit a CSV file containing the values for these variables. MT will create one HIT per row in the CSV file. Amazon will require that the requester put enough money into their account to pay for the number of HITs they have specified. Once the requester pays for the HITs, workers can see the HIT and then begin their submissions. The requester then reviews the submissions as they come in and pays workers. We serve as a requester in the user study we detail in Chapter 5.2.

2.8 SUMMARY

In this chapter, we identified core concepts to understanding the problem space and our solution in social media storytelling. We gave an introduction to the web and web archives. We highlighted Archive-It, the web archive collection platform we will focus on in this research, but also note that our work is not oly applicable to Archive-It. From there we introduced how web page surrogates help users understand the content of web pages without having to load them. We built upon this by detailing the different forms of storytelling and how our visualization of surrogates fits in with those forms.

The latter part of the chapter discussed traditional methods of conveying aboutness in libraries. We highlighted some of the ways machines might provide aboutness for documents and corpora. We discussed some models of understanding and how we can make use of these models for both the development and evaluation of our approaches. We identified some fundamental concepts of user studies because they play a role in evaluating which surrogate is best for understanding web archive collections.

The next chapter discusses the work others have done on different processes in our storytelling model.

CHAPTER 3

RELATED WORK

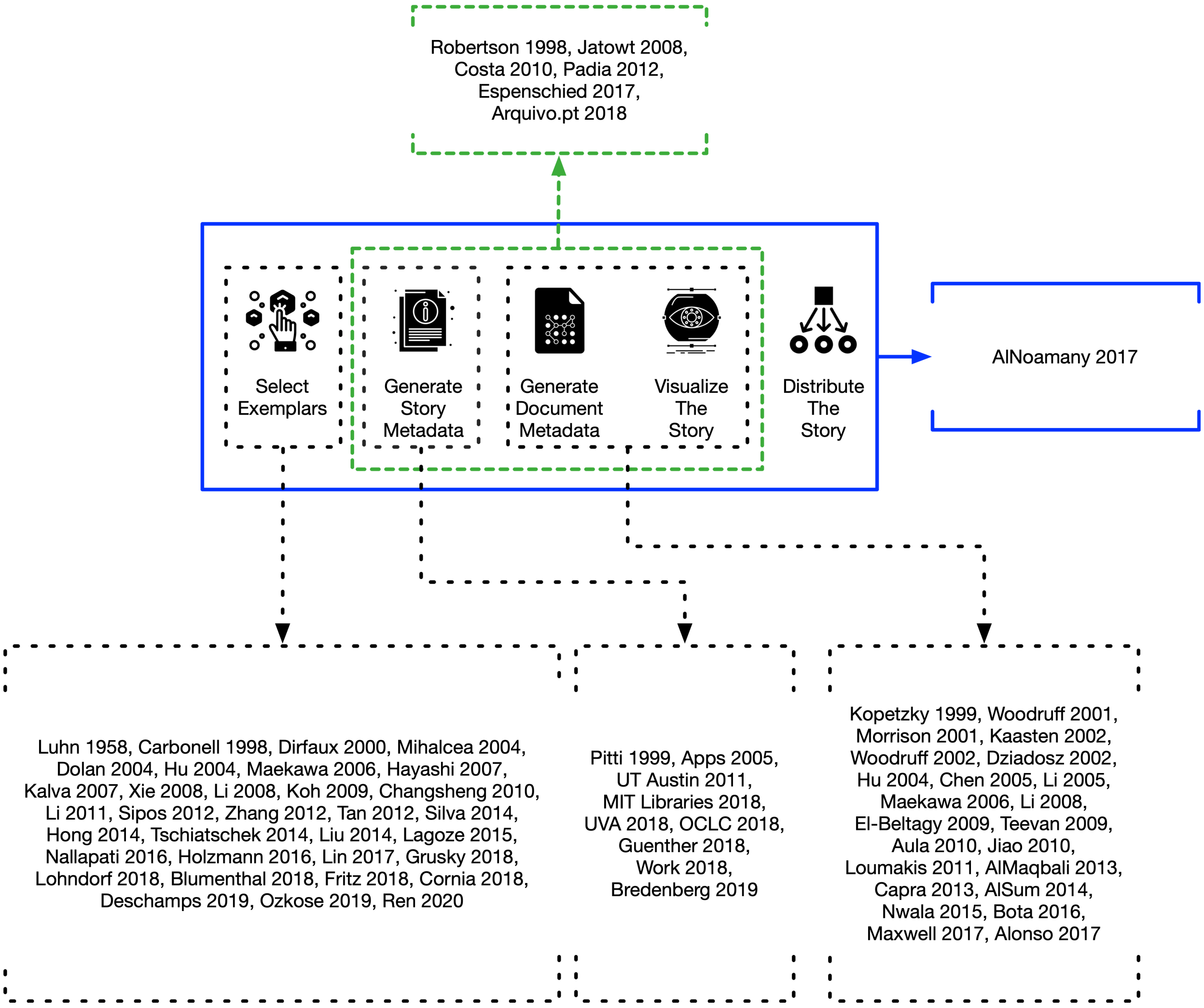

Robertson 1998, Jatowt 2008,
Costa 2010, Padia 2012,
Espenschied 2017,
Arquivo.pt 2018
Select
Exemplars
Generate
Story
Metadata
Generate
Document
Metadata
Visualize
The
Story
Distribute
The
Story
AlNoamany 2017
Luhn 1958, Carbonell 1998, Dirfaux 2000, Mihalcea 2004,
Dolan 2004, Hu 2004, Maekawa 2006, Hayashi 2007,
Kalva 2007, Xie 2008, Li 2008, Koh 2009, Changsheng 2010,
Li 2011, Sipos 2012, Zhang 2012, Tan 2012, Silva 2014,
Hong 2014, Tschiatschek 2014, Liu 2014, Lagoze 2015,
Nallapati 2016, Holzmann 2016, Lin 2017, Grusky 2018,
Lohndorf 2018, Blumenthal 2018, Fritz 2018, Cornia 2018,
Deschamps 2019, Ozkose 2019, Ren 2020
Pitti 1999, Apps 2005,
UT Austin 2011,
MIT Libraries 2018,
UVA 2018, OCLC 2018,
Guenther 2018,
Work 2018,
Bredenberg 2019
Kopetzky 1999, Woodruff 2001,
Morrison 2001, Kaasten 2002,
Woodruff 2002, Dziadosz 2002,
Hu 2004, Chen 2005, Li 2005,
Maekawa 2006, Li 2008,
El-Beltagy 2009, Teevan 2009,
Aula 2010, Jiao 2010,
Loumakis 2011, AlMaqbali 2013,
Capra 2013, AlSum 2014,
Nwala 2015, Bota 2016,
Maxwell 2017, Alonso 2017

Fig. 41. How the related work in this chapter maps to our storytelling model.

Here we highlight work similar to ours (Figure 41), identifying how our work is different or how our work can build upon existing efforts. There have been many attempts to address different parts of our storytelling model. Selecting exemplars is a problem reaching into areas of text summarization and computer vision. Generating story and document metadata has been addressed by librarians for centuries, and we will highlight how they have tried to bring these processes to web archives. As mentioned in the last chapter, surrogates are also an old concept

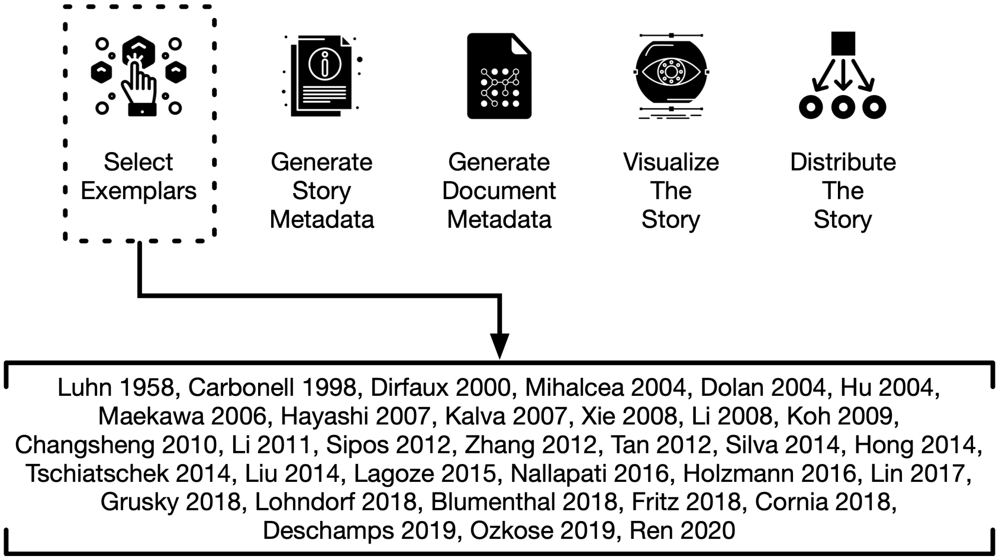

Select Exemplars
Generate Story Metadata
Generate Document Metadata
Visualize The Story
Distribute The Story
Luhn 1958, Carbonell 1998, Dirfaux 2000, Mihalcea 2004, Dolan 2004, Hu 2004,
Maekawa 2006, Hayashi 2007, Kalva 2007, Xie 2008, Li 2008, Koh 2009,
Changsheng 2010, Li 2011, Sipos 2012, Zhang 2012, Tan 2012, Silva 2014, Hong 2014,
Tschiatschek 2014, Liu 2014, Lagoze 2015, Nallapati 2016, Holzmann 2016, Lin 2017,
Grusky 2018, Lohndorf 2018, Blumenthal 2018, Fritz 2018, Cornia 2018,
Deschamps 2019, Ozkose 2019, Ren 2020

Fig. 42. Other work on selecting exemplars inspires methods for addressing this process in our storytelling model.

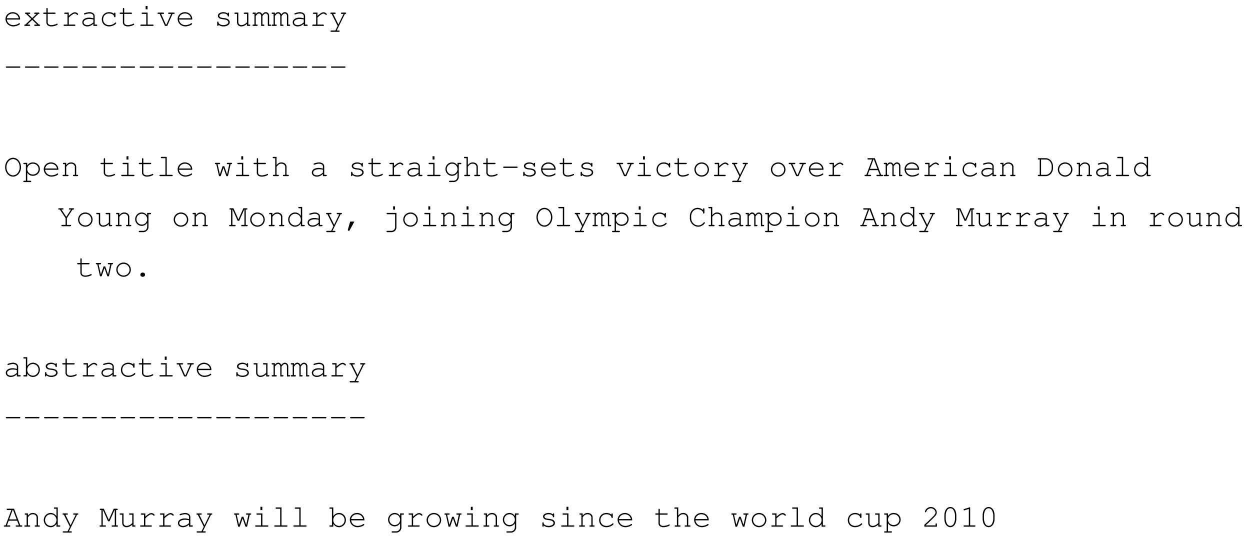
now rendered electronically by the web. Here we will discuss the work done since the dawn of the web to evaluate different surrogates' effectiveness. From there, we move to the process of visualizing collections of web resources and eventually web archive collections. We close this chapter with AlNoamany's work, the first to combine web archives and social media storytelling to summarize web archive collections and the first to use our entire storytelling model.

3.1 SELECTING EXEMPLARS

Storytelling with web archives involves several layers of selecting exemplars. In the storytelling process outlined in Chapter 1, we are referring to choosing exemplar mementos from a web archive collection. However, it does not end there. In Chapter 5.3 we cover creating social cards when no metadata exists to help us. There we select exemplar sentences from text to generate descriptions, and we select exemplar images from the document to generate a striking image. This section – summarized by Figure 42 – covers the work related to selecting exemplars and connects it to work in upcoming chapters.

3.1.1 AUTOMATIC TEXTUAL SUMMARIZATION

We start with text summarization. Most summarization algorithms fall into two categories: **abstractive** and **extractive**, with examples of each in Figure 43. Abstractive summaries contain



```
extractive summary
------------------

Open title with a straight-sets victory over American Donald
   Young on Monday, joining Olympic Champion Andy Murray in round
    two.

abstractive summary
-------------------

Andy Murray will be growing since the world cup 2010
```

Fig. 43. Examples of Extractive (top) vs. Abstractive (bottom) summaries from Mathur et al. [214].

some text from the source document(s) but may differ by word choice to produce text that reads better to the user. Abstractive summaries are typically used to generate headlines rather than complete summaries. Extractive summaries are produced by analyzing text, extracting sentences, and generating a summary based on these sentences. From there, a system generates a smaller document from the larger one. Our problem of selecting exemplars is most similar to these extractive summaries.

Luhn [206] pioneered the idea of automatic "literature abstracts" in 1958. With stop words removed and stemming in place, his algorithm computes the term frequency for each word in the text. Sentences are scored based on how many frequently occurring words they contain, based on the distance of the words from each other. Silva et al. [281] used Luhn's ideas to generate graphs of individual documents, as seen in Figure 44. Other sentence scoring algorithms include LexRank [83], Edmundson's algorithm [79], SumBasic [315] and TextRank [221].

TextRank is an algorithm developed by Mihalcea et al. [221] for keyword extraction and summarization. Brin and Page developed the PageRank algorithm [42] for ranking web page search results, and the PageRank algorithm inspired the authors to create TextRank. For summarization, TextRank breaks a document into its component sentences. These sentences are then compared to each other using a similarity metric such as cosine similarity. The algorithm then builds a weighted graph for the document where the sentences are the nodes, and the similarity scores are

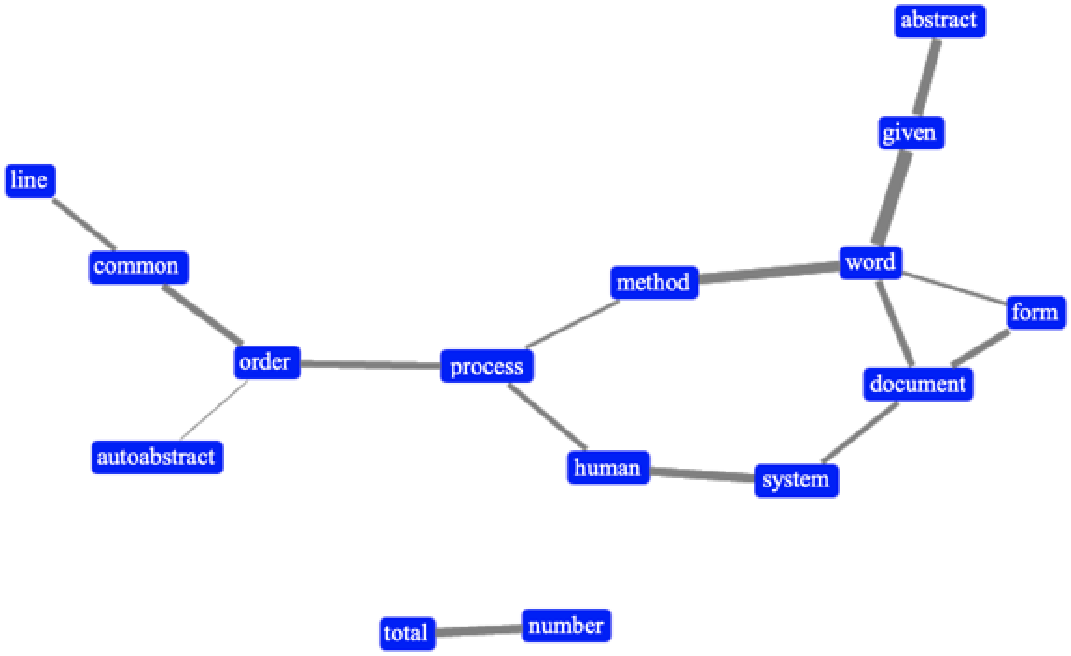


Fig. 44. An example word graph generated by the work of Silva et al. [281] derived from Lunh's algorithm [206]. (Licensed under the Creative Commons Attribution License.)

the weights of the edges. The nodes on the graph are then ranked using Equation 8.

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{8}$$

The TextRank equation is different from PageRank because it takes into account edge weights. In the equation V_i and V_j are nodes, *In* represents the set of all nodes with edges connecting *to* a given node, *Out* represents the set of all nodes reach *from* a given node, and w_{ji} is the weight of the edge from node V_i to V_j. Once sentences are ranked using this equation, the top n (typically $5 \leq n \leq 20$) are chosen for the summarization. TextRank is one of many algorithms that choose exemplar sentences by scoring sentences, but this concept of scoring can be applied to documents as well. We are inspired by this concept and apply various scoring techniques to select exemplar mementos our algorithmic primitives discussed in Chapter 4.3 and realized in the tool Hypercane in Chapter 4.5. Our memento metadata application, MementoEmbed, will be introduced in Chapter 5.4. It can apply TextRank to generate a memento's description if the user requests it. In Chapter 5.3 we mention how many words a human generates when creating descriptions, which is a necessary argument to TextRank implementations, like summa [29].

Summarization efforts have produced several datasets consisting of documents and human-generated summaries for testing. Most recently, Grusky et al. [110] created NEWSROOM which contains 1,321,995 mementos from the Internet Archive with textual summaries extracted from the `description`, `og:description`, and `twitter:description` metadata fields present in online news articles. We apply Grusky's dataset in Chapter 5.3, but we do not repeat their experiments with text summarization. We do, however, apply the metadata from a sample of the NEWSROOM dataset toward the problem of selecting striking images.

Dolan, Quirk, and Brockett [72] evaluated different techniques for acquiring paraphrases from news articles from thousands of news sources. They compare two datasets generated from pre-clustered news articles. The first consists of sentences drawn from articles that have high similarity because they have short Levenshtein distances [192]. The second consists of phrases drawn from the first one or two sentences from each article in a cluster. The system then pairs these phrases with similar phrases in other documents within the cluster. The second dataset uses a common property of news articles: that the first one or two sentences paraphrase the article. We also apply these base concepts of clustering and scoring in Chapters 4.3, 4.4, and 4.5 to select exemplar mementos from a collection.

Choosing the first few sentences seems to perform well for news stories. This simple algorithm is named Lead-3 by Nallapati et al. [231] and Lede-3 by Gruskey et al. [110]. MementoEmbed (Chapter 5.4) applies Lede-3 by default when creating a description for a memento. In Chapter 5.3 we note how many characters, words, or sentences that a human typically selects when producing these descriptions, which is a necessary input for Lede-3.

Carbonell and Goldstein [51] introduced the concept of Maximal Marginal Relevance (MMR) for improving relevance scores for search results. MMR (Equation 9) works by comparing a document to a user-issued query as well as other documents in the results.

$$MMR \overset{\text{def}}{=} Arg \max_{D_i \in R \setminus S} [\lambda(Sim_1(D_i, Q) - (l - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j))] \tag{9}$$

This equation is iteratively executed across all documents and produces ranked results. It works on two sets of documents: R, the set of ranked documents retrieved by the system, and S, the set of documents that the MMR equation has already ranked. The terms D and Q are the document and query. This equation can use any similarity method for Sim_1 and Sim_2, meaning that its ranking method is helpful for systems that employ the Jaccard Index, Cosine Similarity, or other similarity measures. Also important to note: Sim_1 and Sim_2 do not need to be the same similarity measure. Finally, setting the value of λ allows one to optimize further the ranking of documents for their diversity and their relevance to a query. If we set λ to 1, then MMR operates like a standard ranking function. If we set λ to 0, we have approached maximum diversity, where all documents have equal ranking.

Carbonell and Goldstein suggest that MMR has applications in document and corpus summarization because it compares documents to each other. However, it requires a query in order to function. Xie and Liu [333] note how MMR can be used to rank sentences for meeting summarization. They adapt Carbonell and Goldstein's equation such that each sentence is a document in itself. In the equation, the sentence takes the role of the document, and a document vector takes the role of the query. Again, these concepts inspire us. Scoring lexical units (e.g., sentences, documents) and then selecting the highest scoring as exemplars is employed in the algorithms we discuss in Chapter 4.3.

3.1.2 SELECTING IMAGES

In Chapter 5.3 we address the problem of selecting the striking image (i.e., the exemplar image) that best summarizes a given memento. Our work differs in scale and methods from that covered in this section.

Automatic image selection has been applied to the reduction of a large set of images to a small set for building photo albums [255], selecting representative pages from historical manuscripts [62], choosing the best key frame to represent a video [70, 274], generating collages [291] and general image collection summarization [301]. Individual image selection has been applied to detecting specific categories of images, such as spam [202], advertisements [53], or landscapes [115], and coarsely identifying specific principal image subjects, such as *vehicle* or *pet* [170]. None of these solutions attempt to find the striking image that summarizes a single web page.

In 2004, Hu and Bragga [122] analyzed the front page of 25 randomly selected news sites and classified the images into seven categories. A *story image* provides a striking image for a set of news articles covering a specific story. *Preview images* provide striking images for specific articles. *Commercial images* are advertisements. *Host images* provide a photograph of an author. *Heading images* are navigational elements consisting of stylized text. *Icon logos* provide branding for the whole news source or a specific feature of the publication. *Formatting images* perform the function of shaping or arranging a page, and examples include transparent spacing images or graphical horizontal rules, largely an artifact of the limited formatting abilities of the HTML of that era. The authors manually annotated 899 images across these 25 front pages. Their SVM classifier achieved an accuracy of 92.5% when combining the discrete cosine transforms of each image with the values of its color bands and the surrounding text's properties.

In 2006, Maekawa, Hara, and Nishio [208] analyzed forty websites and categorized 3,901 images into eleven categories. They then applied a custom classifier to the problem and achieved an accuracy rate of 83.1% across categories. Maekawa's goal was to classify images so that mobile browsers could avoid the unnecessary download of images that would not display nicely on the smaller displays of mobile devices. Many of their categories are similar to Hu's. Their solution relies on easy-to-calculate features like width, height, byte size, number of colors, content type, and aspect ratio. However, they also include the number of images on the page with similar features and textual information. While their overall accuracy is 83.1%, they do poorly for specific categories, such as an F_1 score of 0.458 for identifying buttons and 0.694 for advertisements.

Li, Shi, and Zhang [197] ran an experiment in 2008 that is more similar to our work. They were interested in identifying striking images for search results. Their *dominant* image is similar to our concept of a striking image. For ground truth, they randomly sampled 3,000 documents from a dataset of pages from msn.com, mit.edu, and cnn.com. They then asked three participants to label each image as *dominant* or *non-dominant*. With this training data, they applied a custom classifier to predict the class of each image on a page. They used the features of pixel size, aspect ratio, sharpness, contrast, number of colors, categorization of photo or graphic with or without a human face, content-type, position on the page, size of image compared to the size of the page rendered in a browser, number of images larger than this image on a page, if the image came from an external site, and if the image repeats across the same web site. Finally, their classifier calculates a relevance score for each image on the page based on the user's query with an accuracy of 0.85.

Koh and Kerne [184] took a different approach. It starts by analyzing the HTML document's DOM. Next, it finds the deepest nodes in the DOM and works its way back up to discover the node that most likely contains the article's content. From the images found in this node, they choose the largest image with the smallest aspect ratio based on empirically determined thresholds. Based on human labeling, their algorithm achieves an accuracy of 0.898 and an F_1 of 0.921 across datasets of web pages consisting of 239 news pages and 254 research pages.

In Chapter 5.3, we evaluate 37,522 mementos of news articles whose authors have specified a striking image in their metadata (e.g., in the `og:image` or `twitter:image` field). Our goal is to automate selecting the same image that the author would have selected for their own page. We evaluate far more articles (37,522) than this existing work, and our ground truth comes directly from each web page's author.

3.1.3 SUMMARIZING CORPORA

Sipos et al. [286] provide an algorithm for summarizing a corpus over time by selecting the most influential documents overall, documents for each era of the corpus, authors within each era, and key phrases within each era. For their purposes, the citation structure within a corpus determines influence. Their work is similar to ours because it selects exemplar documents from a corpus and considers the corpus's temporal dimension, much like we do with web archives. Their work differs, however, because they only focused on scholarly publications.

NIST organized the Text Analysis Conference (TAC). In 2010, TAC established a guided summarization task whereby a system would accept ten news articles and generate a 100 word summary from them based on a set of predefined categories [233]. This task required that the resulting summary for each category of articles contain a set of predefined **aspects**. For example: the category *Accidents and Natural Disasters* contained aspects such as *what happened*, *reasons for accident/disaster*, *who was affected*, and *why did it occur*; and the category *investigations and trials* contained aspects such as *who is the defendant*, *who is the prosecution*, *how did the defendant plead*, and *what was the outcome, including sentencing*. The concept of aspects maps to the *Septem Circumstantiae* mentioned in Chapter 2.7.1 as a model of understanding. Here we present two studies that used the dataset that came out of that workshop.

Li et al. [194] developed an unsupervised approach to mine document collections for these aspects. Their goal was to help generate summaries using these subtopics to answer specific questions about a named entity. Much like our situation, they had existing document collections built around a central theme. They extend LDA (Chapter 2) in order to cluster sentences and words into aspects, creating the entity-aspect model. The entity-aspect model develops probability distributions to assign words to three categories: background word, document word, or aspect word. At the end of this process, the system clusters sentences into aspects of the collection.

They later extended their work [195] and developed an approach to surface aspects of a document collection. Their updated approach has the following steps:

1. identify the aspects of the corpus using an improved version of the entity-aspect model named the event-aspect model

2. cluster the sentences of the corpus around these aspects

3. rank the sentences using the LexRank algorithm [83]

4. compress the sentences to improve their quality

5. select sentences from each cluster

Zhang et al. [338] sought to create an algorithm that provides an improvement over the work presented by Li et al. [195]. They break a given corpus into sentences. The system processes each sentence using part-of-speech tagging and named entity recognition to produce a set of syntactico-semantic patterns. These patterns are then classified into aspects using binary decomposition [302] and semi-supervised learning via a transductive SVM classifier [135]. After this, the authors cluster the sentences by expressing the topics as states and the sentences as observed sequences in a Hidden Markov Model [30]. Sentences are then ranked based on their frequency in the corpus and their aspect score, based on the value calculated by the decision function trained from transductive SVM. The authors annotated the TAC 2010 and TAC 2011 datasets for training and development to train their system.

Hong and Nenkova [120] evaluated different methods of weighting words in a corpus in order to produce better summaries. They consider the following weighting metrics for words:

- word probability, a normalized term frequency based on the total number of words in the input
- log-likelihood ratio, using a large background corpus to identify how likely a word is to occur in a single document or sub-corpus
- a process using the TextRank equation, but where the edges are weighted based on the syntactic dependency between two words, as determined by the Stanford dependency parser [67]

They then introduce a greedy algorithm that weighs all sentences based on the weights of the words. This algorithm chooses the highest ranked sentences and removes highly similar sentences. Comparatively, they use a logistic regression model trained on the DUC 2003 and DUC 2004 datasets [71] for selecting the highest ranked sentences using all of these weights for each word as features. Based on Rouge-N [199] scores on DUC 2003 and DUC 2004 and comparison with other summarizing systems, they determine that their logistic regression model system is comparable, even though it applies relatively simple techniques. The greedy algorithm scored highest when using the TextRank solution alone. Their work implies that even simple techniques may be sufficient for selecting exemplars.

There are key concepts of clustering, scoring, and filtering in each of these cases to acquire the highest scoring documents or sentences from each collection. While we are selecting exemplar mementos and applying the structural features of collections (Chapter 4.1) to do so, we recognize that the repetition of these same algorithmic primitives across solutions indicates a set of possibilities that should work well with web archive collections. We discuss and implement algorithms applying these primitives of filtering, clustering, scoring in Chapter 4.3, 4.4, and 4.5. To support different types of stories (e.g., fixed page sliding time (FPST), sliding page fixed time (SPFT)), we provide users like Rustam and Olayinka with the ability to inform the exemplar selection process with their own aspects (e.g., original resource, time period, keywords).

3.1.4 SELECTING EXEMPLARS FROM WEB ARCHIVE COLLECTIONS

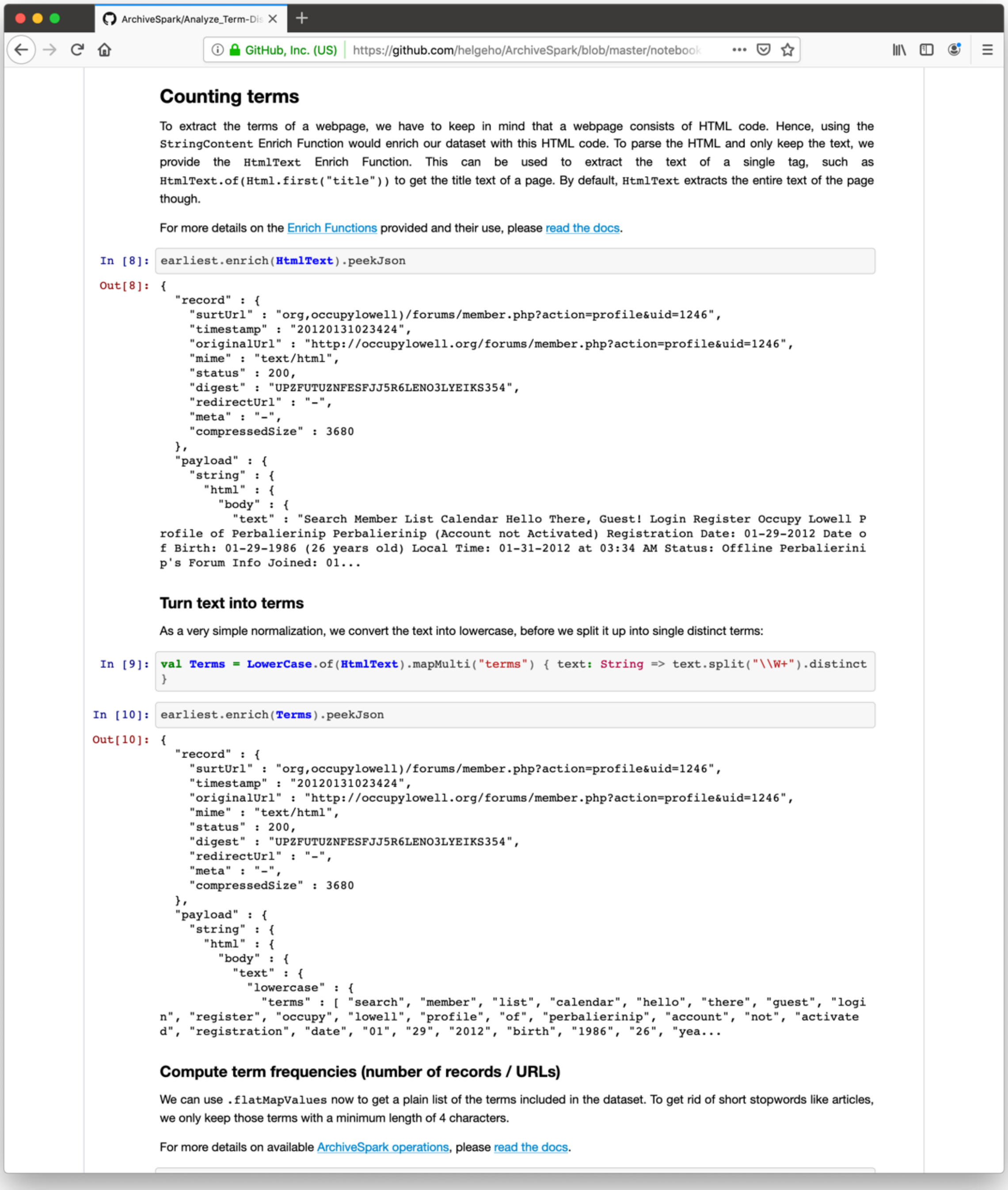
For those users who have access to the WARCs of their collections, a few options exist for exploring those collections and extracting data from them. Holzmann et al. [119] developed the Scala application ArchiveSpark[1]. Rather than reading all WARCs for every query, ArchiveSpark optimizes the ability to query WARCs by first running user queries through CDX files. Users can write their own applications to use the ArchiveSpark API, query the data directly from within the interactive terminal interface offered by Scala, or interact with the data using a Jupyter notebook as shown in Figure 45.

ArchiveSpark offers a variety of operations for extracting data from a web archive corpus. Some examples include:

- extracting titles from mementos
- calculating term distributions from the collection
- discovering named entities
- building a graph of outgoing hyperlinks
- removing duplicate mementos
- extracting and saving all images from the collection

ArchiveSpark is a framework supporting many different operations that can be combined; for example, one can filter a collection by the domain name and then calculate term distributions. ArchiveSpark typically outputs a JSON file consisting of the desired data. It is up to the end-user to further process this data in the desired way. Natively, ArchiveSpark has no visualization capability, but it allows users to filter and explore their WARCs to produce their visualizations with a third-party tool.

[1] `https://github.com/helgeho/ArchiveSpark`


ArchiveSpark/Analyze_Term-Dis
GitHub, Inc. (US)
https://github.com/helgeho/ArchiveSpark/blob/master/notebook

Counting terms

To extract the terms of a webpage, we have to keep in mind that a webpage consists of HTML code. Hence, using the `StringContent` Enrich Function would enrich our dataset with this HTML code. To parse the HTML and only keep the text, we provide the `HtmlText` Enrich Function. This can be used to extract the text of a single tag, such as `HtmlText.of(Html.first("title"))` to get the title text of a page. By default, `HtmlText` extracts the entire text of the page though.

For more details on the Enrich Functions provided and their use, please read the docs.

```
In [8]: earliest.enrich(HtmlText).peekJson
```

```
Out[8]: {
          "record" : {
            "surtUrl" : "org,occupylowell)/forums/member.php?action=profile&uid=1246",
            "timestamp" : "20120131023424",
            "originalUrl" : "http://occupylowell.org/forums/member.php?action=profile&uid=1246",
            "mime" : "text/html",
            "status" : 200,
            "digest" : "UPZFUTUZNFESFJJ5R6LENO3LYEIKS354",
            "redirectUrl" : "-",
            "meta" : "-",
            "compressedSize" : 3680
          },
          "payload" : {
            "string" : {
              "html" : {
                "body" : {
                  "text" : "Search Member List Calendar Hello There, Guest! Login Register Occupy Lowell P
        rofile of Perbalierinip Perbalierinip (Account not Activated) Registration Date: 01-29-2012 Date o
        f Birth: 01-29-1986 (26 years old) Local Time: 01-31-2012 at 03:34 AM Status: Offline Perbalierini
        p's Forum Info Joined: 01...
```

Turn text into terms

As a very simple normalization, we convert the text into lowercase, before we split it up into single distinct terms:

```
In [9]: val Terms = LowerCase.of(HtmlText).mapMulti("terms") { text: String => text.split("\\W+").distinct
        }
```

```
In [10]: earliest.enrich(Terms).peekJson
```

```
Out[10]: {
           "record" : {
             "surtUrl" : "org,occupylowell)/forums/member.php?action=profile&uid=1246",
             "timestamp" : "20120131023424",
             "originalUrl" : "http://occupylowell.org/forums/member.php?action=profile&uid=1246",
             "mime" : "text/html",
             "status" : 200,
             "digest" : "UPZFUTUZNFESFJJ5R6LENO3LYEIKS354",
             "redirectUrl" : "-",
             "meta" : "-",
             "compressedSize" : 3680
           },
           "payload" : {
             "string" : {
               "html" : {
                 "body" : {
                   "text" : {
                     "lowercase" : {
                       "terms" : [ "search", "member", "list", "calendar", "hello", "there", "guest", "logi
         n", "register", "occupy", "lowell", "profile", "of", "perbalierinip", "account", "not", "activate
         d", "registration", "date", "01", "29", "2012", "birth", "1986", "26", "yea...
```

Compute term frequencies (number of records / URLs)

We can use `.flatMapValues` now to get a plain list of the terms included in the dataset. To get rid of short stopwords like articles, we only keep those terms with a minimum length of 4 characters.

For more details on available ArchiveSpark operations, please read the docs.

Fig. 45. An example Jupyter Notebook using ArchiveSpark, rendered by GitHub and Firefox. This notebook evaluates the terms discovered in a subset of WARCs from Archive-It Collection 2950: *Occupy Movement 2011/2012*. (Screenshot taken in July 2018.)

The Archives Unleashed Project[2] [68] provides several tools for exploring collections of WARCs. It provides three applications: the Archives Unleashed Toolkit (AUT), Archives Unleashed Cloud (AUK), and Warclight.

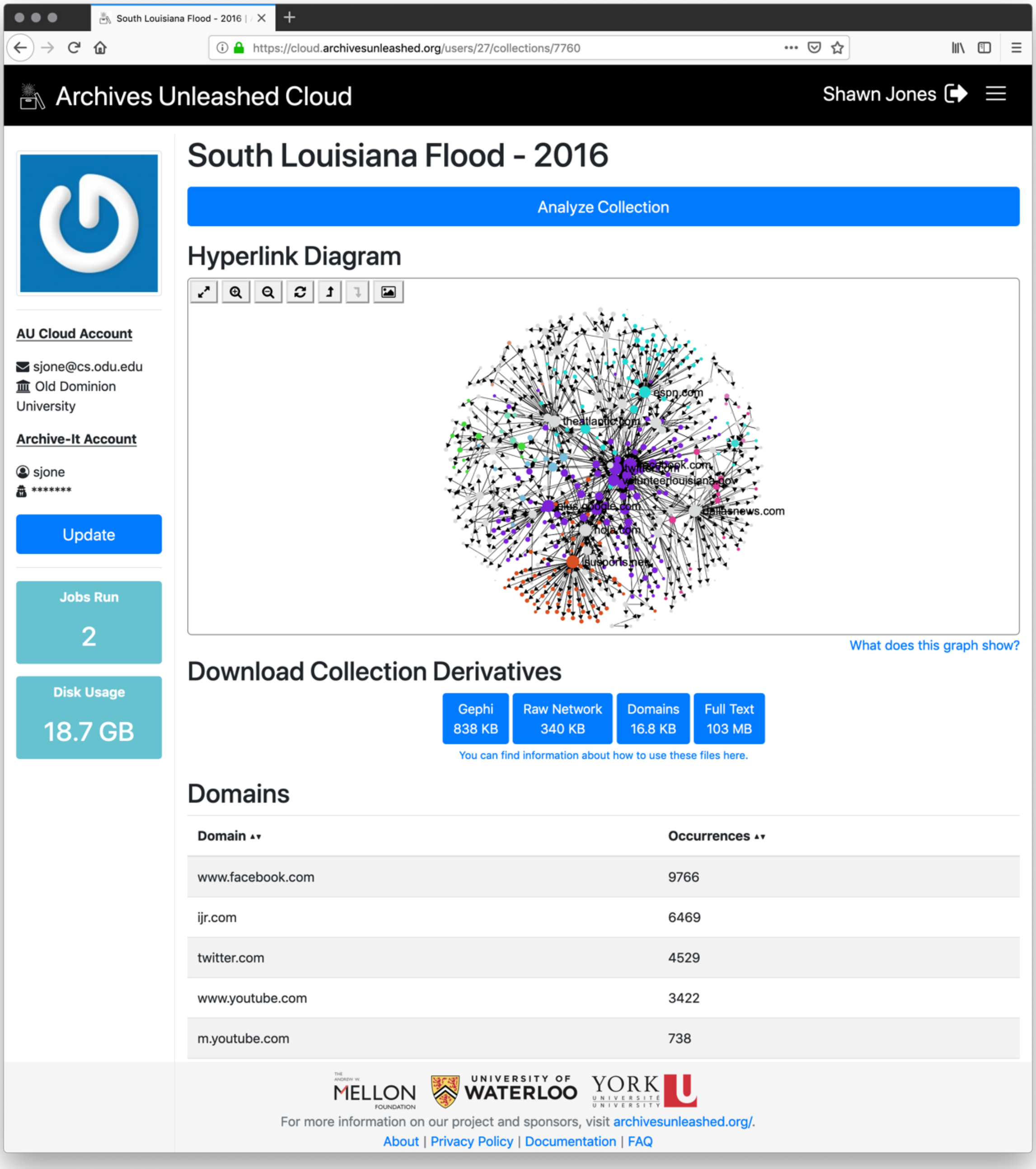
AUT is built upon Warcbase [200], by Lin et al. Warcbase is an earlier tool that provides users with the ability to query WARCs directly. Based on their work with scholars who work with web archives, Lin et al. developed guidance for those trying to work with web archives. This guidance takes the form of four main steps: filter, analyze, aggregate, and visualize (FAAV). AUT helps a user iterate through FAAV. Initially, a scholar with a web archive collection starts with a research question. From there, the scholar would proceed to filter the collection of unnecessary documents, removing duplicates and discovering pages that contain a given search term. This process creates a subcollection that the user can then analyze for named entities, frequent terms, and link graphs. Next, they need to aggregate the data from this analysis to produce statistics and relationships. Finally, the scholar can visualize these relationships to discover other concepts not visible by merely looking at the documents or statistics.

Some examples of AUT's capabilities include:

- filtering by original resource URI, search term, memento-datetime, language, domain, or URI pattern
- listing top-level domains
- extracting all boilerplate from documents
- discovering named entities
- analyzing the link structure of the collection using outgoing links between mementos
- aggregation by domain or memento-datetime
- analysis of images by domain name or MD5 hash
- processing of Tweets from Twitter

Like with ArchiveSpark, operations can be combined. Also, like ArchiveSpark, AUT does not produce visualizations, instead relying upon the user to post-process the data with a third-party tool.

[2] https://archivesunleashed.org/



South Louisiana Flood - 2016

AU Cloud Account

sjone@cs.odu.edu
Old Dominion University

Archive-It Account

sjone
****

Jobs Run 2

Disk Usage 18.7 GB

Hyperlink Diagram

What does this graph show?

Download Collection Derivatives

You can find information about how to use these files here.

Domains

Domain	Occurrences
www.facebook.com	9766
ijr.com	6469
twitter.com	4529
www.youtube.com	3422
m.youtube.com	738

For more information on our project and sponsors, visit archivesunleashed.org/.

Fig. 46. Archives Unleashed Cloud (AUK) run on Archive-It Collection 7760: *South Louisiana Flood – 2016*. (Screenshot taken in July 2018.)

AUK[3] was a cloud platform developed by the Archives Unleashed Team [99]. As seen in Figure 46, end users could import their Archive-It collection and view a Hyperlink Diagram showing the link structure of the collection. AUK provided an easy way for users to generate derivative data for common research questions. The first file provided a link graph for use with Gephi[4]. Gephi allows the end-user to manipulate the nodes and edges in graphs. The second file was a link graph but saved to the standard GraphML file format[5]. The third file contained a distribution of all domain names in the collection. The final file contained the full text of all mementos, allowing further analysis with third-party text analysis tools, like `grep` or Voyant Tools[6]. Armed with AUK and additional tools, like the Archives Unleashed Notebooks[7], a user could answer many questions about a web archive collection if they had access to the WARCs. As of June 2021, AUK has been integrated into Archive-It [297].

Warclight[8] allows a user to query a web archive collection that they previously indexed with the Web Archive Discovery tools' warc-indexer [132] from the UK Web Archive. As seen in Figure 47, Warclight allows the user to drill down into a collection by content type, year of crawl, domain, language, resource name, collecting institution, and collection name. From here, a user can view the pages that match their search terms. For each page, information such as the page title and the domains of outgoing links are available. The titles in the results link to a page providing metadata for the corresponding memento.

We outline primitives in Chapter 4.3 that are inspired by these tools. However, our approach differs by working with publicly accessible mementos, not WARCs. Social media storytelling requires public access to content; thus, WARC filenames and offsets are not suitable for sharing with the public because the public does not typically have access to the WARCs that make up a web archive collection. Moreover, we cannot apply ArchiveSpark, Archives Unleashed, or Warclight to the problem of selecting exemplars because we do not have access to the WARCs that make up the web archive collections we want to summarize. Additionally, because our tools are free from the constraint of WARCs and apply the Memento protocol, users like Rustam and Olayinka can create stories for public web archive collections for which they also cannot access the WARCs.

[3] `http://cloud.archivesunleashed.org/`

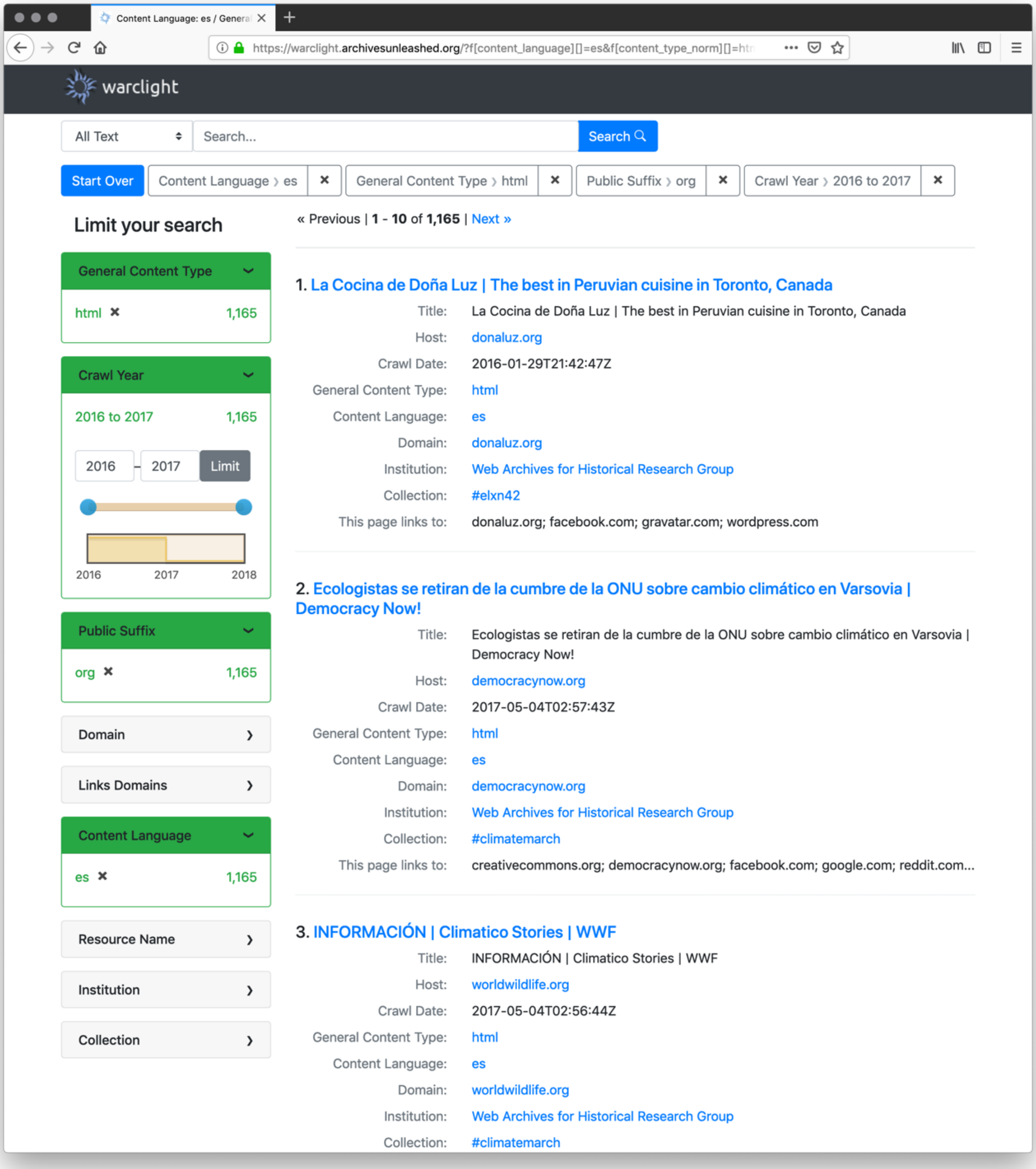
[4] `https://gephi.org/`

[5] `http://graphml.graphdrawing.org/`

[6] `https://voyant-tools.org`

[7] `https://archivesunleashed.org/notebooks/`

[8] `https://github.com/archivesunleashed/warclight`



« Previous | **1** - **10** of **1,165** | Next »

1. La Cocina de Doña Luz | The best in Peruvian cuisine in Toronto, Canada

Title:	La Cocina de Doña Luz | The best in Peruvian cuisine in Toronto, Canada
Host:	donaluz.org
Crawl Date:	2016-01-29T21:42:47Z
General Content Type:	html
Content Language:	es
Domain:	donaluz.org
Institution:	Web Archives for Historical Research Group
Collection:	#elxn42
This page links to:	donaluz.org; facebook.com; gravatar.com; wordpress.com

2. Ecologistas se retiran de la cumbre de la ONU sobre cambio climático en Varsovia | Democracy Now!

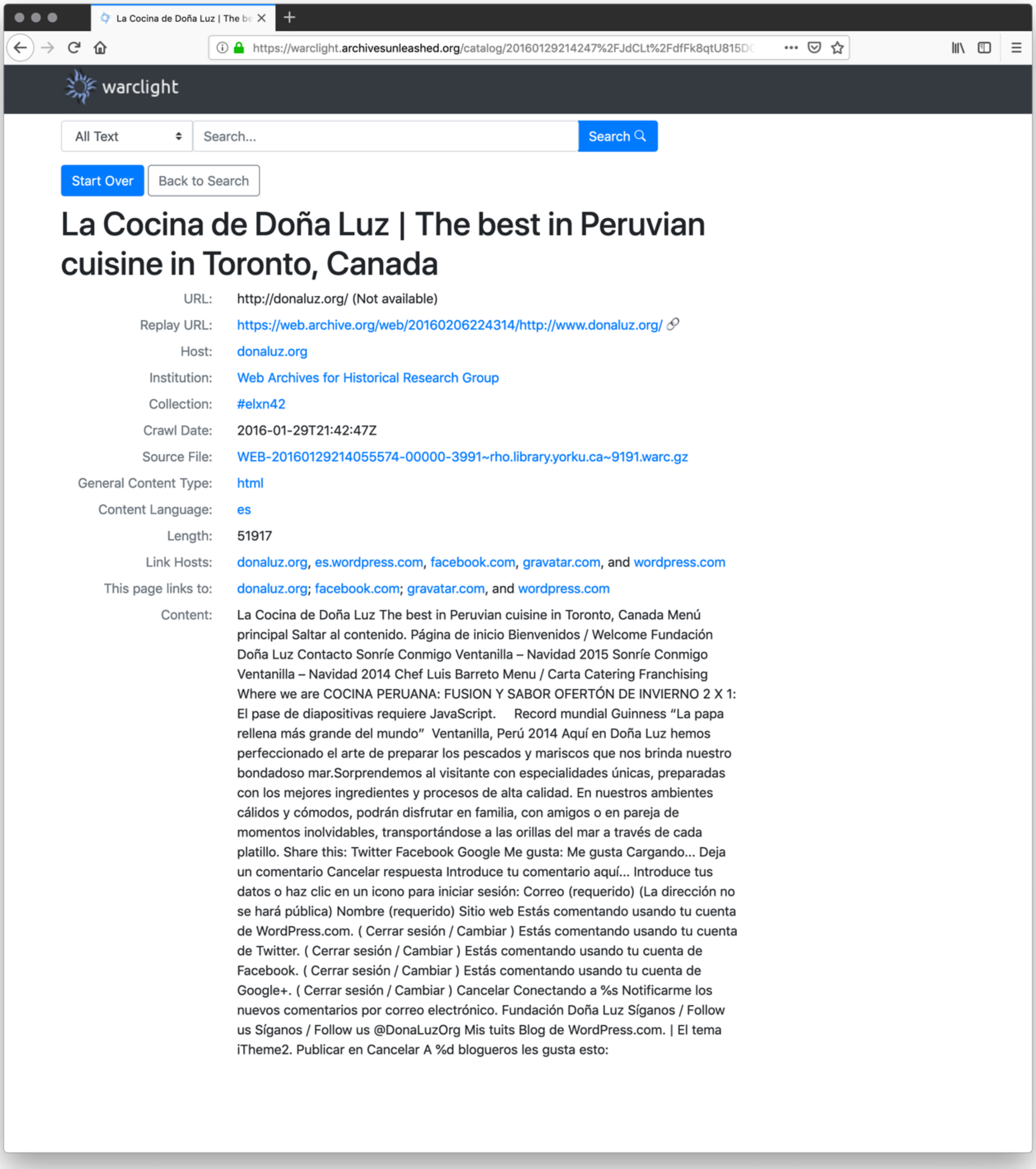
Title:	Ecologistas se retiran de la cumbre de la ONU sobre cambio climático en Varsovia | Democracy Now!
Host:	democracynow.org
Crawl Date:	2017-05-04T02:57:43Z
General Content Type:	html
Content Language:	es
Domain:	democracynow.org
Institution:	Web Archives for Historical Research Group
Collection:	#climatemarch
This page links to:	creativecommons.org; democracynow.org; facebook.com; google.com; reddit.com...

3. INFORMACIÓN | Climatico Stories | WWF

Title:	INFORMACIÓN | Climatico Stories | WWF
Host:	worldwildlife.org
Crawl Date:	2017-05-04T02:56:44Z
General Content Type:	html
Content Language:	es
Domain:	worldwildlife.org
Institution:	Web Archives for Historical Research Group
Collection:	#climatemarch

(a) Filtering Mementos

Fig. 47. Warclight allows users to explore a web archive collection. Examples from: `https://warclight.archivesunleashed.org`


La Cocina de Doña Luz | The be
https://warclight.archivesunleashed.org/catalog/20160129214247%2FJdCLt%2FdfFk8qtU815D
warclight
All Text
Search...
Search
Start Over
Back to Search

La Cocina de Doña Luz | The best in Peruvian cuisine in Toronto, Canada

URL:	http://donaluz.org/ (Not available)
Replay URL:	https://web.archive.org/web/20160206224314/http://www.donaluz.org/
Host:	donaluz.org
Institution:	Web Archives for Historical Research Group
Collection:	#elxn42
Crawl Date:	2016-01-29T21:42:47Z
Source File:	WEB-20160129214055574-00000-3991~rho.library.yorku.ca~9191.warc.gz
General Content Type:	html
Content Language:	es
Length:	51917
Link Hosts:	donaluz.org, es.wordpress.com, facebook.com, gravatar.com, and wordpress.com
This page links to:	donaluz.org; facebook.com; gravatar.com, and wordpress.com
Content:	La Cocina de Doña Luz The best in Peruvian cuisine in Toronto, Canada Menú principal Saltar al contenido. Página de inicio Bienvenidos / Welcome Fundación Doña Luz Contacto Sonríe Conmigo Ventanilla – Navidad 2015 Sonríe Conmigo Ventanilla – Navidad 2014 Chef Luis Barreto Menu / Carta Catering Franchising Where we are COCINA PERUANA: FUSION Y SABOR OFERTÓN DE INVIERNO 2 X 1: El pase de diapositivas requiere JavaScript. Record mundial Guinness "La papa rellena más grande del mundo" Ventanilla, Perú 2014 Aquí en Doña Luz hemos perfeccionado el arte de preparar los pescados y mariscos que nos brinda nuestro bondadoso mar.Sorprendemos al visitante con especialidades únicas, preparadas con los mejores ingredientes y procesos de alta calidad. En nuestros ambientes cálidos y cómodos, podrán disfrutar en familia, con amigos o en pareja de momentos inolvidables, transportándose a las orillas del mar a través de cada platillo. Share this: Twitter Facebook Google Me gusta: Me gusta Cargando... Deja un comentario Cancelar respuesta Introduce tu comentario aquí... Introduce tus datos o haz clic en un icono para iniciar sesión: Correo (requerido) (La dirección no se hará pública) Nombre (requerido) Sitio web Estás comentando usando tu cuenta de WordPress.com. (Cerrar sesión / Cambiar) Estás comentando usando tu cuenta de Twitter. (Cerrar sesión / Cambiar) Estás comentando usando tu cuenta de Facebook. (Cerrar sesión / Cambiar) Estás comentando usando tu cuenta de Google+. (Cerrar sesión / Cambiar) Cancelar Conectando a %s Notificarme los nuevos comentarios por correo electrónico. Fundación Doña Luz Síganos / Follow us Síganos / Follow us @DonaLuzOrg Mis tuits Blog de WordPress.com. | El tema iTheme2. Publicar en Cancelar A %d blogueros les gusta esto:

(b) Viewing Memento Metadata

Fig. 47. (Continued) Warclight allows users to explore a web archive collection. Examples from: `https://warclight.archivesunleashed.org`

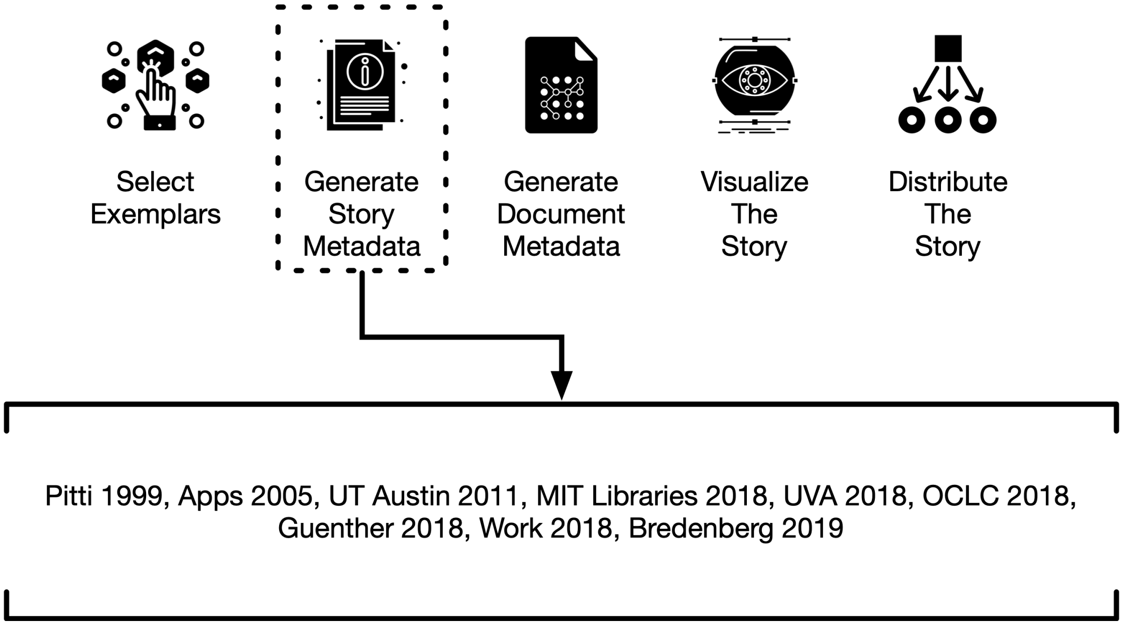

Select Exemplars
Generate Story Metadata
Generate Document Metadata
Visualize The Story
Distribute The Story
Pitti 1999, Apps 2005, UT Austin 2011, MIT Libraries 2018, UVA 2018, OCLC 2018, Guenther 2018, Work 2018, Bredenberg 2019

Fig. 48. Others have worked on generating story metadata, which is important to our storytelling model.

3.2 GENERATING STORY METADATA

We are not the first to try to generate metadata from collections (Figure 48). As noted in Chapter 2, librarians have been applying metadata to collections for centuries. We need it for our storytelling processes because it may help us select exemplars and it improves the information scent of the stories we produce.

At first glance, Encoded Archival Description (EAD) [40] appears to be a good fit for providing metadata for web archive collections. EAD provides a finding aid, or an index of the resources within a collection [308]. The creators of EAD originally conceived of it for use with special collections consisting of all kinds of documents, including personal correspondence, charts, maps, and electronic records [268]. EAD attempts to provide metadata describing the whole collection as well as each item within the collection. Figure 49 shows an example of an EAD record for a single item.

Similar solutions to EAD include the IASC Web Archiving Metadata Application Profile [223], the University of Virginia Library Web Archiving Metadata Application Profile [309], OCLC's Web Archiving Metadata Working Group[9] Descriptive Metadata for Web Archiving [73],

[9] https://www.oclc.org/research/themes/research-collections/wam.html

```
<archdesc level="collection" type="inventory" relatedencoding="
   MARC21">
<did>
    <head>Overview of the Collection</head>
    <repository encodinganalog="852$a" label="Repository: ">
       Blank University</repository>
    <origination label="Creator: ">
        <persname encodinganalog="100">Brightman, Samuel C. (
           Samuel Charles), 1911-1992</persname>
    </origination>
    <unittitle encodinganalog="245$a" label="Title: ">Samuel C.
        Brightman Papers</unittitle>
    <unitdate encodinganalog="245$f" normal="1932/1992" type="
       inclusive" label="Inclusive Dates: ">1932-1992</unitdate
       >
    <physdesc encodinganalog="300$a" label="Quantity: ">
        <extent>6 linear ft.</extent>
    </physdesc>
    <abstract encodinganalog="520$a" label="Abstract: ">
        Papers of the American journalist including some war
           correspondence,
        political and political humor writings, and adult
           education material
    </abstract>
    <unitid encodinganalog="099" label="Identification: "
       countrycode="us" repositorycode="NSyU">2458163</unitid>
    <langmaterial label="Language: " encodinganalog="546">
        <language langcode="eng">English</language>
    </langmaterial>
</did>
</archdesc>
```

Fig. 49. An example from Wikipedia [325] of an `archdesc` record in EAD.

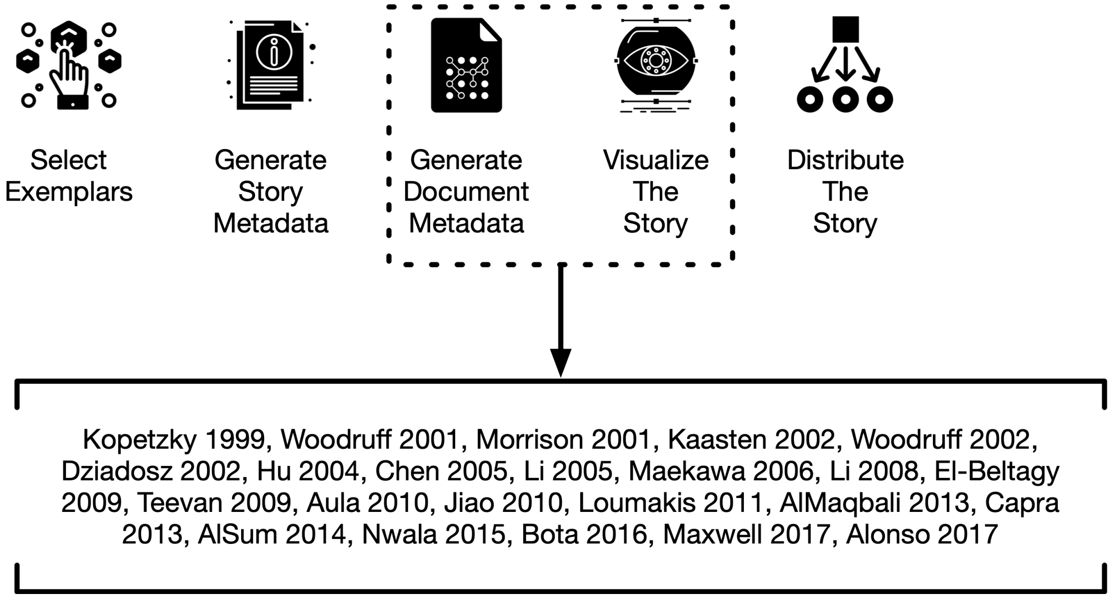

Select Exemplars
Generate Story Metadata
Generate Document Metadata
Visualize The Story
Distribute The Story
Kopetzky 1999, Woodruff 2001, Morrison 2001, Kaasten 2002, Woodruff 2002, Dziadosz 2002, Hu 2004, Chen 2005, Li 2005, Maekawa 2006, Li 2008, El-Beltagy 2009, Teevan 2009, Aula 2010, Jiao 2010, Loumakis 2011, AlMaqbali 2013, Capra 2013, AlSum 2014, Nwala 2015, Bota 2016, Maxwell 2017, Alonso 2017

Fig. 50. Studying document metadata on the web through surrogates.

and Metadata Application Profile for Description of Websites with Archived Versions [111].

Scale is the greatest challenge to implementing solutions like these for web archive collections. These standards were designed with the intention that humans would carefully review each item in the collection, gather that item's metadata using some cataloging standard, and then manually produce its record. As noted, web archive collections can consist of thousands of documents, and manually applying this metadata is an expensive proposition. In Chapter 4.5 we show how Hypercane can automatically acquire human-created metadata while also generating some metadata that can provide statistics and aboutness for a collection.

3.3 GENERATING DOCUMENT METADATA AND VISUALIZING DOCUMENTS AS SURROGATES

As we covered in Chapter 2.4 and will further discuss in Chapter 5, document metadata for a web page is often visualized as a surrogate. Through studying the different types of surrogates created for web pages, we get an idea of the possibilities of the type of document metadata to generate for storytelling. The studies in this section – visualized in Figure 50 – try to determine which surrogate works best for a user selecting a search engine result. We have a slightly different use case. Surrogates for both search engines and social media try to answer the question of "Will this meet my information need?", but search engine surrogates work to differentiate themselves from their neighbors on the same page whereas social media surrogates focus on encouraging readers to visit their page. Search engine surrogates for the same document can change with respect to a query. Social media surrogates have no query as input and are only informed by their document's contents. With this in mind, we cover the history of surrogate studies here.

3.3.1 TEXT SNIPPETS

Text snippets were one of the earliest surrogates. They require fetching a given web page and selecting the text to be used in the snippet. Many different methods exist for text selection, like those employed by El-Beltagy and Rafea [81] and Chen et al. [55]. Search engines typically use text snippets for displaying results. Figure 51 displays different search engine results for the URI `https://www.cs.odu.edu/~mweigle`. Most text snippets extract the title and some of the text of the underlying page, displaying it with the URI. Some text snippets, like the Google[10] surrogate shown in Figure 51a provide a date estimating the last modification date of the page. Bing[11] surrogates may not contain the date, but they display menu links from the page, as seen in Figure 51b. As Figure 51c shows, DuckDuckGo[12] surrogates only show the title and text, along with the URI and favicon.

Despite their age, text snippets are still undergoing research. In 2017, Maxwell et al. [216] evaluated snippet length for search queries. In Maxwell's experiment, as snippet length increased, participants "issued fewer queries, examined fewer pages; but clicked more documents, i.e., they spent more of their time assessing documents at higher ranks." Even though users preferred longer snippets, their accuracy did not improve. In Chapter 5.3 we establish the sizes of the descriptions that human authors create, indicating a desired length for text snippets independent of queries.

As a group, text snippets are listed one per row on a web page. This visualization is optimal for search results, as the position of the result conveys its relevancy. However, this format affects how many surrogates a user can view at once. Systems typically display one text snippet per row, whereas more thumbnails can fit into the same amount of space. In social media storytelling, surrogates are ordered not to convey relevance but often to convey the order of events in time. Thus, though the concept of arrangement is similar, the meaning is quite different.

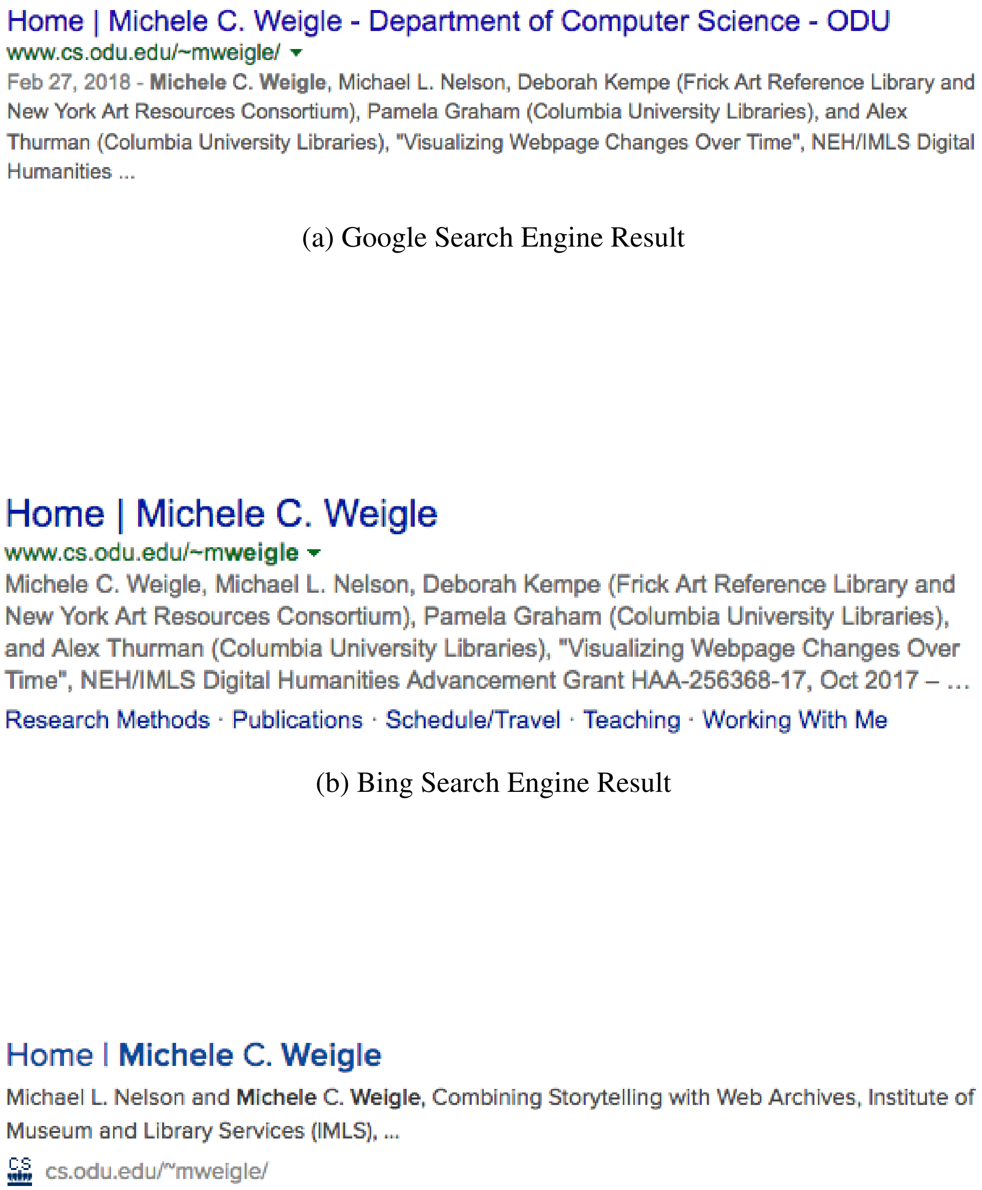
3.3.2 THUMBNAILS

A thumbnail is produced by loading a given web page in a browser and taking a screenshot of the contents of the browser window. Different systems use thumbnails in different ways. For

[10] `https://www.google.com/`

[11] `https://www.bing.com/`

[12] `https://duckduckgo.com/`



Home | Michele C. Weigle - Department of Computer Science - ODU
www.cs.odu.edu/~mweigle/ ▾
Feb 27, 2018 - **Michele C. Weigle**, Michael L. Nelson, Deborah Kempe (Frick Art Reference Library and New York Art Resources Consortium), Pamela Graham (Columbia University Libraries), and Alex Thurman (Columbia University Libraries), "Visualizing Webpage Changes Over Time", NEH/IMLS Digital Humanities ...

(a) Google Search Engine Result

Home | Michele C. Weigle
www.cs.odu.edu/~m**weigle** ▾
Michele C. Weigle, Michael L. Nelson, Deborah Kempe (Frick Art Reference Library and New York Art Resources Consortium), Pamela Graham (Columbia University Libraries), and Alex Thurman (Columbia University Libraries), "Visualizing Webpage Changes Over Time", NEH/IMLS Digital Humanities Advancement Grant HAA-256368-17, Oct 2017 – ...
Research Methods · Publications · Schedule/Travel · Teaching · Working With Me

(b) Bing Search Engine Result

Home | **Michele C. Weigle**
Michael L. Nelson and **Michele** C. **Weigle**, Combining Storytelling with Web Archives, Institute of Museum and Library Services (IMLS), ...
cs.odu.edu/~mweigle/

(c) DuckDuckGo Search Engine Result

Fig. 51. Text snippet surrogates provided by different search engines for `https://www.cs.odu.edu/~mweigle` collected on April 9, 2019. All were collected on the same date at around the same time, yet they have different content.

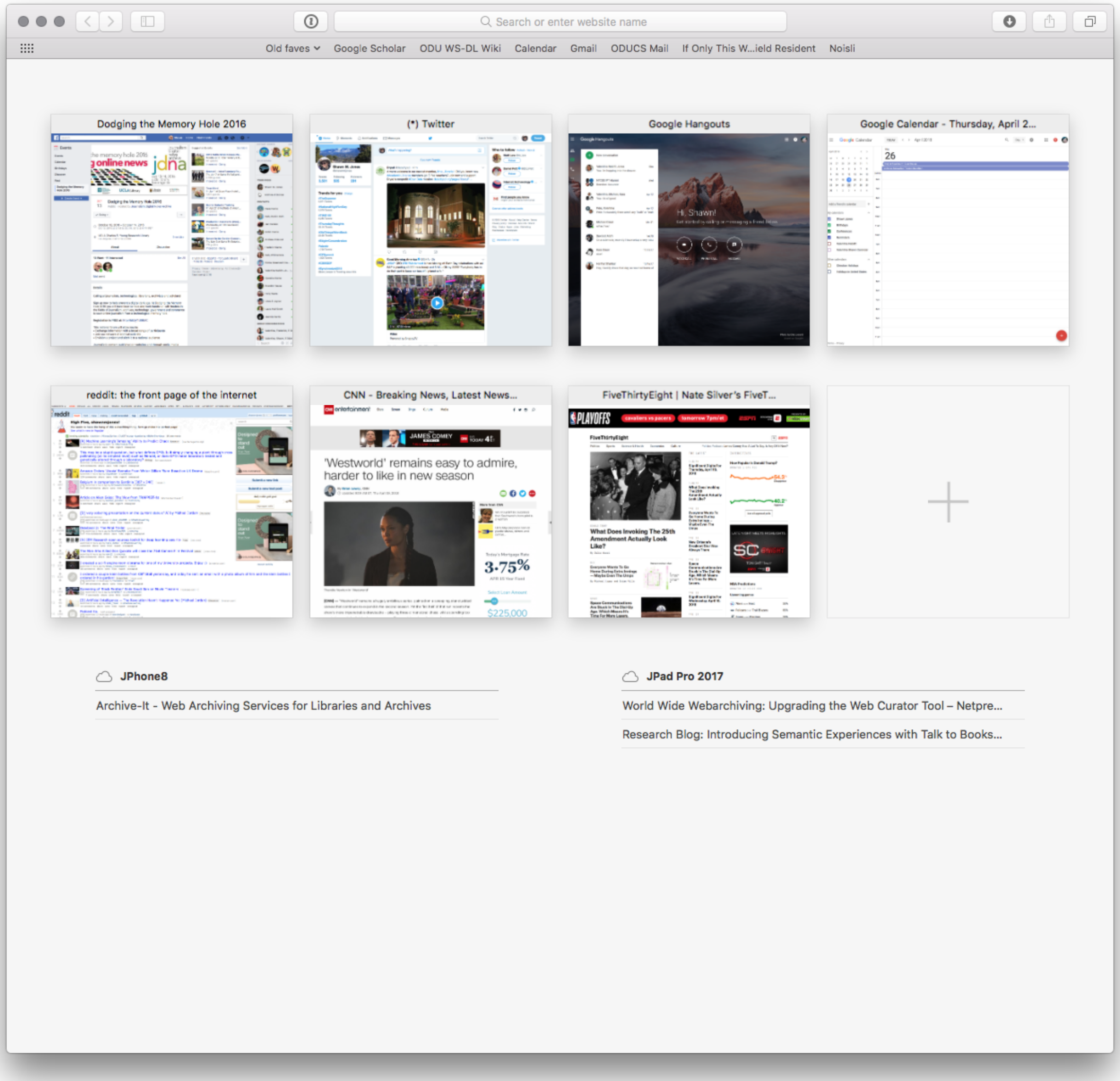

Search or enter website name
Old faves
Google Scholar
ODU WS-DL Wiki
Calendar
Gmail
ODUCS Mail
If Only This W...ield Resident
Noisli
Dodging the Memory Hole 2016
(*) Twitter
Google Hangouts
Google Calendar - Thursday, April 2...
reddit: the front page of the internet
CNN - Breaking News, Latest News...
FiveThirtyEight | Nate Silver's FiveT...
JPhone8
Archive-It - Web Archiving Services for Libraries and Archives
JPad Pro 2017
World Wide Webarchiving: Upgrading the Web Curator Tool – Netpre...
Research Blog: Introducing Semantic Experiences with Talk to Books...

Fig. 52. The Safari web browser uses thumbnails to show surrogates for web pages that are currently loaded in its tabs. (Screenshot taken in April 2018.)

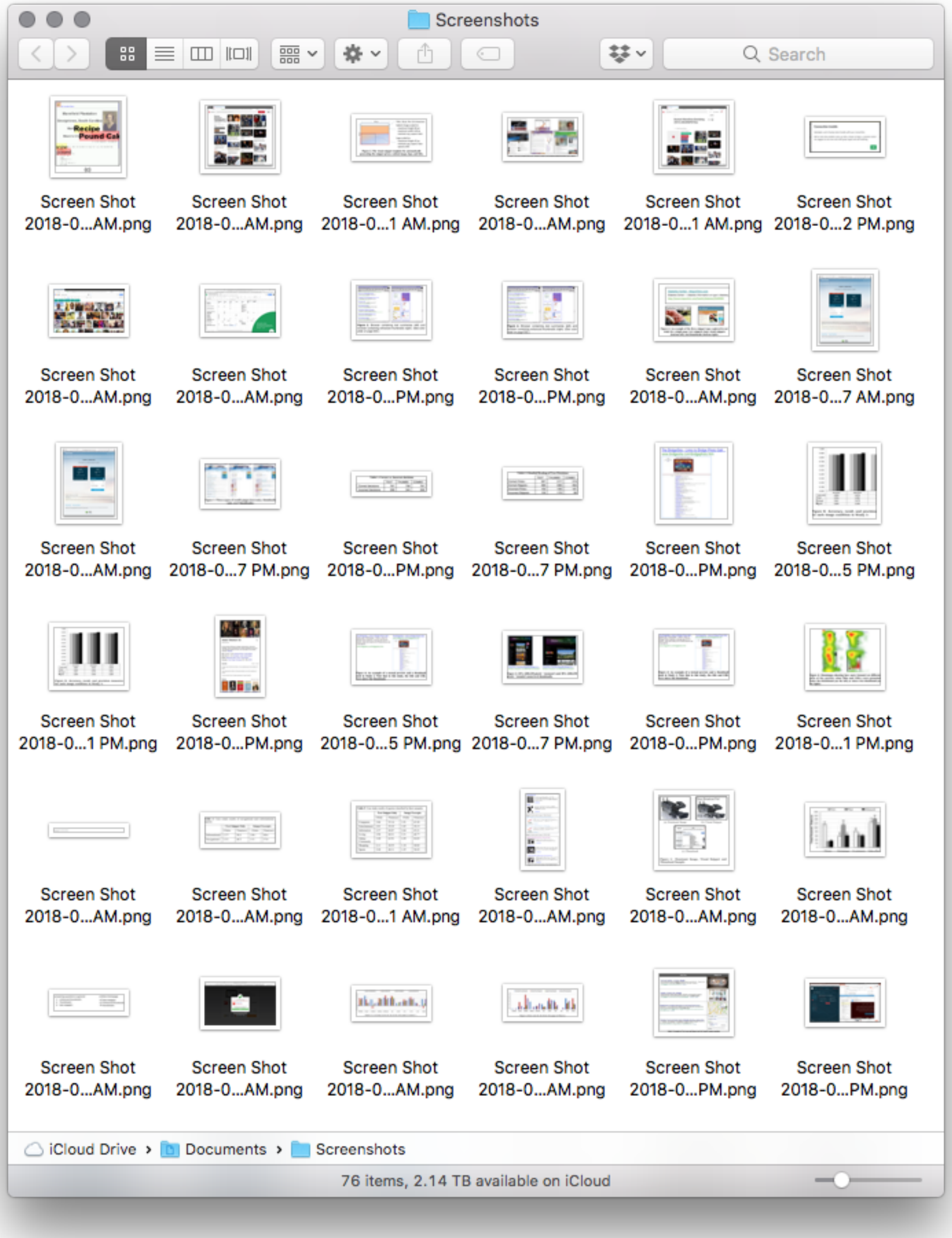

Screenshots
Search
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...1 AM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...1 AM.png
Screen Shot 2018-0...2 PM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...PM.png
Screen Shot 2018-0...PM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...7 AM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...7 PM.png
Screen Shot 2018-0...PM.png
Screen Shot 2018-0...7 PM.png
Screen Shot 2018-0...PM.png
Screen Shot 2018-0...5 PM.png
Screen Shot 2018-0...1 PM.png
Screen Shot 2018-0...PM.png
Screen Shot 2018-0...5 PM.png
Screen Shot 2018-0...7 PM.png
Screen Shot 2018-0...PM.png
Screen Shot 2018-0...1 PM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...1 AM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...AM.png
Screen Shot 2018-0...PM.png
Screen Shot 2018-0...PM.png
iCloud Drive › Documents › Screenshots
76 items, 2.14 TB available on iCloud

Fig. 53. macOS Finder displaying thumbnails of file contents. (Screenshot taken in April 2018.)

example, as seen in Figure 52, the Safari web browser uses them to display the content of tabs.

Kopetzky and Mühlhäuser [186] demonstrated that systems could use thumbnails to provide a preview of a linked page via a mouseover effect so that users could decide if a link was worth clicking. We create a similar surrogate with thumbnails and social cards in 5.2. Robertson et al. [279] proposed using a 3D virtual environment for organizing a corpus of web pages where they visualize each page as a thumbnail. Outside of the web, file management tools, such as macOS's Finder, seen in Figure 53, use thumbnails to provide visual previews of documents.

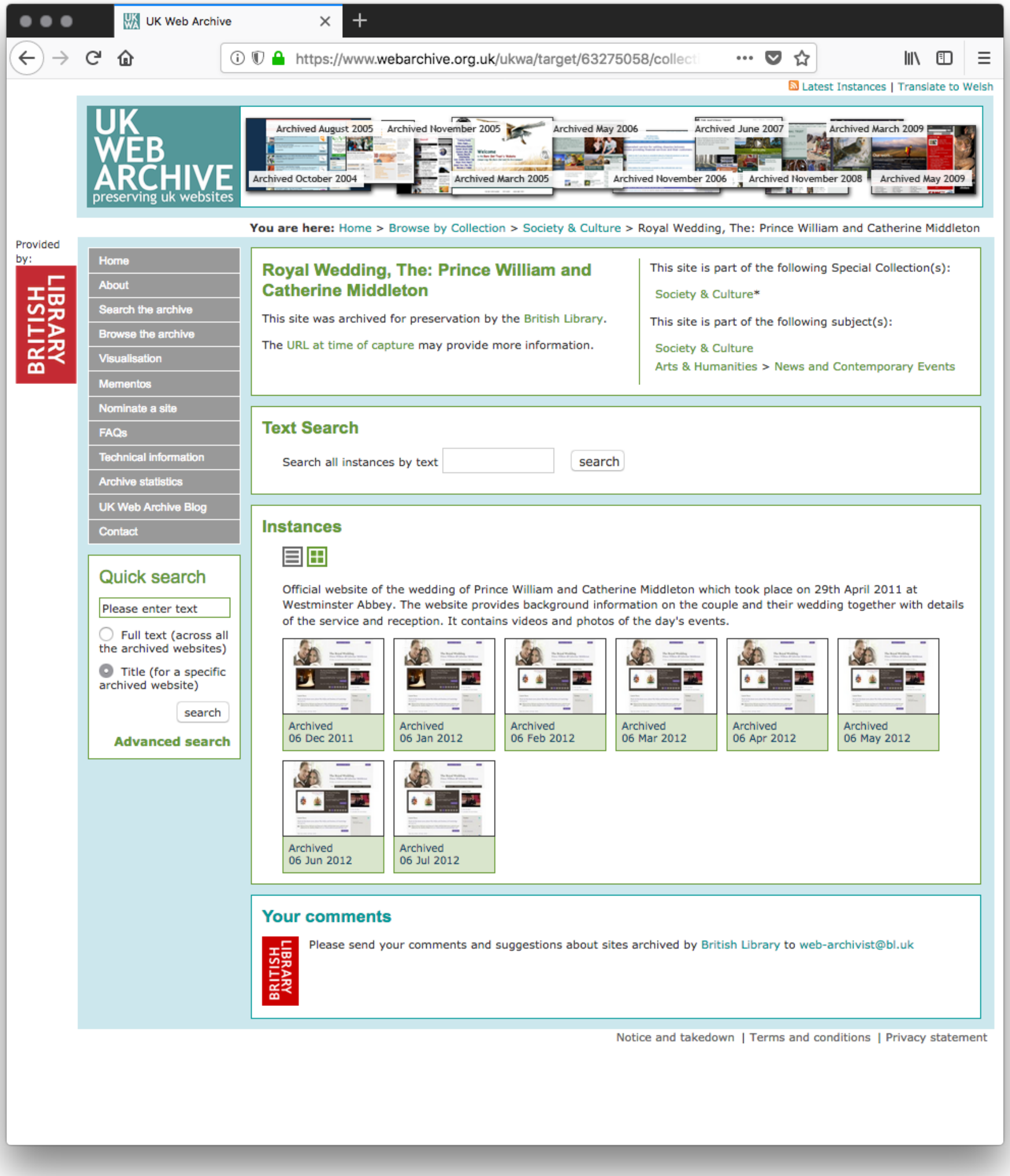
In the web archiving world, the UK Web Archive uses thumbnails (Figure 54) to show a series of mementos so one can compare the content of each memento, viewing the content drift over time. Nwala's "What Did It Look Like?" [243] is a platform that animates thumbnails so one can watch the changes to a web page over the years. Even though systems can automate their generation with tools like Puppeteer[13], browser thumbnails require significant resources to create. Generation involves launching a browser, loading the page, including all images and scripts, and then taking a screenshot of that page. In addition to the memory and processing needed, thumbnails also require multiple network connections to retrieve all resources for a page. Individually, thumbnails may not require a lot of processing or storage, but browser thumbnails can also be costly to produce and store for thousands of pages. This cost in time and resources has led to research by AlSum et al. [16] later implemented in the TMVis project [207] that focuses on optimizing the selection of mementos worthy of thumbnails. The UK Web Archive stores thumbnails in its WARCs [131], but only for seed mementos and not the deep mementos of its crawls.

The size of thumbnails has a severe effect on their utility. If the thumbnail is too large, it does not provide room for the comparison of surrogates. If the thumbnail is too small, users cannot see what is in the image. Thumbnails are also difficult for users to understand if a page consists mostly of text or has no unique features. Kaasten and Greenberg [168] established that the optimal thumbnail size is 208x208 pixels.

The viewport of a thumbnail is also an essential part of its construction. Depending on what we want to emphasize on a web page, we may need to generate a thumbnail from content "below the fold". For example, Aula et al. [24] evaluated the use of thumbnails that were the same size but had magnified a portion of a web page at 20% versus 38%. She found that users performed better with thumbnails at a magnification of 20%.

In Chapter 5.2 we evaluate if thumbnails work well for understanding web archive collections.

[13] `https://pptr.dev`



You are here: Home > Browse by Collection > Society & Culture > Royal Wedding, The: Prince William and Catherine Middleton

Royal Wedding, The: Prince William and Catherine Middleton

This site was archived for preservation by the British Library.

The URL at time of capture may provide more information.

This site is part of the following Special Collection(s):

Society & Culture*

This site is part of the following subject(s):

Society & Culture
Arts & Humanities > News and Contemporary Events

Quick search

Please enter text

○ Full text (across all the archived websites)

● Title (for a specific archived website)

search

Advanced search

Text Search

Search all instances by text [ ] search

Instances

Official website of the wedding of Prince William and Catherine Middleton which took place on 29th April 2011 at Westminster Abbey. The website provides background information on the couple and their wedding together with details of the service and reception. It contains videos and photos of the day's events.

Archived 06 Dec 2011
Archived 06 Jan 2012
Archived 06 Feb 2012
Archived 06 Mar 2012
Archived 06 Apr 2012
Archived 06 May 2012
Archived 06 Jun 2012
Archived 06 Jul 2012

Your comments

Please send your comments and suggestions about sites archived by British Library to web-archivist@bl.uk

Notice and takedown | Terms and conditions | Privacy statement

Fig. 54. The UK Web Archive uses thumbnails to show different mementos for the same resource, allowing the user to view web page changes over time. (Screenshot taken in April 2018.)

3.3.3 SOCIAL CARDS

The social card goes by many names: rich link, snippet [109], social snippet, social media card [254], Twitter card [304], embedded representation [117], rich object [87], or social card [230]. The social card typically consists of an image, a title, and a text snippet from the web page it visualizes.

Social media is not the only place where social cards are employed. As seen in Figure 55, Wikipedia uses social cards to provide a preview of links if the user hovers over the link, similar to what Kopetzky had envisioned with thumbnails. Google News[14] (Figure 56) often uses social cards for individual stories. Social cards sometimes include additional information beyond text snippets and images. Alonso et al. [15] used social cards in a prototype for Bing search results. Those cards also incorporated lists of users who shared the target web page as well as associated hashtags.

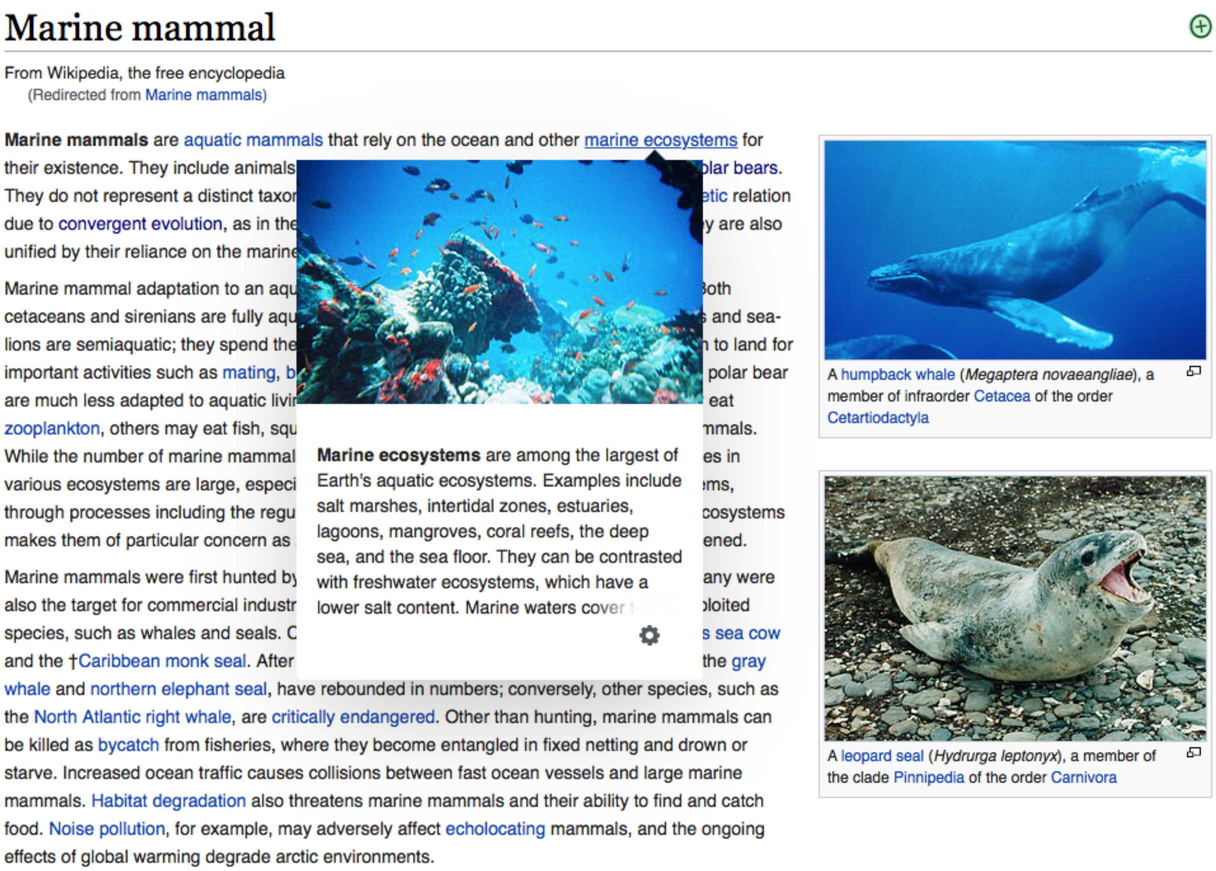
There are similar concepts that are not instances of the social card. As shown in Figure 57, some of the cards used by Google News are not social cards because each is a surrogate for a news story spanning multiple resources, rather than a single resource. Likewise, search engines use entity cards to display information about a specific entity drawn from multiple sources (Figure 58). Bota et al. [39] established that entity cards are useful to search engine users. We do not consider entity cards to be social cards because each social card is a surrogate for a single web resource. An entity card is a surrogate for a conceptual entity, and a system can construct it from multiple sources.

In Chapter 5.2 we evaluate if social cards or a combination of social cards and thumbnails work well for understanding web archive collections.

3.3.4 SURROGATE EVALUATION

Web page surrogates have been of great interest to those studying search engine result pages. Here we detail eight studies on web surrogates, most of which we already mentioned above. Table 9 displays the different surrogates evaluated by these studies.

[14] `https://news.google.com`



Marine mammal

From Wikipedia, the free encyclopedia
(Redirected from Marine mammals)

Marine mammals are aquatic mammals that rely on the ocean and other marine ecosystems for
their existence. They include animals plar bears.
They do not represent a distinct taxor etic relation
due to convergent evolution, as in the y are also
unified by their reliance on the marine

Marine mammal adaptation to an aqu Both
cetaceans and sirenians are fully aqu s and sea-
lions are semiaquatic; they spend the n to land for
important activities such as mating, b polar bear
are much less adapted to aquatic livir eat
zooplankton, others may eat fish, squ mmals.
While the number of marine mammal es in
various ecosystems are large, especi ems,
through processes including the regu cosystems
makes them of particular concern as ened.

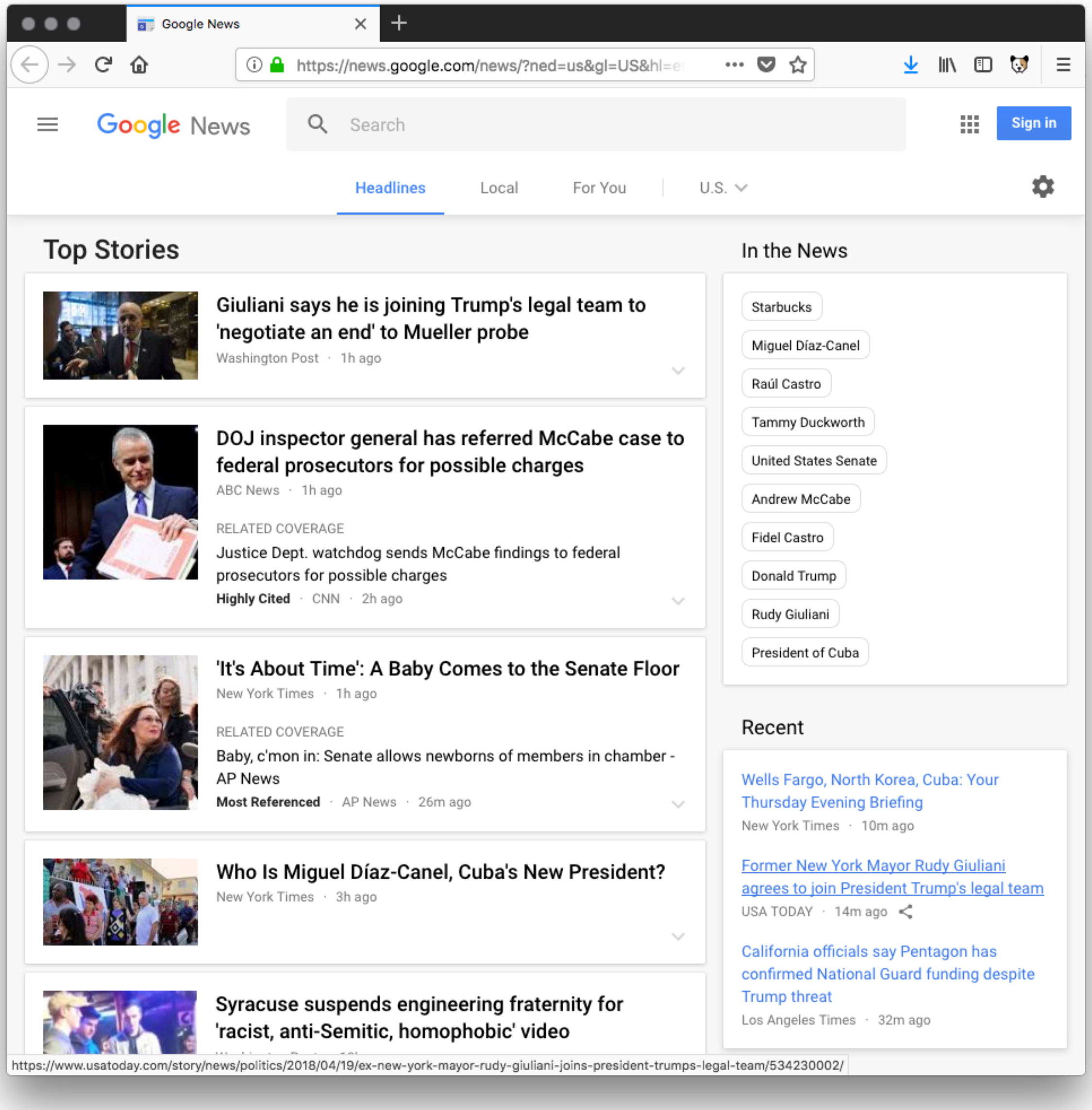
Marine mammals were first hunted by any were
also the target for commercial industr ploited
species, such as whales and seals. C s sea cow
and the †Caribbean monk seal. After the gray
whale and northern elephant seal, have rebounded in numbers; conversely, other species, such as the North Atlantic right whale, are critically endangered. Other than hunting, marine mammals can be killed as bycatch from fisheries, where they become entangled in fixed netting and drown or starve. Increased ocean traffic causes collisions between fast ocean vessels and large marine mammals. Habitat degradation also threatens marine mammals and their ability to find and catch food. Noise pollution, for example, may adversely affect echolocating mammals, and the ongoing effects of global warming degrade arctic environments.

Marine ecosystems are among the largest of Earth's aquatic ecosystems. Examples include salt marshes, intertidal zones, estuaries, lagoons, mangroves, coral reefs, the deep sea, and the sea floor. They can be contrasted with freshwater ecosystems, which have a lower salt content. Marine waters cover t

A humpback whale (*Megaptera novaeangliae*), a member of infraorder Cetacea of the order Cetartiodactyla

A leopard seal (*Hydrurga leptonyx*), a member of the clade Pinnipedia of the order Carnivora

Fig. 55. When a user hovers over an internal link, Wikipedia uses social cards to display a preview of the linked web page [322]. (Screenshot taken in April 2018.)



Top Stories

Giuliani says he is joining Trump's legal team to 'negotiate an end' to Mueller probe
Washington Post · 1h ago

DOJ inspector general has referred McCabe case to federal prosecutors for possible charges
ABC News · 1h ago

RELATED COVERAGE
Justice Dept. watchdog sends McCabe findings to federal prosecutors for possible charges
Highly Cited · CNN · 2h ago

'It's About Time': A Baby Comes to the Senate Floor
New York Times · 1h ago

RELATED COVERAGE
Baby, c'mon in: Senate allows newborns of members in chamber - AP News
Most Referenced · AP News · 26m ago

Who Is Miguel Díaz-Canel, Cuba's New President?
New York Times · 3h ago

Syracuse suspends engineering fraternity for 'racist, anti-Semitic, homophobic' video

In the News

- Starbucks
- Miguel Díaz-Canel
- Raúl Castro
- Tammy Duckworth
- United States Senate
- Andrew McCabe
- Fidel Castro
- Donald Trump
- Rudy Giuliani
- President of Cuba

Recent

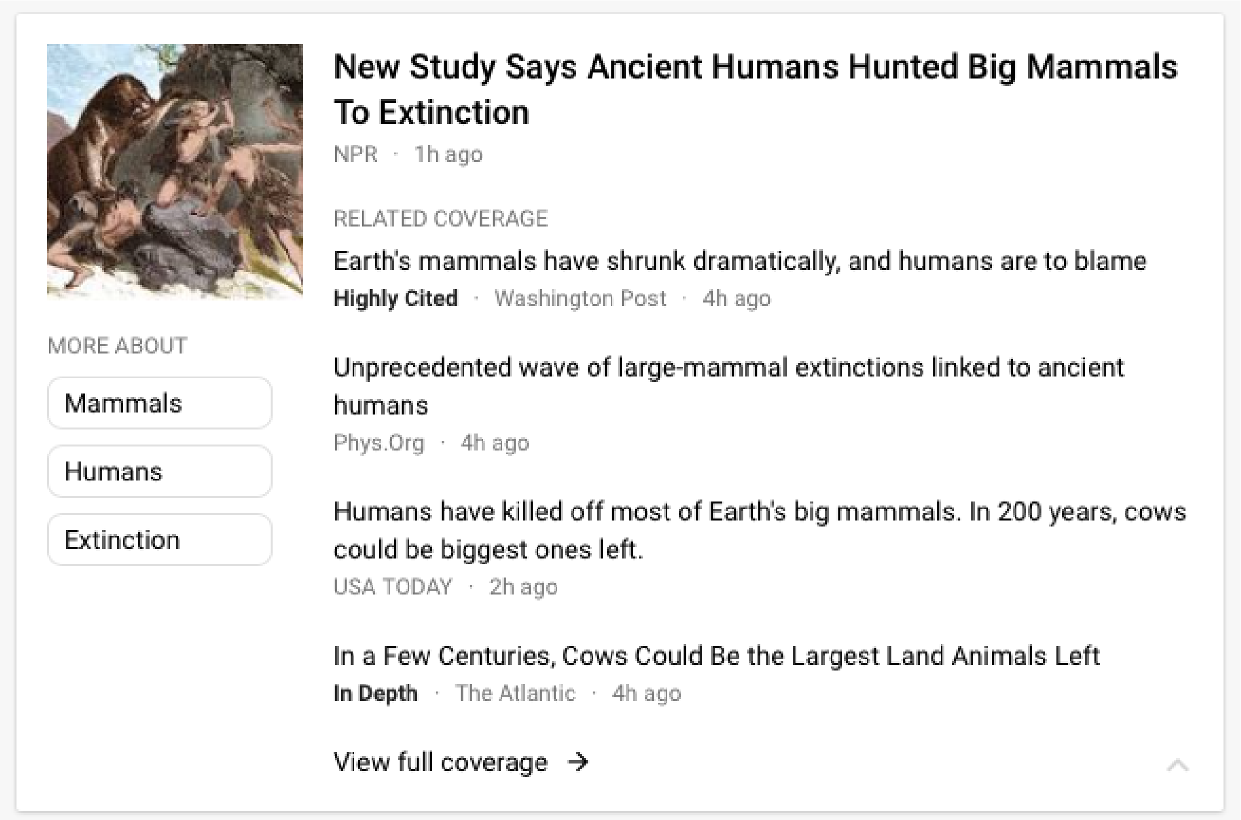
Wells Fargo, North Korea, Cuba: Your Thursday Evening Briefing
New York Times · 10m ago

Former New York Mayor Rudy Giuliani agrees to join President Trump's legal team
USA TODAY · 14m ago

California officials say Pentagon has confirmed National Guard funding despite Trump threat
Los Angeles Times · 32m ago

https://www.usatoday.com/story/news/politics/2018/04/19/ex-new-york-mayor-rudy-giuliani-joins-president-trumps-legal-team/534230002/

Fig. 56. Google News uses other types of cards to group news articles. (Screenshot taken in April 2018.)



New Study Says Ancient Humans Hunted Big Mammals To Extinction

NPR · 1h ago

RELATED COVERAGE

Earth's mammals have shrunk dramatically, and humans are to blame

Highly Cited · Washington Post · 4h ago

Unprecedented wave of large-mammal extinctions linked to ancient humans

Phys.Org · 4h ago

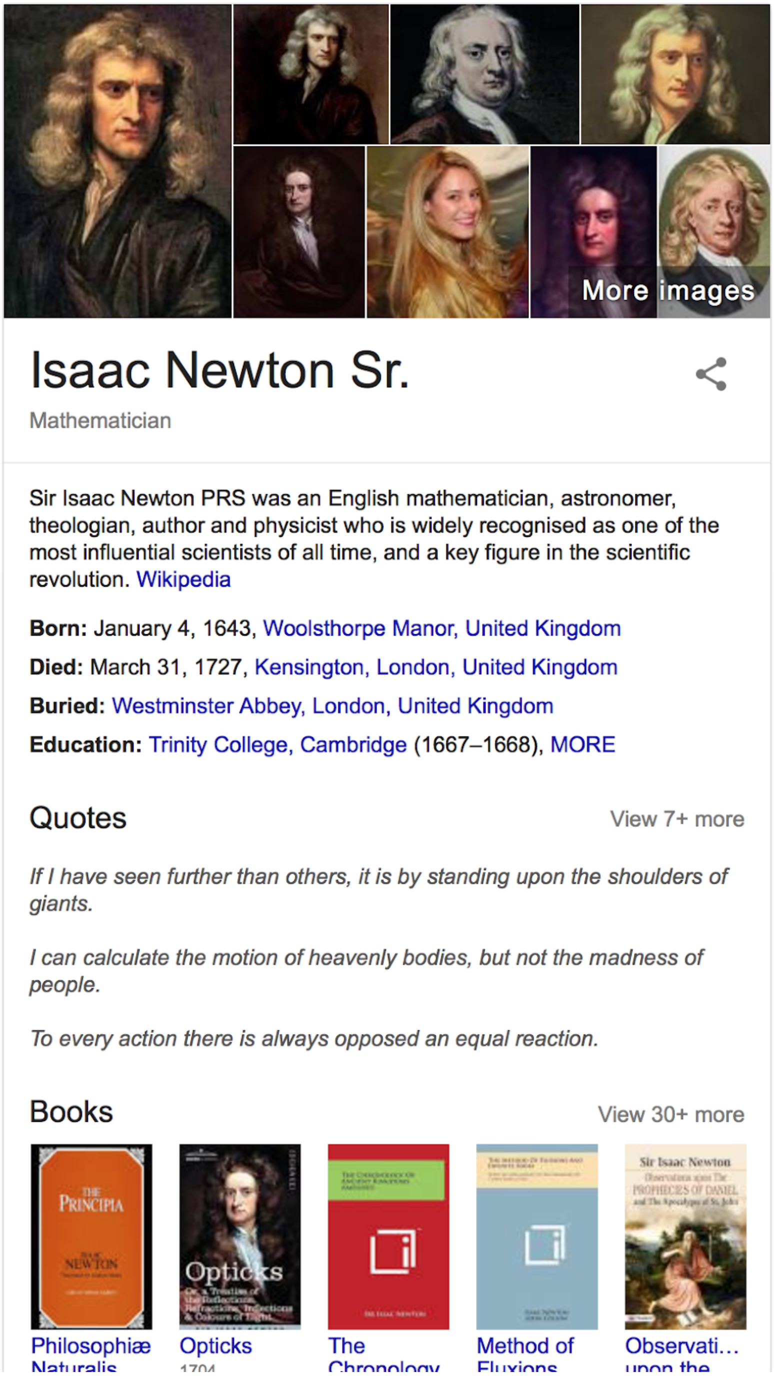
Humans have killed off most of Earth's big mammals. In 200 years, cows could be biggest ones left.

USA TODAY · 2h ago

In a Few Centuries, Cows Could Be the Largest Land Animals Left

In Depth · The Atlantic · 4h ago

Fig. 57. This card used by Google News is not a surrogate for a single web resource, and hence we do not consider it a social card. (Screenshot taken in April 2018.)



Isaac Newton Sr.

Mathematician

Sir Isaac Newton PRS was an English mathematician, astronomer, theologian, author and physicist who is widely recognised as one of the most influential scientists of all time, and a key figure in the scientific revolution. Wikipedia

Born: January 4, 1643, Woolsthorpe Manor, United Kingdom

Died: March 31, 1727, Kensington, London, United Kingdom

Buried: Westminster Abbey, London, United Kingdom

Education: Trinity College, Cambridge (1667–1668), MORE

Quotes — View 7+ more

If I have seen further than others, it is by standing upon the shoulders of giants.

I can calculate the motion of heavenly bodies, but not the madness of people.

To every action there is always opposed an equal reaction.

Books — View 30+ more

Philosophiæ Naturalis | Opticks 1704 | The Chronology | Method of Fluxions | Observati… upon the

Fig. 58. This card format, used by Google Search is also not a surrogate for a single web resource. This is an entity card, drawing from multiple web resources. (Screenshot taken in April 2018.)

Woodruff et al. [331, 332] introduced the concept of enhanced thumbnails. These thumbnails are a snapshot of the page in a browser, but only after a system has increased the font size of keywords to make them more visible in the thumbnail. To evaluate their effectiveness, she generated questions based on work by Morrison et al. [227] about tasks users commonly perform on the web. The authors divided the questions into four categories, and three questions per category were each given to 18 participants. The authors presented the participants with search engine result pages consisting of 100 text snippets, thumbnails, or enhanced thumbnails. In their attempt to find web resources that would address their assigned questions, participants were evaluated based on their response times. The results indicated that enhanced thumbnails provided the fastest response times overall, but the results varied depending on the type of task. For example, for locating an entity's homepage, text snippets and enhanced thumbnails performed roughly the same. Likewise, finding the picture of an entity performed roughly the same between thumbnails and enhanced thumbnails. All three surrogate types performed just as well for e-commerce or medical side-effect questions.

Dziadosz and Chandrasekar [77] tested the concept of text snippets combined with thumbnails. In this study of 35 participants, the authors gave each of them two queries and two tasks. The authors gave each participant a different surrogate type. The participant's first task was to identify all search engine results on the page they assumed to be relevant to their query. Their second task was to visit the pages being the surrogates and identify which were relevant. The number of correct decisions for text snippets combined with thumbnails was higher than just for text or just for thumbnails. Aula et al. [24] also evaluated the results of 223 participants viewing text snippets, thumbnails, or a combination. In their study they discovered that a combination was more effective in making relevance judgments.

Teevan et al. [293] evaluated the effectiveness of visual snippets — single images combining a striking image with a page's site logo and title text. Her study consisted of 276 participants who were each given 12 search tasks and a set of 20 search results, with 4 of the 12 tasks completed with different surrogates. She discovered that text snippets required the fewest clicks compared to thumbnails, which required the most. This result indicates a lot of false-positive matches for participants when using thumbnails. Participants preferred visual snippets or text snippets equally over thumbnails and preferred visual snippets for shopping tasks. Most participants found thumbnails to be too small to be helpful.

Jiao et al. [134] introduced the concept of using external images generated by search results as a surrogate. He compared the use of internal images, external images, thumbnails, and visual snippets by 51 participants. Like Dziadosz's study, they asked participants to guess the web page's

TABLE 9

The surrogates evaluated by different user studies.

First Author & Year	**Text Snippet**	**Internal/ External Image**	**Visual Snippet**	**Thumbnail**	**Enhanced Thumbnail/ Visual Tags**	**Text + Thumbnail**	**Social Card**
Woodruff 2001	X			X	X		
Dziadosz 2002	X			X		X	
Li 2008	X						X
Teevan 2009	X		X	X			
Jiao 2010		X	X	X			
Aula 2010	X			X			
Al Maqbali 2010	X		X		X	X	X
Loumakis 2011	X	X					X
Capra 2013	X	X					X

relevance behind the surrogate and then later evaluate if their earlier guess was correct. To generate search results, they randomly sampled 100 queries from the KDD CUP '05 dataset [196] and submitted them to Bing. His results show that none of the surrogates works for all types of pages. For example, overall, internal images were best for pages that contained a dominant image. In contrast, thumbnails or external images were best for understanding pages that did not contain a dominant image.

Li et al. [197] was interested in identifying dominant images in web pages. We focus here on the second study in their work which compares text snippets and social cards. First, the authors randomly sampled 100 queries from the KDD CUP '05 dataset [196] and submitted them to Google. Next, the authors evaluated the search engine results and reformatted them into either text snippets or social cards. Then, the authors gave two groups of 12 students the queries either classified by their functionalities or semantic categories. Finally, the authors evaluated participants based on the number of clicks of relevant results and also on the amount of time they took with each search. Social cards were the clear winner over text snippets in terms of time and clicks.

Loumakis [204] attempted to compare the performance of images, text snippets, and social cards. Using pre-selected queries and 81 participants, Loumakis also reformatted Google search results. Unfortunately, he did not get the same level of performance in his study, noting "Adding an image to a SERP result will not significantly help users in identifying correct results, but neither will it significantly hinder them if an image is placed with text cues where the scents may conflict."

Al Maqbali [4] explored the use of different image augmentations for visual snippets, text + thumbnail, social card, text + visual snippet, and a text + tag cloud/thumbnail combination. Al Maqbali had 65 participants evaluate the relevance of search engine result pages as in the prior studies. This study reached the same conclusion as Loumakis: adding images to text snippets did not make a difference in search engine users' performance.

To further understand the disagreement between the results of Loumakis [204], Al Maqbali [4], and Li [197], Capra [50] explored the effectiveness of text snippets and social cards. He wanted to determine if the quality or relevance of the image used in the social card affected performance. Before any relevance study, he had the participants rate individual internal images for a social card as good, bad, and mixed. Capra discovered that text snippets with good images have a slightly higher statistically significant accuracy score than just text snippets alone, at the cost of judgment duration for each surrogate. The accuracy for text snippets was 0.864, the accuracy for social cards with bad images was also 0.864, and the accuracy for social cards with good images was 0.884. From the perspective of search engine result pages overall, there was evidence that good images showed improvement in accuracy with ambiguous queries (e.g., jaguar the car or the cat?).

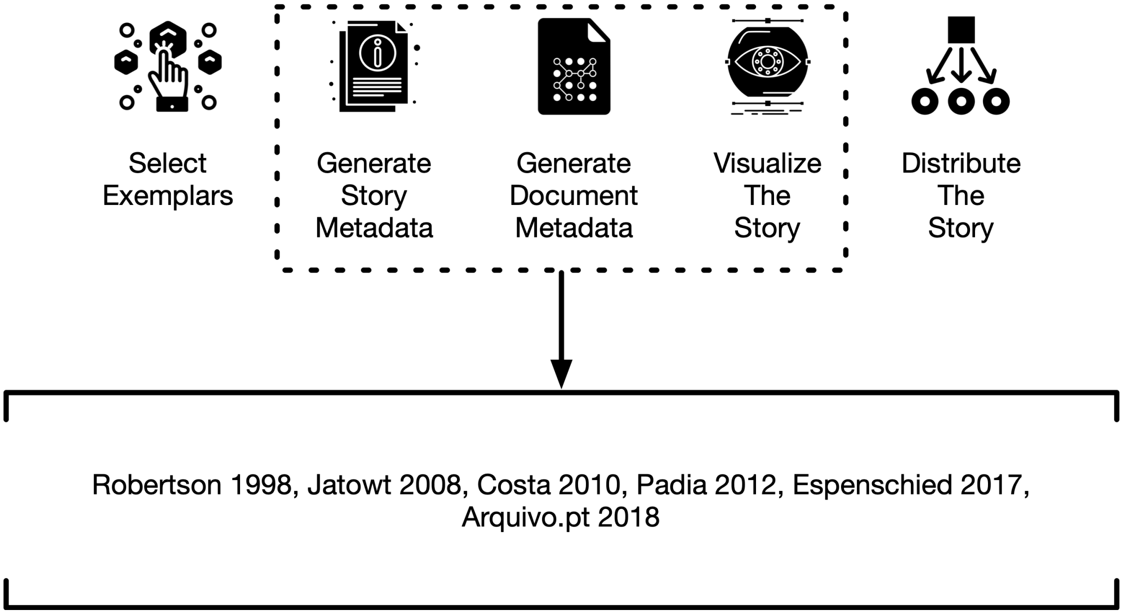

Select Exemplars
Generate Story Metadata
Generate Document Metadata
Visualize The Story
Distribute The Story
Robertson 1998, Jatowt 2008, Costa 2010, Padia 2012, Espenschied 2017, Arquivo.pt 2018

Fig. 59. Others have tried to make sense of web page collections.

However, in this case, the improvements were not statistically significant.

Where these studies focused on whether or not users selected relevant search engine results with different surrogate types, in Chapter 5.2 we evaluate how well users understand collections through different types of surrogates. We find, much like Capra, that social cards are better for understanding collections than the existing Archive-It surrogate.

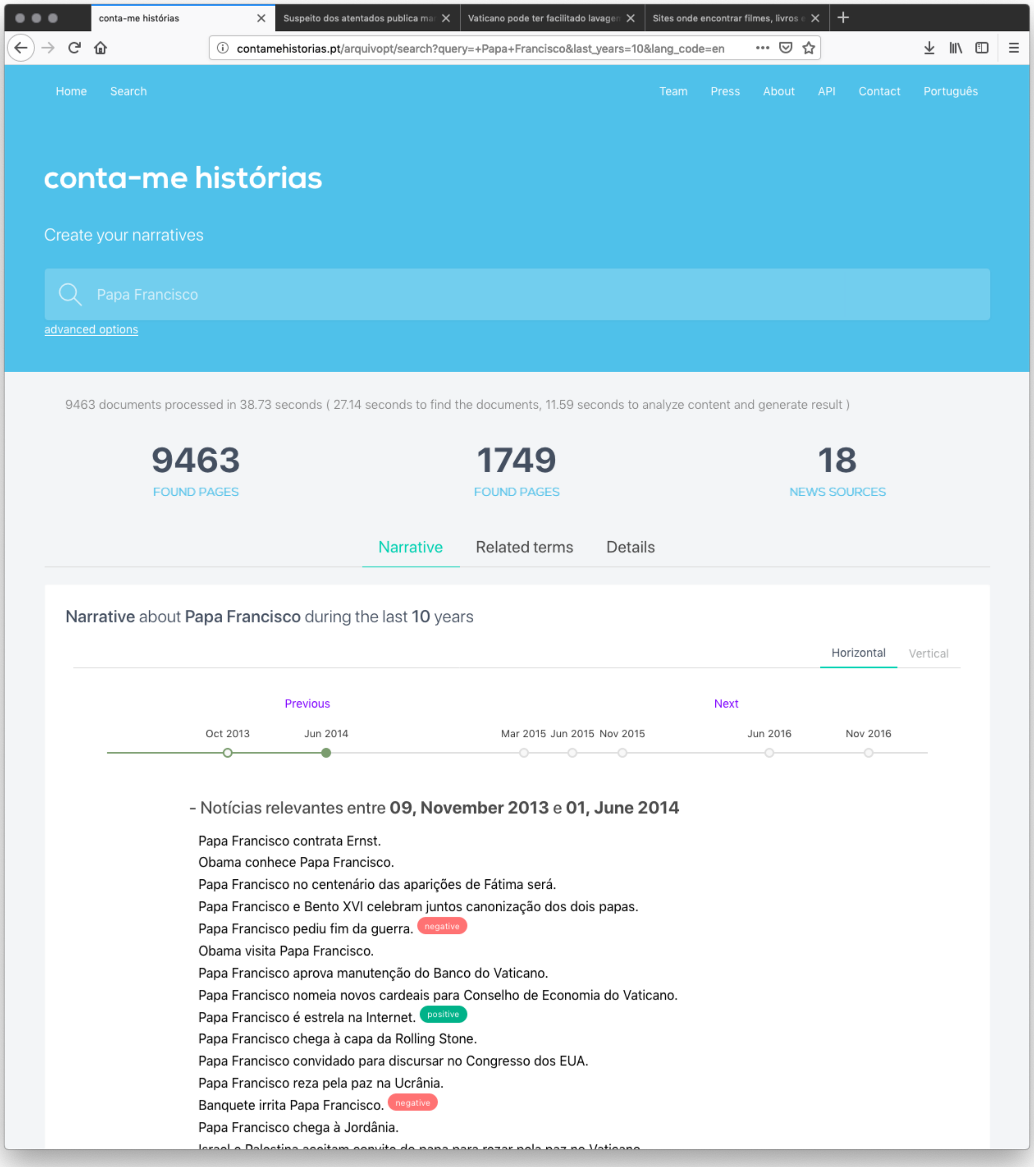
3.4 VISUALIZING WEB ARCHIVE COLLECTIONS AS A WHOLE

Several projects – visualized in Figure 59 – have worked on visualizing web archive collections as a whole. They do not create stories. These efforts were trying to take into account everything in the collection rather than selecting exemplars. They often use story metadata and document metadata but are difficult for users to understand without training.

Specific to the Portuguese Web Archive (Arquivo.pt), the Laboratory of Artificial Intelligence and Decision Support has developed Conta Me Histórias (Tell Me Stories) [49] Temporal Summarization Framework[15,16]. It is a storytelling tool that retrieves mementos across that web archive and generates a listing of headlines for a given topic over time. To use Conta Me Histórias, a user types in a search term, and the system finds mementos fitting that term. From there, the system

[15] `https://github.com/LIAAD/TemporalSummarizationFramework`

[16] `http://contamehistorias.pt/arquivopt/`



conta-me histórias

Create your narratives

9463 documents processed in 38.73 seconds (27.14 seconds to find the documents, 11.59 seconds to analyze content and generate result)

9463	1749	18
FOUND PAGES	FOUND PAGES	NEWS SOURCES

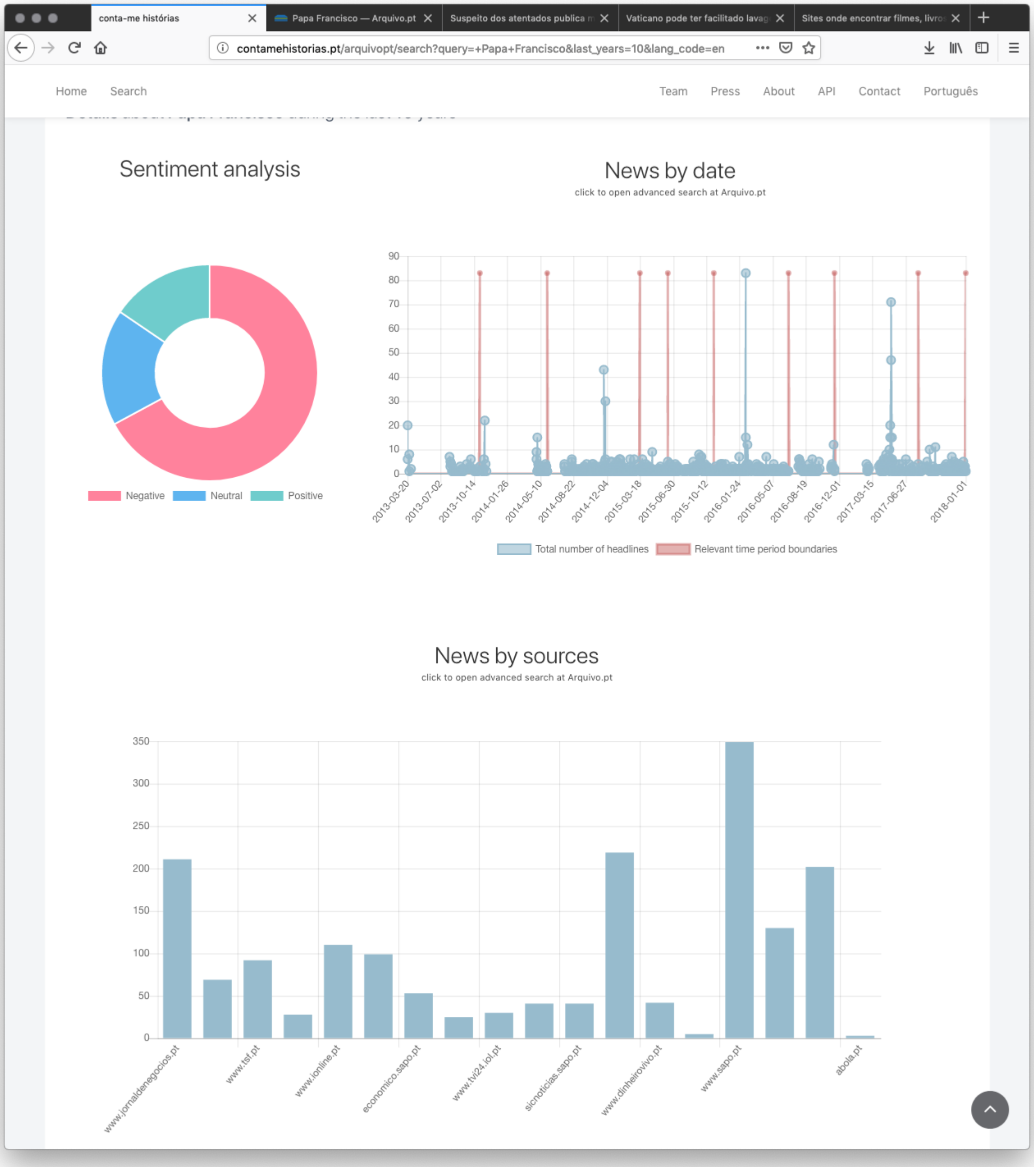
Narrative about **Papa Francisco** during the last **10** years

- Notícias relevantes entre **09, November 2013** e **01, June 2014**

Papa Francisco contrata Ernst.
Obama conhece Papa Francisco.
Papa Francisco no centenário das aparições de Fátima será.
Papa Francisco e Bento XVI celebram juntos canonização dos dois papas.
Papa Francisco pediu fim da guerra. negative
Obama visita Papa Francisco.
Papa Francisco aprova manutenção do Banco do Vaticano.
Papa Francisco nomeia novos cardeais para Conselho de Economia do Vaticano.
Papa Francisco é estrela na Internet. positive
Papa Francisco chega à capa da Rolling Stone.
Papa Francisco convidado para discursar no Congresso dos EUA.
Papa Francisco reza pela paz na Ucrânia.
Banquete irrita Papa Francisco. negative
Papa Francisco chega à Jordânia.
Israel e Palestina aceitam convite do papa para rezar pela paz no Vaticano.

(a) Timeline Interface

Fig. 60. Conta Me Histórias displays a timeline of results for Papa Franciso (Pope Francis) drawn from the Portuguese Web Archive (Arquivo.pt). Examples from: `http://contamehistorias.pt/arquivopt/`



(b) Statistics Interface

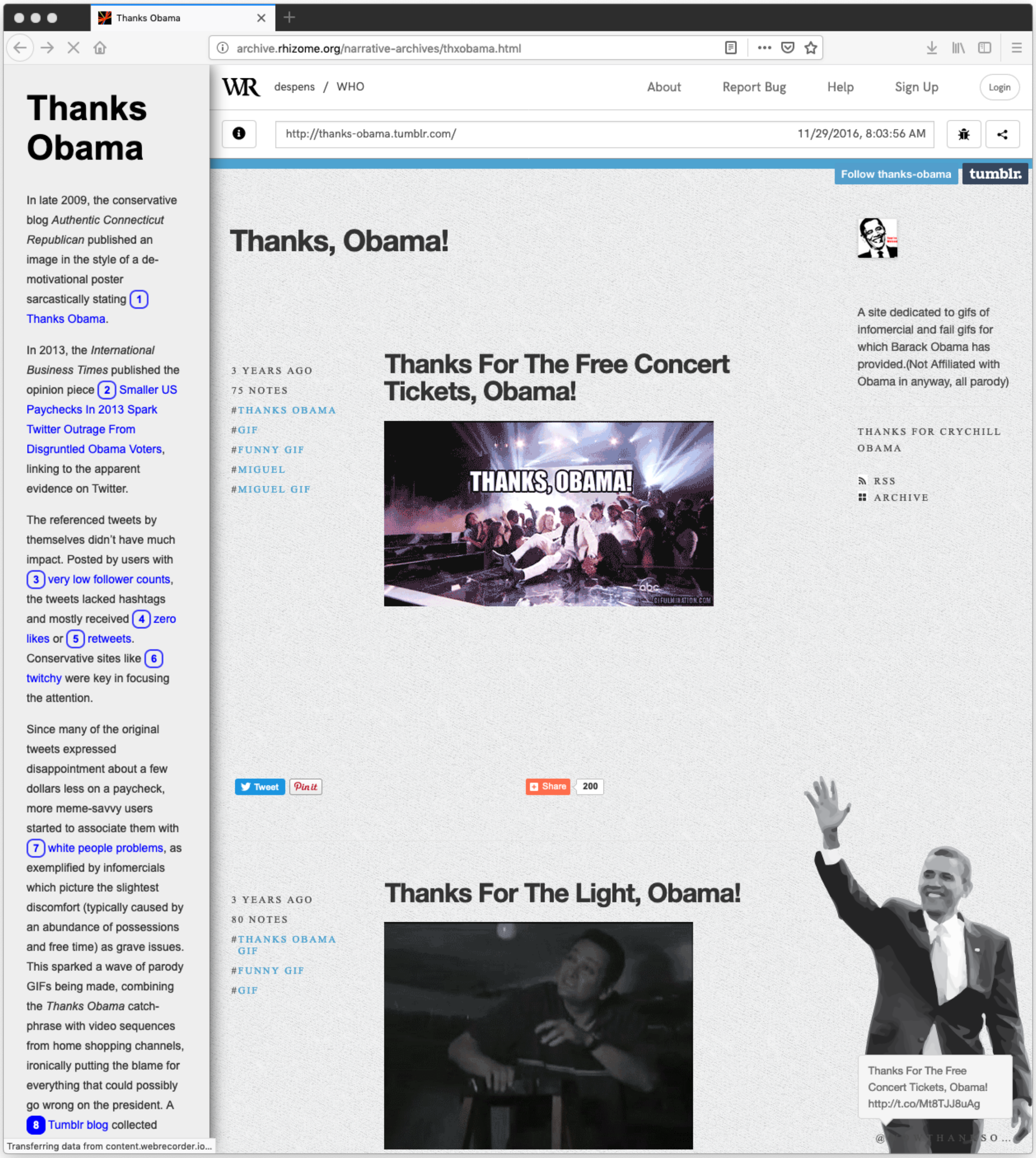
Fig. 60. (Continued) Conta Me Histórias displays a timeline of results for Papa Franciso (Pope Francis) drawn from the Portuguese Web Archive (Arquivo.pt). Examples from: `http://contamehistorias.pt/arquivopt/`

generates a story consisting of a timeline and a set of headlines corresponding to mementos for the search term, seen in Figure 60a. A user can click on the headline and view the memento. Clicking on the *Details* tab brings the user to a set of statistics about the mementos, seen in Figure 60b. From here, we can see sentiment analysis, the distribution of the mementos, and the distribution of domain names. These visualizations combine information retrieval and visualization to deliver relevant search terms while preserving the available temporal information. In our case, we are focusing on a single web archive collection rather than only representing mementos that match a search term across an entire web archive.

Espenschied [276] worked with data from the Obama White House to craft stories combining mementos from Webrecorder with the Obama social media archive. He prototyped a narrative framework for three different events in the Obama presidency, shown in Figure 61. Humans created these stories. On the left side exists a set of paragraphs detailing how a given social media story unfolded. These paragraphs contain links to mementos gathered using Webrecorder. Espenschied notes how this is an example of "how a curated web archive can clear a path through otherwise overwhelming data and really tell a story." While we believe that these narrative features are a meaningful future for web archives, our goal is to generate automated stories that summarize web archive collections. Our goals align in the desire to help users find the signal in overwhelming data.

Jatowt et al. [133] were one of the first to visualize web pages over time. They demonstrated the Page History Explorer (PHE), a tool that downloads mementos from web archives and provides a visualization. PHE shows all mementos of a given URI as thumbnails from oldest to newest, spanning left to right. For several user-specified time periods, a term cloud displays up to 20 top-scored terms from the web page during that time. Clicking on any of the thumbnails gives the user a larger window holding the thumbnail. If the user supplies a keyword, then the system arranges the pages along the y-axis in the order of frequency of that keyword among the mementos. Jatowt's visualization worked for a single original resource and not a web archive collection, but this visualization influenced Padia and AlNoamany.

Padia and AlNoamany [256, 257] developed several visualizations of web archive collections. Some of these visualizations rely upon the *Group* seed metadata field, as generated by the curator of the collection. As noted in earlier chapters, most Archive-It collections lack seed metadata; thus, some of these visualizations may not work collections outside of these examples. For collections lacking this metadata, Padia developed a set of rules to categorize websites based on the URI features work of Kan and Thi [172]. Padia [256] created rules to fit URIs into the categories of social media, news, blogs, videos, and others.



Thanks Obama

In late 2009, the conservative blog *Authentic Connecticut Republican* published an image in the style of a de-motivational poster sarcastically stating (1) Thanks Obama.

In 2013, the *International Business Times* published the opinion piece (2) Smaller US Paychecks In 2013 Spark Twitter Outrage From Disgruntled Obama Voters, linking to the apparent evidence on Twitter.

The referenced tweets by themselves didn't have much impact. Posted by users with (3) very low follower counts, the tweets lacked hashtags and mostly received (4) zero likes or (5) retweets. Conservative sites like (6) twitchy were key in focusing the attention.

Since many of the original tweets expressed disappointment about a few dollars less on a paycheck, more meme-savvy users started to associate them with (7) white people problems, as exemplified by infomercials which picture the slightest discomfort (typically caused by an abundance of possessions and free time) as grave issues. This sparked a wave of parody GIFs being made, combining the *Thanks Obama* catch-phrase with video sequences from home shopping channels, ironically putting the blame for everything that could possibly go wrong on the president. A (8) Tumblr blog collected

Thanks, Obama!

A site dedicated to gifs of infomercial and fail gifs for which Barack Obama has provided.(Not Affiliated with Obama in anyway, all parody)

THANKS FOR CRYCHILL OBAMA

3 YEARS AGO
75 NOTES
#THANKS OBAMA
#GIF
#FUNNY GIF
#MIGUEL
#MIGUEL GIF

Thanks For The Free Concert Tickets, Obama!

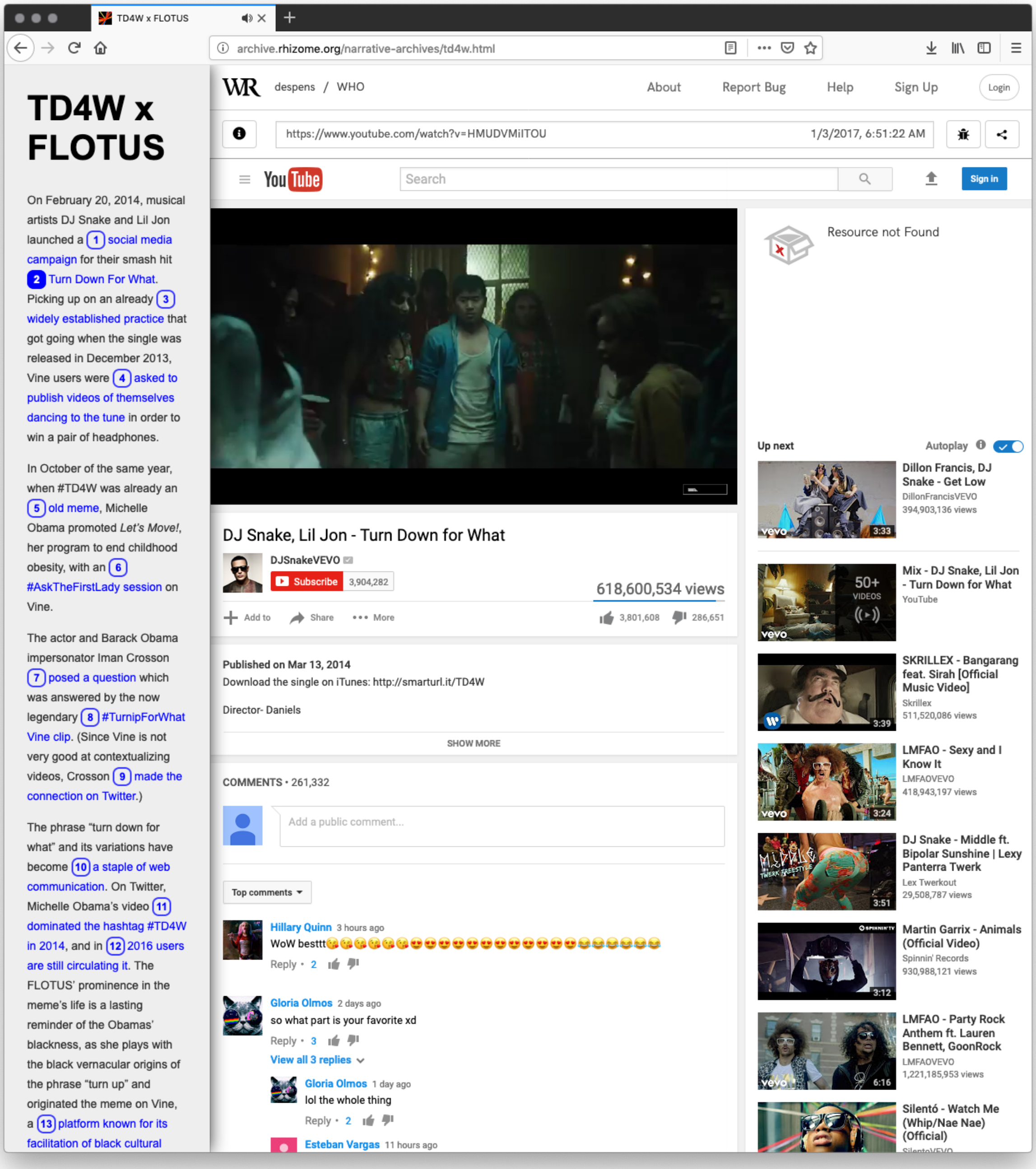
3 YEARS AGO
80 NOTES
#THANKS OBAMA GIF
#FUNNY GIF
#GIF

Thanks For The Light, Obama!

Thanks For The Free Concert Tickets, Obama! http://t.co/Mt8TJJ8uAg

(a) "Thanks Obama"

Fig. 61. Stories produced from Webrecorder as part of a prototype of its narrative features. Examples from: `http://archive.rhizome.org/narrative-archives/`. (Screenshots taken July 2018.)



TD4W x FLOTUS

On February 20, 2014, musical artists DJ Snake and Lil Jon launched a (1) social media campaign for their smash hit (2) Turn Down For What. Picking up on an already (3) widely established practice that got going when the single was released in December 2013, Vine users were (4) asked to publish videos of themselves dancing to the tune in order to win a pair of headphones.

In October of the same year, when #TD4W was already an (5) old meme, Michelle Obama promoted *Let's Move!*, her program to end childhood obesity, with an (6) #AskTheFirstLady session on Vine.

The actor and Barack Obama impersonator Iman Crosson (7) posed a question which was answered by the now legendary (8) #TurnipForWhat Vine clip. (Since Vine is not very good at contextualizing videos, Crosson (9) made the connection on Twitter.)

The phrase "turn down for what" and its variations have become (10) a staple of web communication. On Twitter, Michelle Obama's video (11) dominated the hashtag #TD4W in 2014, and in (12) 2016 users are still circulating it. The FLOTUS' prominence in the meme's life is a lasting reminder of the Obamas' blackness, as she plays with the black vernacular origins of the phrase "turn up" and originated the meme on Vine, a (13) platform known for its facilitation of black cultural

DJ Snake, Lil Jon - Turn Down for What

DJSnakeVEVO

3,904,282

618,600,534 views

3,801,608 286,651

Published on Mar 13, 2014
Download the single on iTunes: http://smarturl.it/TD4W

Director- Daniels

COMMENTS • 261,332

Hillary Quinn 3 hours ago
WoW besttt😘😘😘😘😘😘😍😍😍😍😍😍😍😍😍😍😍😍😂😂😂😂😂😂
Reply • 2

Gloria Olmos 2 days ago
so what part is your favorite xd
Reply • 3

Gloria Olmos 1 day ago
lol the whole thing
Reply • 2

Esteban Vargas 11 hours ago

Dillon Francis, DJ Snake - Get Low
DillonFrancisVEVO
394,903,136 views
3:33

Mix - DJ Snake, Lil Jon - Turn Down for What
YouTube
50+ VIDEOS

SKRILLEX - Bangarang feat. Sirah [Official Music Video]
Skrillex
511,520,086 views
3:39

LMFAO - Sexy and I Know It
LMFAOVEVO
418,943,197 views
3:24

DJ Snake - Middle ft. Bipolar Sunshine | Lexy Panterra Twerk
Lex Twerkout
29,508,787 views
3:51

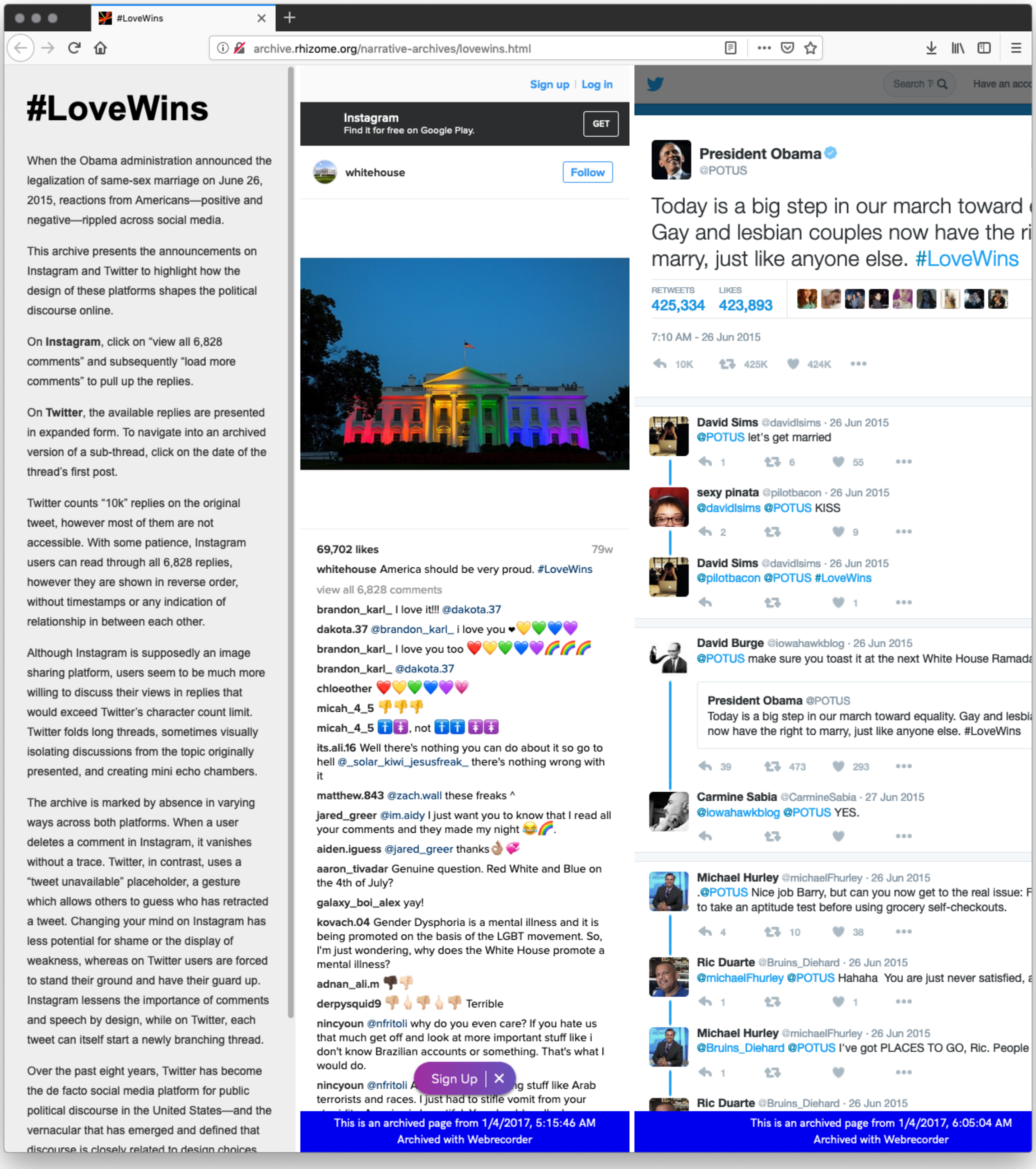
Martin Garrix - Animals (Official Video)
Spinnin' Records
930,988,121 views
3:12

LMFAO - Party Rock Anthem ft. Lauren Bennett, GoonRock
LMFAOVEVO
1,221,185,953 views
6:16

Silentó - Watch Me (Whip/Nae Nae) (Official)
SilentoVEVO

(b) TD4W x FLOTUS

Fig. 61. Stories produced from Webrecorder as part of a prototype of its narrative features. Examples from: `http://archive.rhizome.org/narrative-archives/`. (Screenshots taken July 2018.)


#LoveWins
archive.rhizome.org/narrative-archives/lovewins.html
Sign up | Log in
Instagram
Find it for free on Google Play.
GET
Follow
Search T
Have an acc

#LoveWins

When the Obama administration announced the legalization of same-sex marriage on June 26, 2015, reactions from Americans—positive and negative—rippled across social media.

This archive presents the announcements on Instagram and Twitter to highlight how the design of these platforms shapes the political discourse online.

On **Instagram**, click on “view all 6,828 comments” and subsequently “load more comments” to pull up the replies.

On **Twitter**, the available replies are presented in expanded form. To navigate into an archived version of a sub-thread, click on the date of the thread’s first post.

Twitter counts “10k” replies on the original tweet, however most of them are not accessible. With some patience, Instagram users can read through all 6,828 replies, however they are shown in reverse order, without timestamps or any indication of relationship in between each other.

Although Instagram is supposedly an image sharing platform, users seem to be much more willing to discuss their views in replies that would exceed Twitter’s character count limit. Twitter folds long threads, sometimes visually isolating discussions from the topic originally presented, and creating mini echo chambers.

The archive is marked by absence in varying ways across both platforms. When a user deletes a comment in Instagram, it vanishes without a trace. Twitter, in contrast, uses a “tweet unavailable” placeholder, a gesture which allows others to guess who has retracted a tweet. Changing your mind on Instagram has less potential for shame or the display of weakness, whereas on Twitter users are forced to stand their ground and have their guard up. Instagram lessens the importance of comments and speech by design, while on Twitter, each tweet can itself start a newly branching thread.

Over the past eight years, Twitter has become the de facto social media platform for public political discourse in the United States—and the vernacular that has emerged and defined that discourse is closely related to design choices

whitehouse

69,702 likes 79w

whitehouse America should be very proud. #LoveWins

view all 6,828 comments

brandon_karl_ I love it!!! @dakota.37

dakota.37 @brandon_karl_ i love you ❤💛💚💙💜

brandon_karl_ I love you too ❤️💛💚💙💜🌈🌈🌈

brandon_karl_ @dakota.37

chloeother ❤️💛💚💙💜💗

micah_4_5 👎👎👎

micah_4_5 🚹🚺, not 🚹🚹 🚺🚺

its.ali.16 Well there's nothing you can do about it so go to hell @_solar_kiwi_jesusfreak_ there's nothing wrong with it

matthew.843 @zach.wall these freaks ^

jared_greer @im.aidy I just want you to know that I read all your comments and they made my night 😂🌈.

aiden.iguess @jared_greer thanks👌💞

aaron_tivadar Genuine question. Red White and Blue on the 4th of July?

galaxy_boi_alex yay!

kovach.04 Gender Dysphoria is a mental illness and it is being promoted on the basis of the LGBT movement. So, I'm just wondering, why does the White House promote a mental illness?

adnan_ali.m 👎👎

derpysquid9 👎🖕👎🖕👎 Terrible

nincyoun @nfritoli why do you even care? If you hate us that much get off and look at more important stuff like i don't know Brazilian accounts or something. That's what I would do.

nincyoun @nfritoli A Sign Up | ✕ g stuff like Arab terrorists and races. I just had to stifle vomit from your

This is an archived page from 1/4/2017, 5:15:46 AM
Archived with Webrecorder

President Obama @POTUS

Today is a big step in our march toward
Gay and lesbian couples now have the ri
marry, just like anyone else. #LoveWins

RETWEETS 425,334 LIKES 423,893

7:10 AM - 26 Jun 2015

10K 425K 424K

David Sims @davidlsims · 26 Jun 2015
@POTUS let's get married
1 6 55

sexy pinata @pilotbacon · 26 Jun 2015
@davidlsims @POTUS KISS
2 9

David Sims @davidlsims · 26 Jun 2015
@pilotbacon @POTUS #LoveWins
1

David Burge @iowahawkblog · 26 Jun 2015
@POTUS make sure you toast it at the next White House Ramada

> **President Obama** @POTUS
> Today is a big step in our march toward equality. Gay and lesbi
> now have the right to marry, just like anyone else. #LoveWins

39 473 293

Carmine Sabia @CarmineSabia · 27 Jun 2015
@iowahawkblog @POTUS YES.

Michael Hurley @michaelFhurley · 26 Jun 2015
.@POTUS Nice job Barry, but can you now get to the real issue: F
to take an aptitude test before using grocery self-checkouts.
4 10 38

Ric Duarte @Bruins_Diehard · 26 Jun 2015
@michaelFhurley @POTUS Hahaha You are just never satisfied, a
1 1

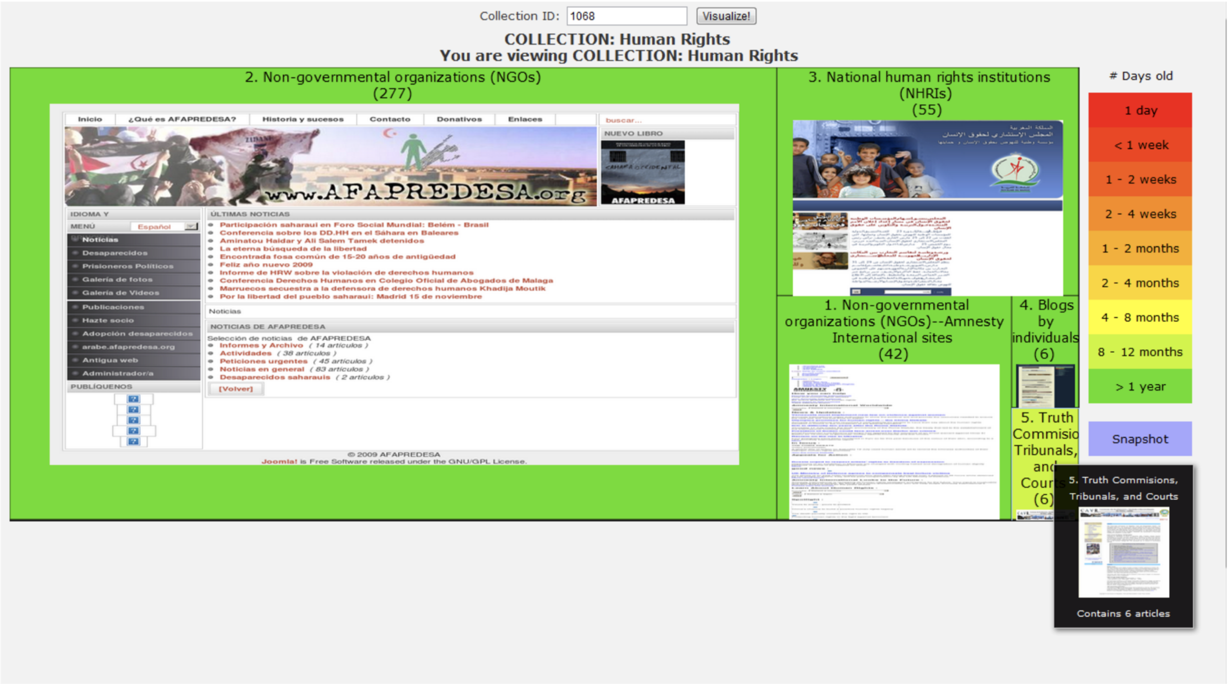
Michael Hurley @michaelFhurley · 26 Jun 2015
@Bruins_Diehard @POTUS I've got PLACES TO GO, Ric. People
1

Ric Duarte @Bruins_Diehard · 26 Jun 2015

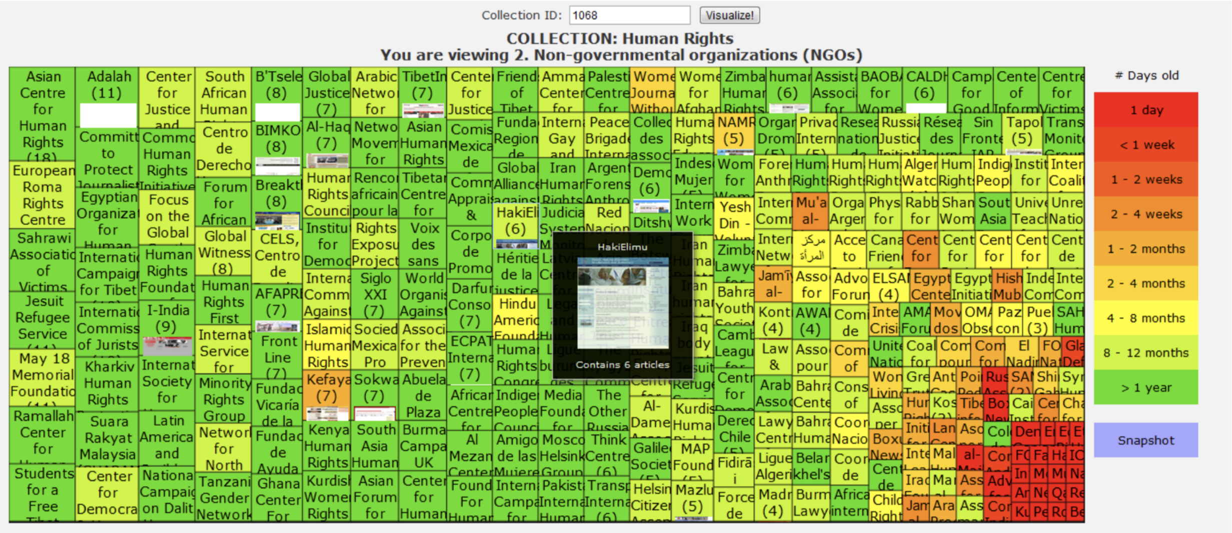
This is an archived page from 1/4/2017, 6:05:04 AM
Archived with Webrecorder

(c) #LoveWins

Fig. 61. Stories produced from Webrecorder as part of a prototype of its narrative features. Examples from: `http://archive.rhizome.org/narrative-archives/`. (Screenshots taken July 2018.)

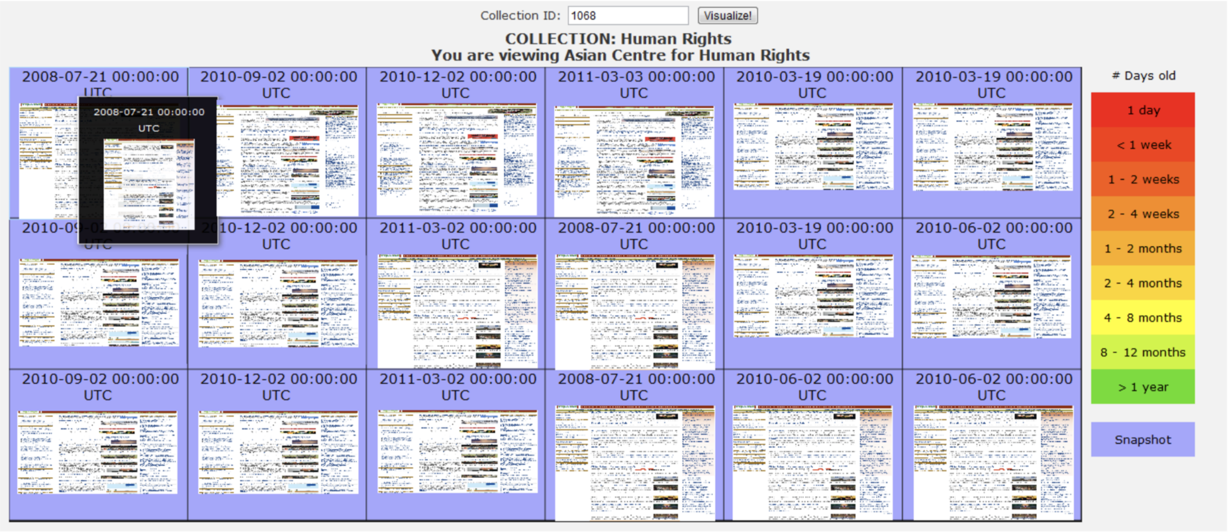

Collection ID: 1068
Visualize!
COLLECTION: Human Rights
You are viewing COLLECTION: Human Rights
2. Non-governmental organizations (NGOs) (277)
3. National human rights institutions (NHRIs) (55)
1. Non-governmental organizations (NGOs)--Amnesty International sites (42)
4. Blogs by individuals (6)
5. Truth Commissio Tribunals, and Cour (6)
5. Truth Commisions, Tribunals, and Courts
Contains 6 articles
www.AFAPREDESA.org
# Days old
1 day
< 1 week
1 - 2 weeks
2 - 4 weeks
1 - 2 months
2 - 4 months
4 - 8 months
8 - 12 months
> 1 year
Snapshot

(a) Category Level of Padia's TreeMap Visualzation.



(b) Web Page Level of Padia's TreeMap Visualzation.

Fig. 62. Padia et al.'s [256] interactive TreeMap visualization presented a web archive collection at three different levels.



Collection ID: 1068 Visualize!

COLLECTION: Human Rights
You are viewing Asian Centre for Human Rights

(c) Memento of Padia's TreeMap Visualzation.

Fig. 62. (Continued) Padia et al.'s [256] interactive TreeMap visualization presented a web archive collection at three different levels.

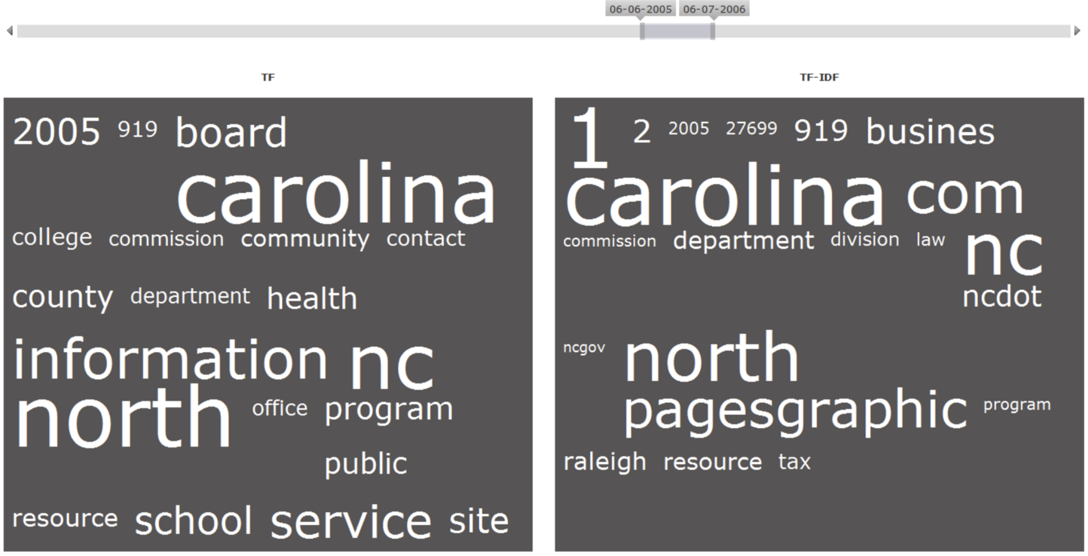
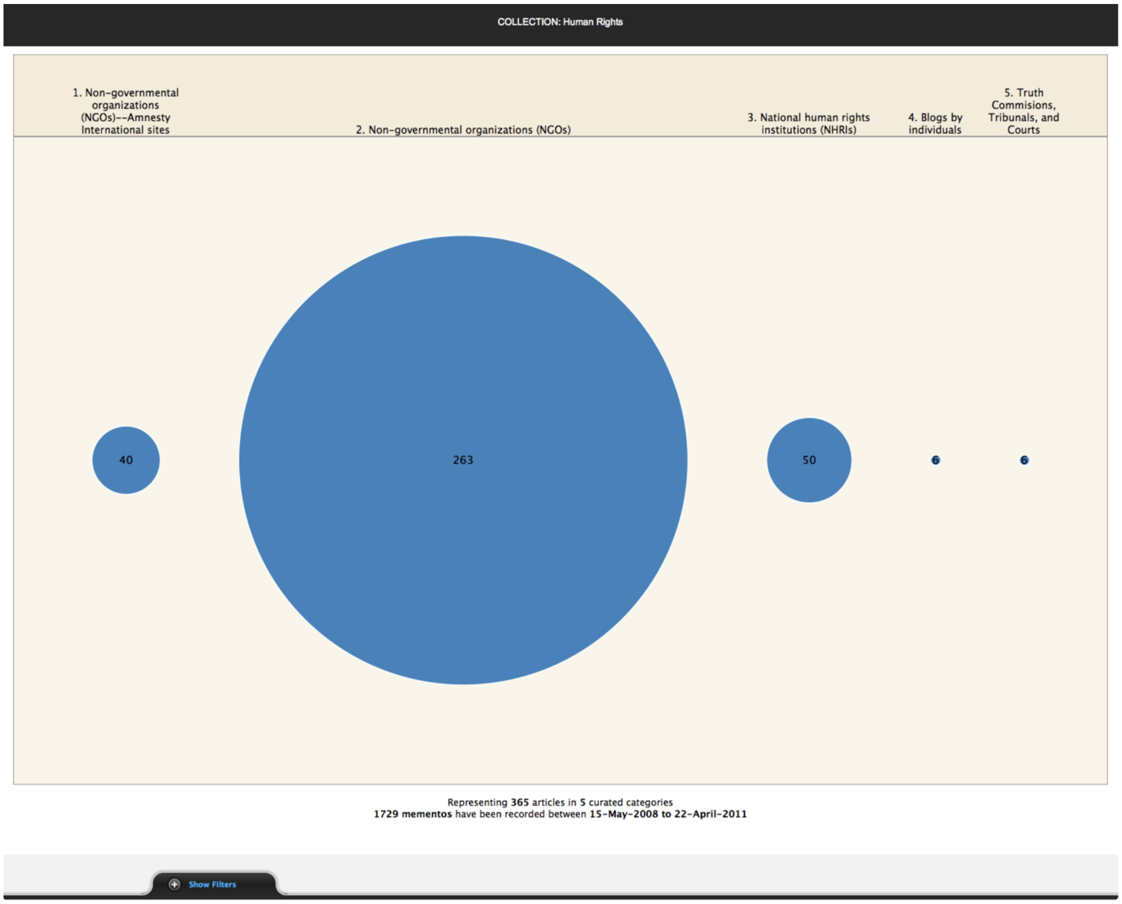


Fig. 63. Padia et al.'s [256] TimeCloud Visualization of Archive-It Collection 194: *North Carolina State Government Web Site Archive*. The left shows the top 20 terms weighted only by term frequency, whereas the cloud on the right shows the top 20 scoring terms by TF-IDF.



COLLECTION: Human Rights

Representing **365** articles in **5** curated categories
1729 mementos have been recorded between **15-May-2008 to 22-April-2011**

Fig. 64. Padia et al.'s [256] BubbleChart Visualization of Archive-It Collection 1068: *Human Rights*. All possible values for the *Group* seed metadata field in the given collection are shown across the top. The size of the bubble indicates the number of mementos in each group.

The TreeMap shown in Figure 62 displays Padia's interactive visualization of Archive-It Collection 1068: *Human Rights*. This visualization consists of multiple levels, allowing the user to navigate from web page category (Figure 62a) to original resource (Figure 62b) to memento (Figure 62c). They designed this layout to provide a hierarchical view of the collection, grouping original resources into categories and mementos under original resources.

Figure 63 displays Padia's TimeCloud for Archive-It Collection 194: *North Carolina State Government Web Site Archive*. Using the sliders on top to change the range of start and end time, a user can view the 20 most frequent terms encountered in the collection at that time period.

Padia also developed a non-interactive visualization that provides the user with a quick summary of the types of web pages found in the collection. Figure 64 displays this bubble chart. The size of the bubbles indicates how many mementos fit into each category. A user can click on the bubbles and get to the archived copies. This view provides the end-user with a quick idea of the number of mementos in a particular *Group*. The bottom of the chart provides a summary of the collection, including the number of original resources, mementos, groups, and the time span of the first and last memento-datetimes.

Padia's Image Plot visualization, seen in Figure 65, gives a more interactive view of the size of different groups. Again, the visualization breaks the collection into its defined categories, with thumbnails of the latest memento for each original resource shown below each category. The user can scroll through these thumbnails and click on them to visit Archive-It's page listing all mementos for that original resource. The visualization includes a smaller bar plot on the lower right. The size of each bar in this plot serves a similar purpose to the size of the bubbles shown in Figure 64, allowing the user to see how many mementos fit into each category. If a user hovers over the thumbnail of a memento, a wordle (word cloud) is generated, showing the top 20 terms for that page.

Costa and Silva [64] point out that timelines can help users understand the temporal nature of mementos. With this in mind, Padia developed a Timeline visualization, as seen in Figure 66. Each category has a corresponding color. Each line in the visualization represents a single original resource. Dots on each line indicate captured mementos. When a user hovers over a time, a window displays the other original resources that the web archive captured on the same day. Padia notes that this visualization allows users to see the structure of crawls and the evolution of the collection over time.

These visualizations require training and description to use correctly. The Bubble Chart and the Image Plot attempt to summarize a single metadata field for the collection but do not provide any additional insight into the collection content. The TreeMap visualization allows a user to

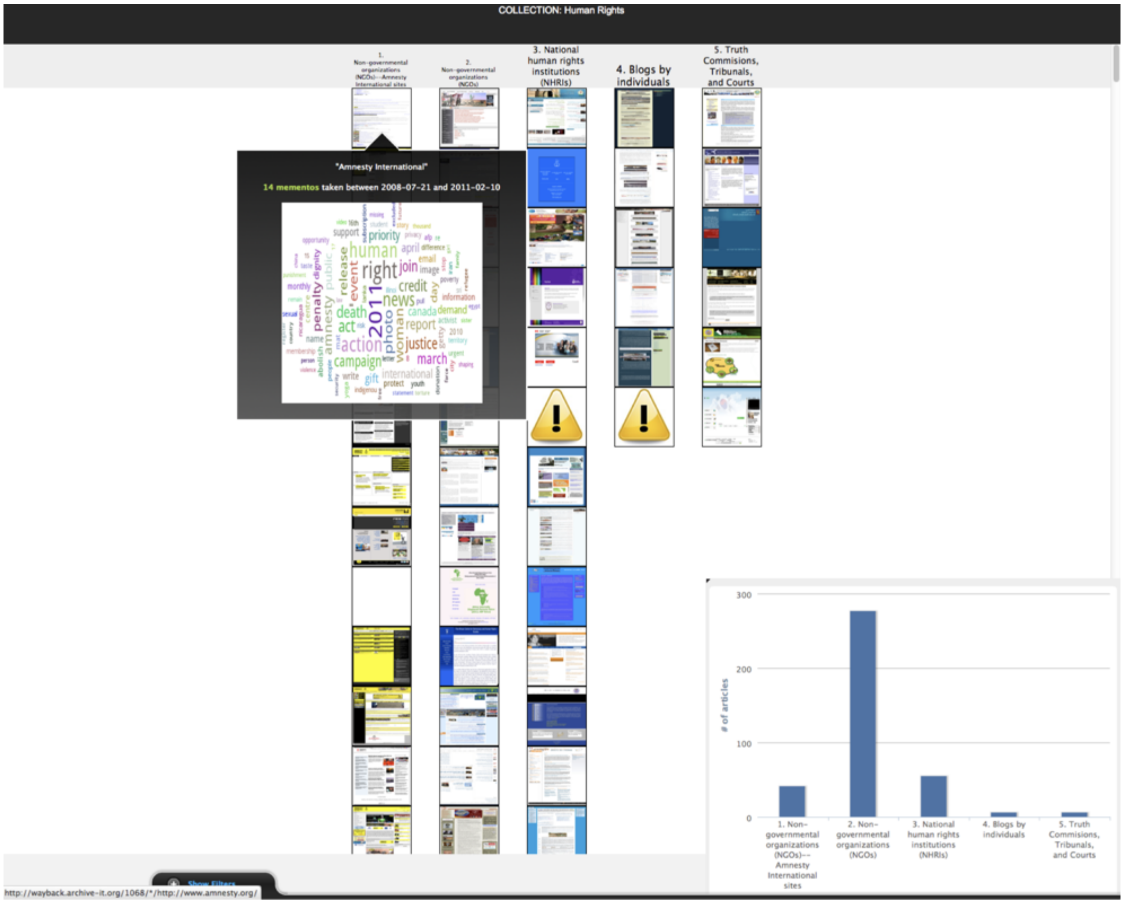
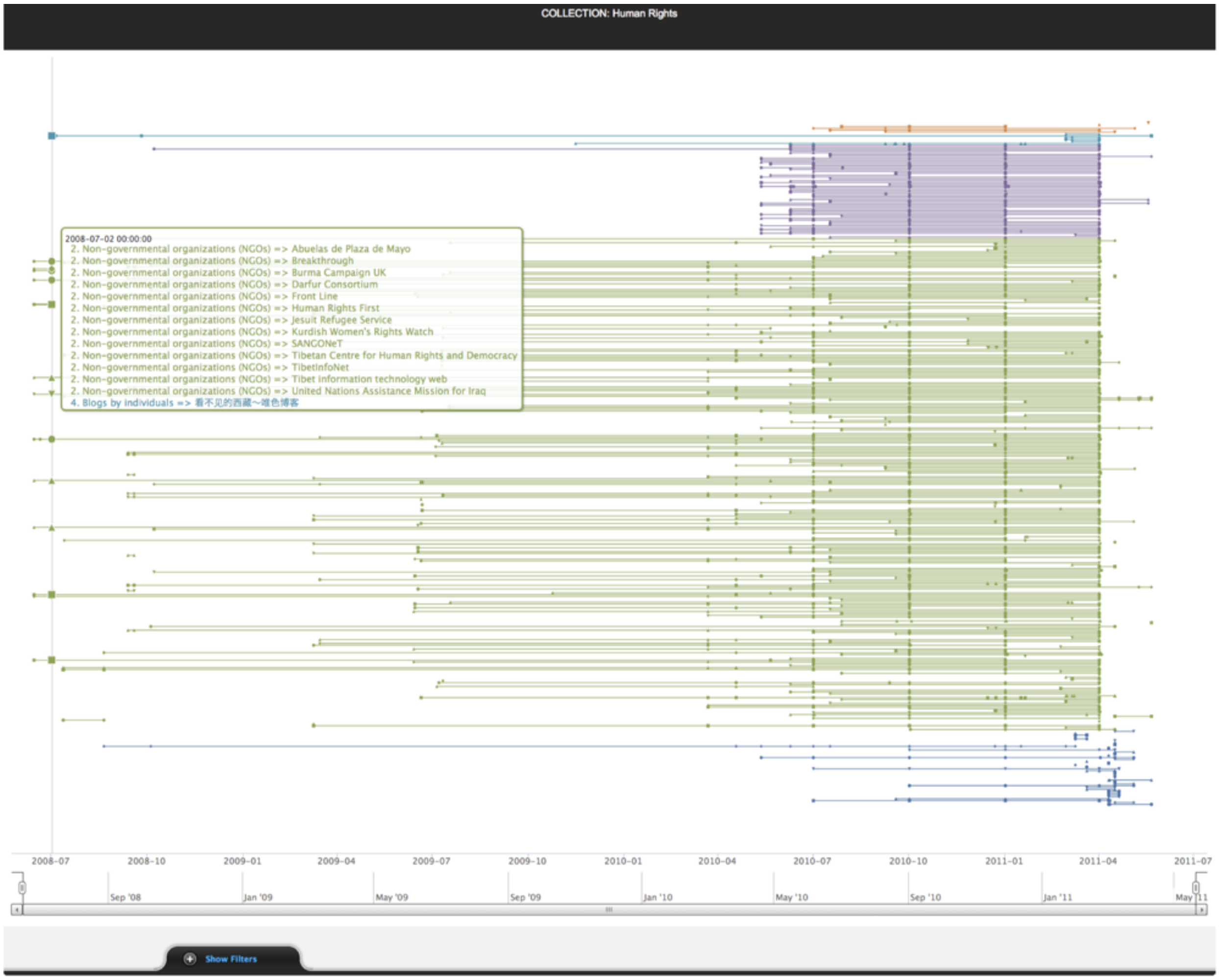

COLLECTION: Human Rights
1. Non-governmental organizations (NGOs)---Amnesty International sites
2. Non-governmental organizations (NGOs)
3. National human rights institutions (NHRIs)
4. Blogs by individuals
5. Truth Commisions, Tribunals, and Courts
"Amnesty International"
14 mementos taken between 2008-07-21 and 2011-02-10
human
right
2011
news
death
action
justice
campaign
amnesty
penalty
release
event
woman
photo
credit
join
priority
support
report
march
international
300
200
100
0
# of articles
1. Non-governmental organizations (NGOs)--Amnesty International sites
2. Non-governmental organizations (NGOs)
3. National human rights institutions (NHRIs)
4. Blogs by individuals
5. Truth Commisions, Tribunals, and Courts
Show Filters
http://wayback.archive-it.org/1068/*/http://www.amnesty.org/

Fig. 65. Padia et al.'s [256] Image Plot Visualization of Archive-It Collection 1068: *Human Rights*. All possible values for the *Group* seed metadata field in the given collection are shown across the top.

explore the mementos of the collection from page categories and seeds but requires much interaction. These visualizations try to visualize all of the mementos in a collection at once. Our goal is different. We seek to provide a story that allows the user to quickly get the general idea of a collection, saving the user time. We also apply visualization techniques (surrogates) that users are exposed to all of the time, thus requiring little training and less cognitive load.

3.5 STORYTELLING WITH WEB ARCHIVES

AlNoamany et al. [7, 11, 12, 13, 13, 14, 14] pioneered the use of social media storytelling for web archive summarization. As shown in Figure 67, AlNoamany's work covers our entire storytelling model. Her work was instrumental in promoting the idea of combining exemplar



2008-07-02 00:00:00
2. Non-governmental organizations (NGOs) => Abuelas de Plaza de Mayo
2. Non-governmental organizations (NGOs) => Breakthrough
2. Non-governmental organizations (NGOs) => Burma Campaign UK
2. Non-governmental organizations (NGOs) => Darfur Consortium
2. Non-governmental organizations (NGOs) => Front Line
2. Non-governmental organizations (NGOs) => Human Rights First
2. Non-governmental organizations (NGOs) => Jesuit Refugee Service
2. Non-governmental organizations (NGOs) => Kurdish Women's Rights Watch
2. Non-governmental organizations (NGOs) => SANGONeT
2. Non-governmental organizations (NGOs) => Tibetan Centre for Human Rights and Democracy
2. Non-governmental organizations (NGOs) => TibetInfoNet
2. Non-governmental organizations (NGOs) => Tibet information technology web
2. Non-governmental organizations (NGOs) => United Nations Assistance Mission for Iraq
4. Blogs by individuals => 看不见的西藏~唯色博客

Fig. 66. Padia et al.'s [256] Timeline Visualization of Archive-It Collection 1068: *Human Rights*.

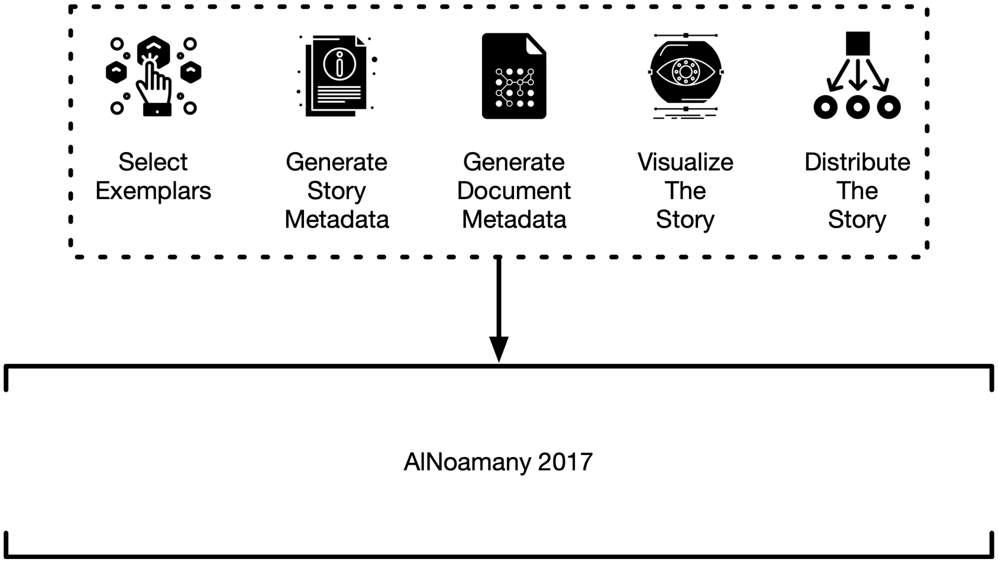
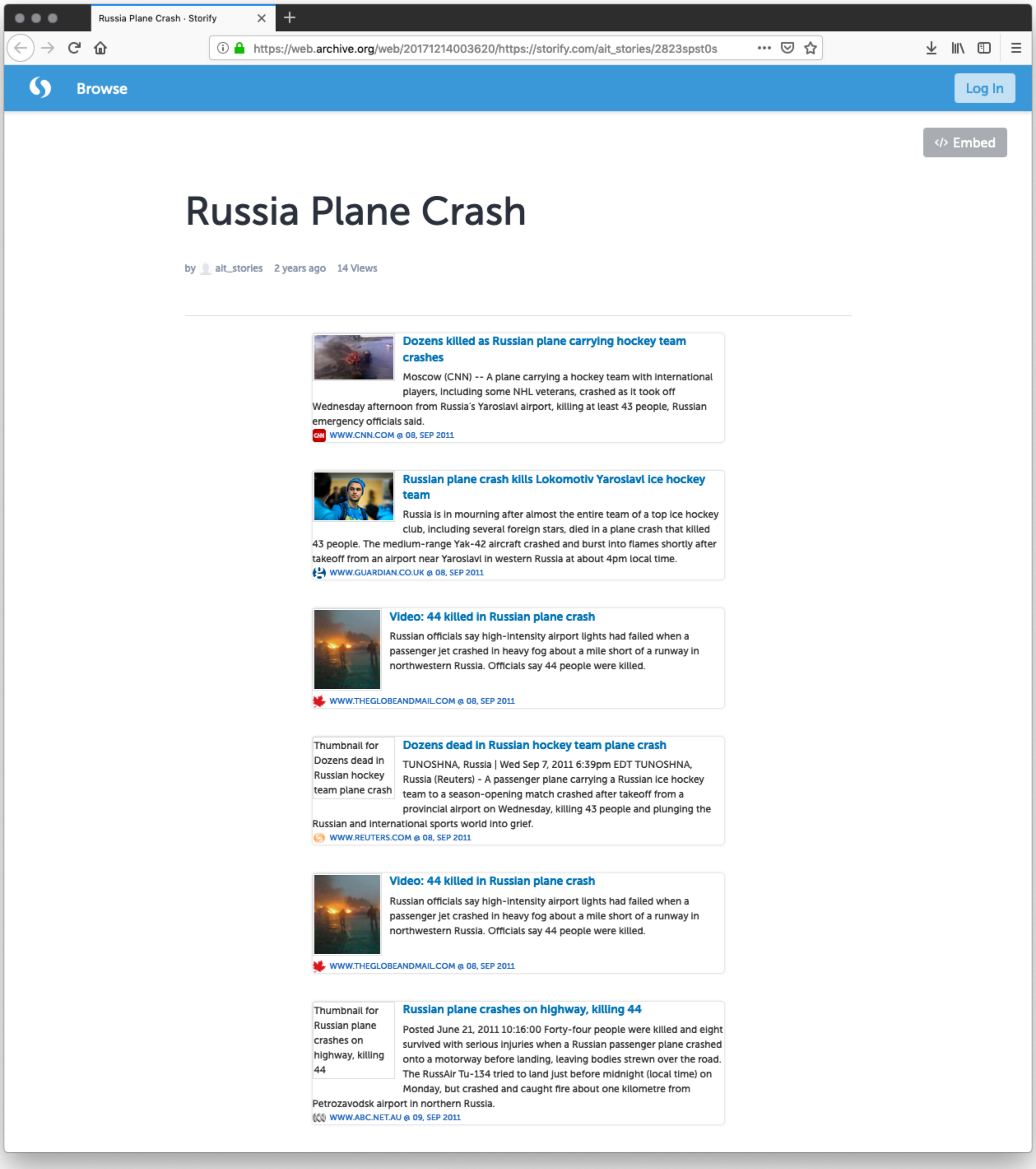


Fig. 67. AlNoamany et al.'s [11, 12, 13, 14] work combines all of the storytelling processes in our model.

mementos with surrogates as a visualization. Figure 68 displays four stories produced by her first *Dark and Stormy Archives* (DSA) system. These stories were submitted to and visualized by the now-defunct [289] social media storytelling service Storify.

She applied several approaches to select exemplars. To reduce the number of mementos from consideration, AlNoamany chose first to identify mementos that were off-topic [10]. Within web archive collections, the system crawls seeds at various intervals. Each memento is an observation of that seed at a particular point in time. Sometimes, the seed itself contains content that no longer matches the topic of the web archive collection, as seen in Figure 69. This example relays a somewhat complicated history for the web page at hamdeensabhay.com. It starts as on-topic but goes off-topic due to database errors, the inability of the company to pay for hosting, website changes, hacking, and, finally, domain expiration. Many seeds exhibit different kinds of off-topic behavior, and the next chapter contains more examples of this behavior. AlNoamany discovered that off-topic seeds exhibit several patterns. *Always On* seeds never go off-topic. *Step Function On* seeds are on-topic initially and stop being on-topic after some point; they never come back on-topic. *Step Function Off* seeds are off-topic initially and then start being on-topic past a certain point. *Oscillating* seeds are on and off-topic at various points throughout their life. Finally, *Always Off* seeds are never on-topic.

Using human review of 15,760 mementos across three Archive-It collections, AlNoamany et



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

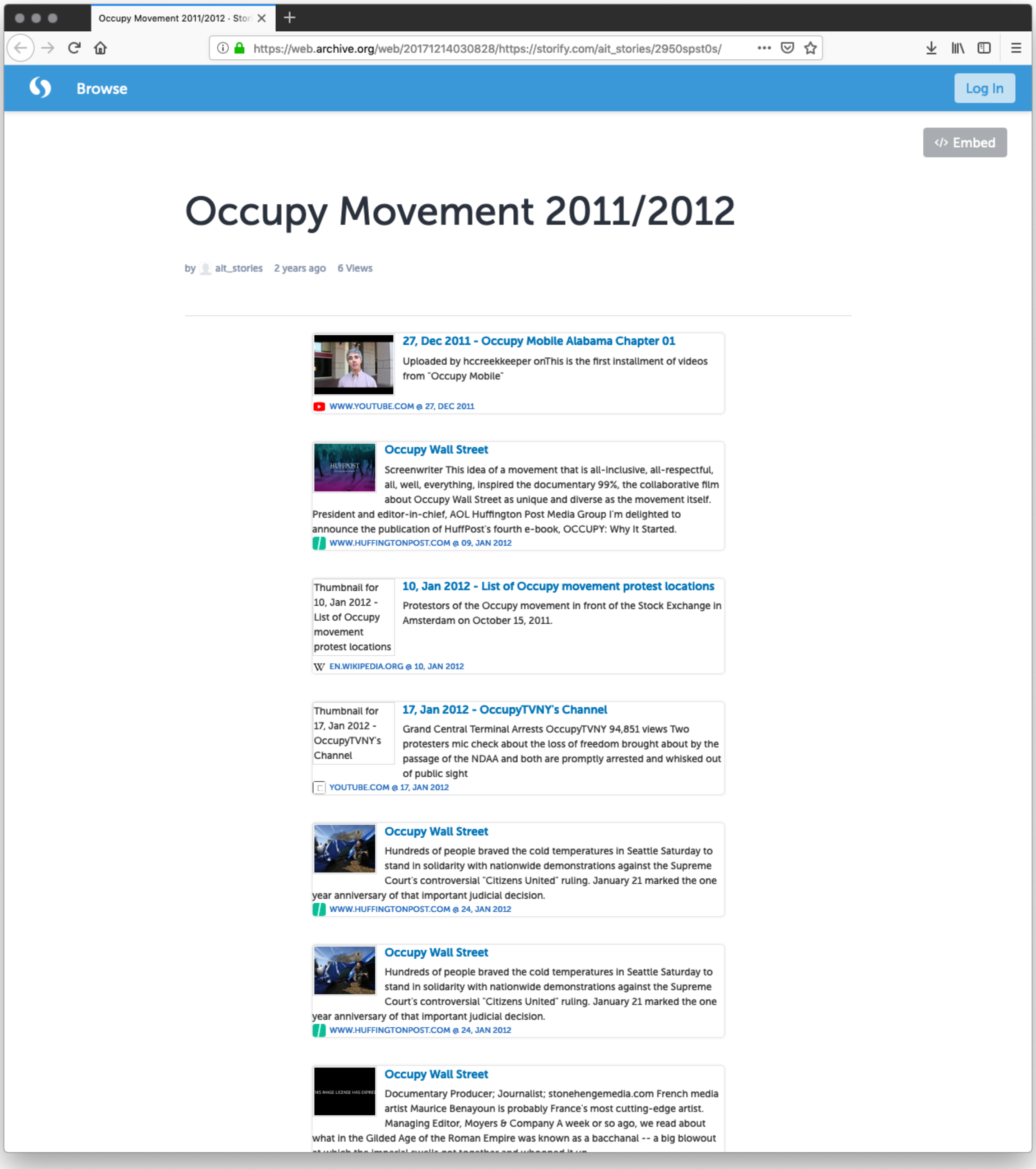
Russian plane crashes on highway, killing 44

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

(a) Russia Plane Crash Sept 7,2011

Fig. 68. Stories produced as part of AlNoamany's Dark and Stormy Archives [11], captured as mementos by the Internet Archive.


Occupy Movement 2011/2012 - Stor
https://web.archive.org/web/20171214030828/https://storify.com/ait_stories/2950spst0s/
Browse
Log In
</> Embed

Occupy Movement 2011/2012

by alt_stories 2 years ago 6 Views

27, Dec 2011 - Occupy Mobile Alabama Chapter 01

Uploaded by hccreekkeeper onThis is the first installment of videos from "Occupy Mobile"

WWW.YOUTUBE.COM @ 27, DEC 2011

Occupy Wall Street

Screenwriter This idea of a movement that is all-inclusive, all-respectful, all, well, everything, inspired the documentary 99%, the collaborative film about Occupy Wall Street as unique and diverse as the movement itself. President and editor-in-chief, AOL Huffington Post Media Group I'm delighted to announce the publication of HuffPost's fourth e-book, OCCUPY: Why It Started.

WWW.HUFFINGTONPOST.COM @ 09, JAN 2012

Thumbnail for 10, Jan 2012 - List of Occupy movement protest locations

10, Jan 2012 - List of Occupy movement protest locations

Protestors of the Occupy movement in front of the Stock Exchange in Amsterdam on October 15, 2011.

EN.WIKIPEDIA.ORG @ 10, JAN 2012

Thumbnail for 17, Jan 2012 - OccupyTVNY's Channel

17, Jan 2012 - OccupyTVNY's Channel

Grand Central Terminal Arrests OccupyTVNY 94,851 views Two protesters mic check about the loss of freedom brought about by the passage of the NDAA and both are promptly arrested and whisked out of public sight

YOUTUBE.COM @ 17, JAN 2012

Occupy Wall Street

Hundreds of people braved the cold temperatures in Seattle Saturday to stand in solidarity with nationwide demonstrations against the Supreme Court's controversial "Citizens United" ruling. January 21 marked the one year anniversary of that important judicial decision.

WWW.HUFFINGTONPOST.COM @ 24, JAN 2012

Occupy Wall Street

Hundreds of people braved the cold temperatures in Seattle Saturday to stand in solidarity with nationwide demonstrations against the Supreme Court's controversial "Citizens United" ruling. January 21 marked the one year anniversary of that important judicial decision.

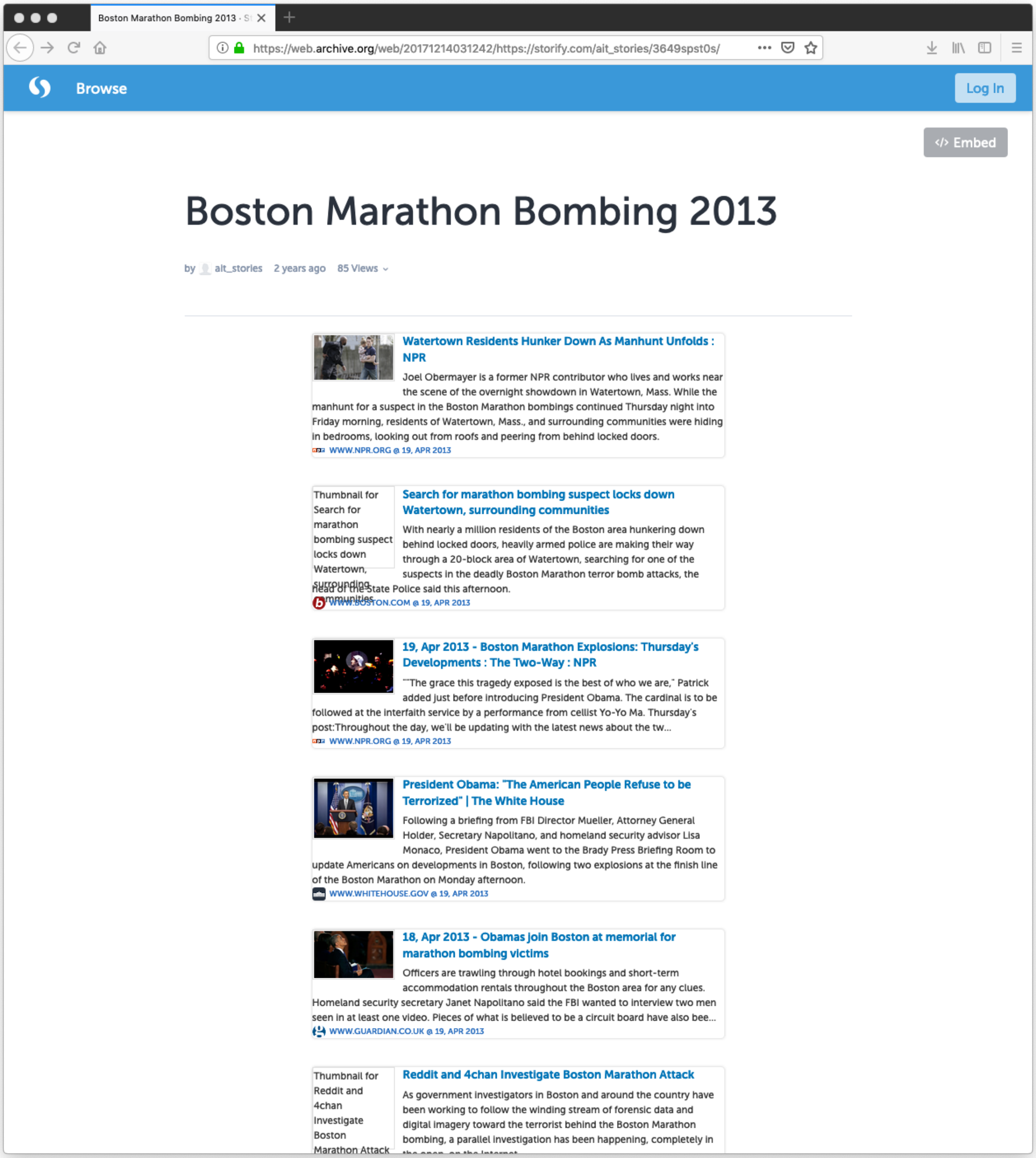
WWW.HUFFINGTONPOST.COM @ 24, JAN 2012

Occupy Wall Street

Documentary Producer; Journalist; stonehengemedia.com French media artist Maurice Benayoun is probably France's most cutting-edge artist. Managing Editor, Moyers & Company A week or so ago, we read about what in the Gilded Age of the Roman Empire was known as a bacchanal -- a big blowout at which the imperial swells got together and whooped it up

(b) Occupy Movement 2011/2012

Fig. 68. (Continued) Stories produced as part of AlNoamany's Dark and Stormy Archives [11], captured as mementos by the Internet Archive.


Boston Marathon Bombing 2013 · S
https://web.archive.org/web/20171214031242/https://storify.com/alt_stories/3649spst0s/
Browse
Log In
</> Embed

Boston Marathon Bombing 2013

by alt_stories 2 years ago 85 Views

Watertown Residents Hunker Down As Manhunt Unfolds : NPR

Joel Obermayer is a former NPR contributor who lives and works near the scene of the overnight showdown in Watertown, Mass. While the manhunt for a suspect in the Boston Marathon bombings continued Thursday night into Friday morning, residents of Watertown, Mass., and surrounding communities were hiding in bedrooms, looking out from roofs and peering from behind locked doors.

WWW.NPR.ORG @ 19, APR 2013

Thumbnail for Search for marathon bombing suspect locks down Watertown, surrounding communities

Search for marathon bombing suspect locks down Watertown, surrounding communities

With nearly a million residents of the Boston area hunkering down behind locked doors, heavily armed police are making their way through a 20-block area of Watertown, searching for one of the suspects in the deadly Boston Marathon terror bomb attacks, the head of the State Police said this afternoon.

WWW.BOSTON.COM @ 19, APR 2013

19, Apr 2013 - Boston Marathon Explosions: Thursday's Developments : The Two-Way : NPR

""The grace this tragedy exposed is the best of who we are," Patrick added just before introducing President Obama. The cardinal is to be followed at the interfaith service by a performance from cellist Yo-Yo Ma. Thursday's post:Throughout the day, we'll be updating with the latest news about the tw...

WWW.NPR.ORG @ 19, APR 2013

President Obama: "The American People Refuse to be Terrorized" | The White House

Following a briefing from FBI Director Mueller, Attorney General Holder, Secretary Napolitano, and homeland security advisor Lisa Monaco, President Obama went to the Brady Press Briefing Room to update Americans on developments in Boston, following two explosions at the finish line of the Boston Marathon on Monday afternoon.

WWW.WHITEHOUSE.GOV @ 19, APR 2013

18, Apr 2013 - Obamas join Boston at memorial for marathon bombing victims

Officers are trawling through hotel bookings and short-term accommodation rentals throughout the Boston area for any clues. Homeland security secretary Janet Napolitano said the FBI wanted to interview two men seen in at least one video. Pieces of what is believed to be a circuit board have also bee...

WWW.GUARDIAN.CO.UK @ 19, APR 2013

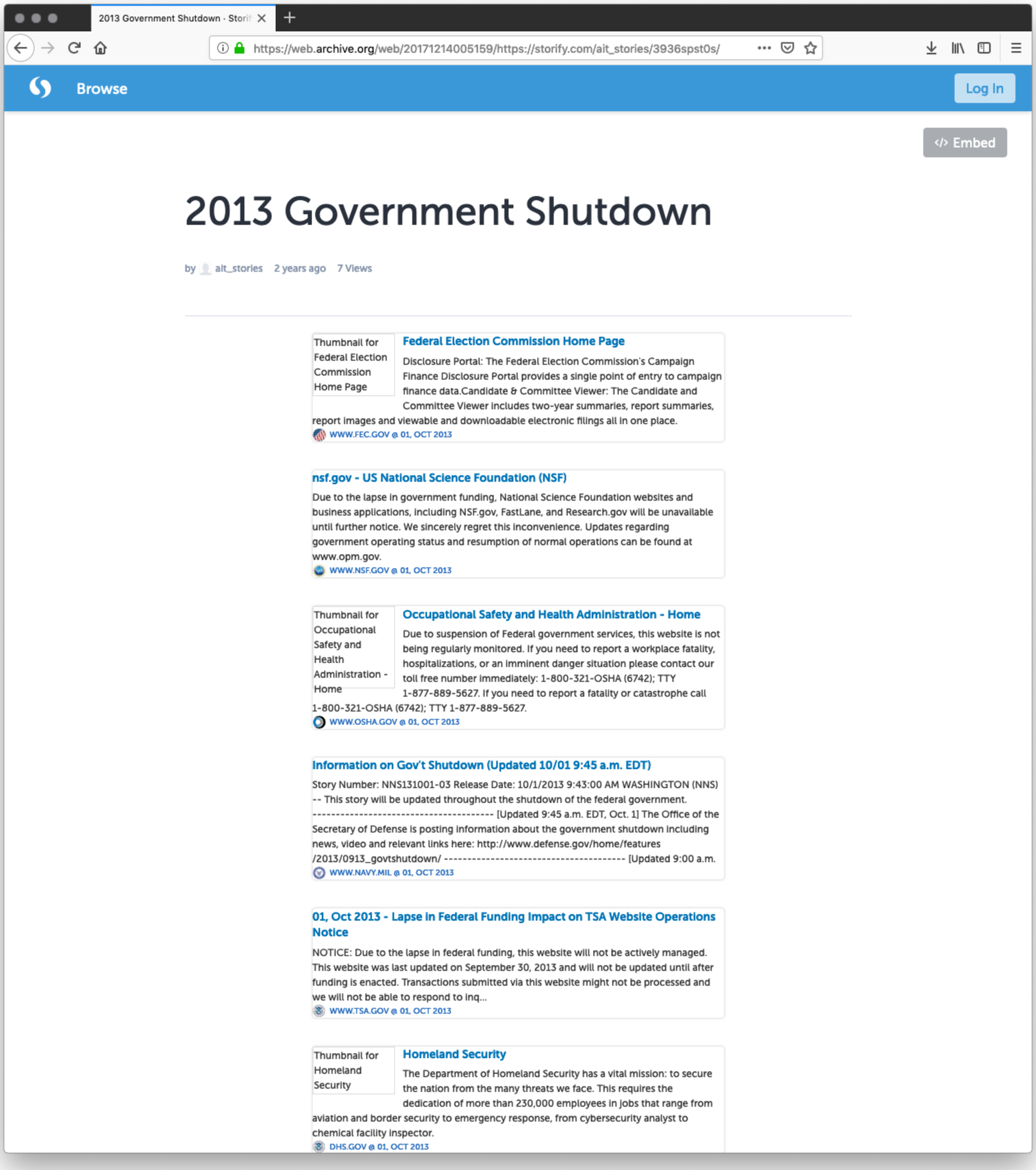
Thumbnail for Reddit and 4chan Investigate Boston Marathon Attack

Reddit and 4chan Investigate Boston Marathon Attack

As government investigators in Boston and around the country have been working to follow the winding stream of forensic data and digital imagery toward the terrorist behind the Boston Marathon bombing, a parallel investigation has been happening, completely in the open, on the Internet.

(c) 2013 Boston Marathon Bombing

Fig. 68. (Continued) Stories produced as part of AlNoamany's Dark and Stormy Archives [11], captured as mementos by the Internet Archive.


2013 Government Shutdown · Storif
https://web.archive.org/web/20171214005159/https://storify.com/ait_stories/3936spst0s/
Browse
Log In
</> Embed

2013 Government Shutdown

by ait_stories 2 years ago 7 Views

Thumbnail for Federal Election Commission Home Page

Federal Election Commission Home Page

Disclosure Portal: The Federal Election Commission's Campaign Finance Disclosure Portal provides a single point of entry to campaign finance data.Candidate & Committee Viewer: The Candidate and Committee Viewer includes two-year summaries, report summaries, report images and viewable and downloadable electronic filings all in one place.

WWW.FEC.GOV @ 01, OCT 2013

nsf.gov - US National Science Foundation (NSF)

Due to the lapse in government funding, National Science Foundation websites and business applications, including NSF.gov, FastLane, and Research.gov will be unavailable until further notice. We sincerely regret this inconvenience. Updates regarding government operating status and resumption of normal operations can be found at www.opm.gov.

WWW.NSF.GOV @ 01, OCT 2013

Thumbnail for Occupational Safety and Health Administration - Home

Occupational Safety and Health Administration - Home

Due to suspension of Federal government services, this website is not being regularly monitored. If you need to report a workplace fatality, hospitalizations, or an imminent danger situation please contact our toll free number immediately: 1-800-321-OSHA (6742); TTY 1-877-889-5627. If you need to report a fatality or catastrophe call 1-800-321-OSHA (6742); TTY 1-877-889-5627.

WWW.OSHA.GOV @ 01, OCT 2013

Information on Gov't Shutdown (Updated 10/01 9:45 a.m. EDT)

Story Number: NNS131001-03 Release Date: 10/1/2013 9:43:00 AM WASHINGTON (NNS) -- This story will be updated throughout the shutdown of the federal government. ------------------------------------- [Updated 9:45 a.m. EDT, Oct. 1] The Office of the Secretary of Defense is posting information about the government shutdown including news, video and relevant links here: http://www.defense.gov/home/features /2013/0913_govtshutdown/ ------------------------------------- [Updated 9:00 a.m.

WWW.NAVY.MIL @ 01, OCT 2013

01, Oct 2013 - Lapse in Federal Funding Impact on TSA Website Operations Notice

NOTICE: Due to the lapse in federal funding, this website will not be actively managed. This website was last updated on September 30, 2013 and will not be updated until after funding is enacted. Transactions submitted via this website might not be processed and we will not be able to respond to inq...

WWW.TSA.GOV @ 01, OCT 2013

Thumbnail for Homeland Security

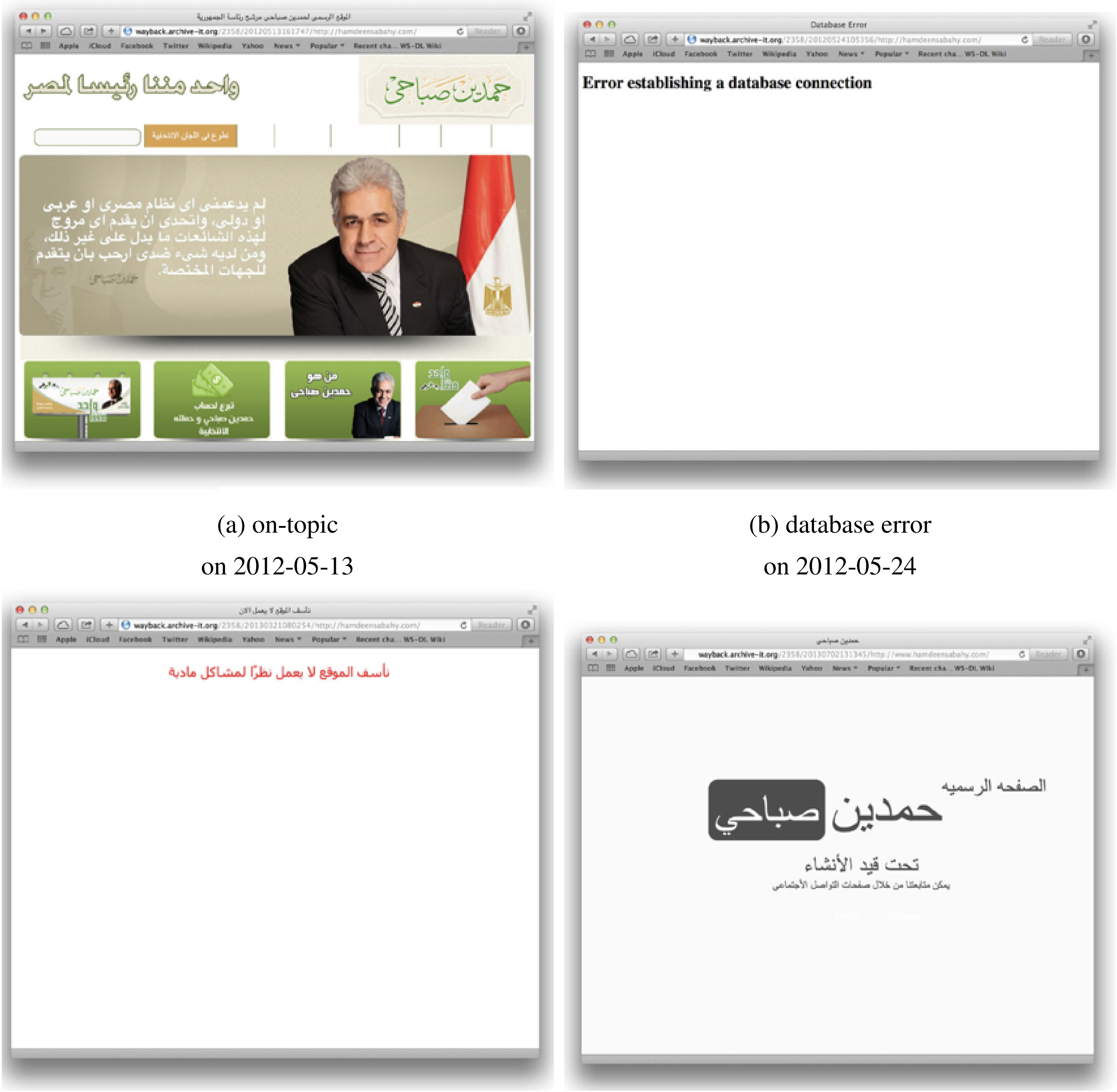
Homeland Security

The Department of Homeland Security has a vital mission: to secure the nation from the many threats we face. This requires the dedication of more than 230,000 employees in jobs that range from aviation and border security to emergency response, from cybersecurity analyst to chemical facility inspector.

DHS.GOV @ 01, OCT 2013

(d) 2013 Government Shutdown

Fig. 68. (Continued) Stories produced as part of AlNoamany's Dark and Stormy Archives [11], captured as mementos by the Internet Archive.


wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/
Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
واحد مننا رئيسا لمصر
حمدين صباحي
Database Error
wayback.archive-it.org/2358/20120524105356/http://hamdeensabahy.com/
wayback.archive-it.org/2358/20130321080254/http://hamdeensabahy.com/
wayback.archive-it.org/2358/20130702131345/http://www.hamdeensabahy.com/

لم يدعمني اى نظام مصري او عربي او دولي، واتحدى ان يقدم اى مروج لهذه الشائعات ما يدل على غير ذلك، ومن لديه شيء ضدى ارحب بان يتقدم للجهات المختصة.

Error establishing a database connection

نأسف الموقع لا يعمل نظرًا لمشاكل مادية

الصفحه الرسميه حمدين صباحي

تحت قيد الأنشاء

يمكن متابعتنا من خلال صفحات التواصل الأجتماعي

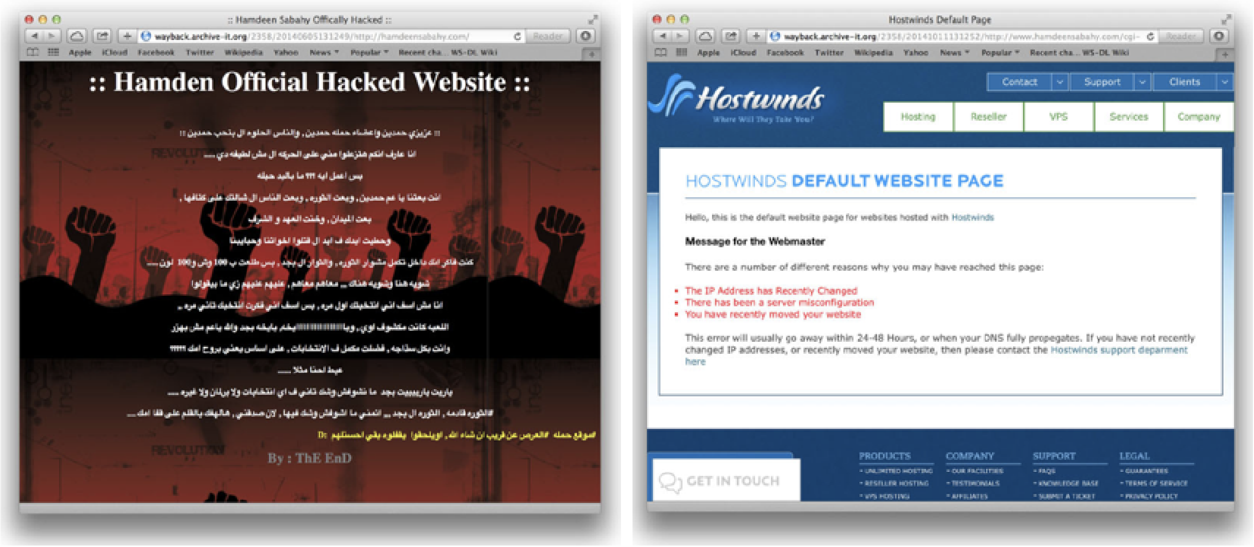
(a) on-topic
on 2012-05-13

(b) database error
on 2012-05-24

(c) financial issues
on 2013-03-21

(d) under construction
on 2013-07-02

Fig. 69. An example from AlNoamany et al. [10] of the hamdeensabhay.com seed going off-topic. Each thumbnail corresponds to a memento captured by Archive-It because the site is re-crawled periodically.



:: Hamden Official Hacked Website ::

By : ThE EnD

HOSTWINDS DEFAULT WEBSITE PAGE

Hello, this is the default website page for websites hosted with Hostwinds

Message for the Webmaster

There are a number of different reasons why you may have reached this page:

- The IP Address has Recently Changed
- There has been a server misconfiguration
- You have recently moved your website

This error will usually go away within 24-48 Hours, or when your DNS fully propegates. If you have not recently changed IP addresses, or recently moved your website, then please contact the Hostwinds support deparment here

:: Hamdeen Sabahy Offically Hacked ::
wayback.archive-it.org/2358/20140605131249/http://hamdeensabahy.com/
Hostwinds Default Page
wayback.archive-it.org/2358/20141011131252/http://www.hamdeensabahy.com/cgi-
Hostwinds
Where Will They Take You?
Contact
Support
Clients
Hosting
Reseller
VPS
Services
Company
GET IN TOUCH
PRODUCTS
COMPANY
SUPPORT
LEGAL

(e) hacked
on 2014-06-05

(f) domain expiration
on 2014-10-11

Fig. 69. (Continued) An example from AlNoamany et al. [10] of the hamdeensabhay.com seed going off-topic. Each thumbnail corresponds to a memento captured by Archive-It because the site is re-crawled periodically.

al. manually created a gold standard dataset [152] to evaluate several algorithms for detecting off-topic data. This gold standard dataset contains the URI-M of each memento and whether it is on or off-topic compared to the rest of the collection. Separately, they selected the first memento from each seed's TimeMap for these collections and computed similarity measures comparing each subsequent memento to the first. The similarity measures they used were cosine similarity, Jaccard coefficient, the intersection of most frequent terms, web-based kernel function, change in size by bytes, and change in size by words. Comparing these results to the gold standard dataset, they computed accuracy and F_1 scores. They discovered that cosine similarity was the best algorithm for discovering off-topic mementos. A combination of word count and cosine similarity performed even better. In Chapter 4.2, we evaluate other similarity measures and more precise threshold values beyond her study to see if we can improve upon this off-topic detection and, in our case, word count performed best.

AlNoamany et al. [9] evaluated both 3,109 Archive-It Collections and 14,568 Storify social media stories. In Storify, users could employ several elements as part of their narrative, including links represented as social cards, images, video, and freeform text. AlNoamany discovered that 48% of stories contained no text elements. When grouping elements from all stories, only 8.1% contained text, making text elements quite rare. In contrast, 70.8% of elements were links. Across all stories, the top 5 domains from links are to social media sites, with 74.92% to twitter.com, 29.09% to youtube.com, 12.64% to instagram.com, 12.37% to facebook.com, and 7.41% to flickr.com.

Interestingly enough, most of the tweets in these stories contained their own embedded resources, producing a nesting relationship whereby the tweet's card contains its own social cards, images, and videos. Using the number of views for each story as a proxy for popularity, they discovered that popular stories contain a median of 10 images and a median of 20 links. Overall, popular stories have a median value of 28 elements.

They evaluated all collections Archive-It to understand the platform's makeup. They discovered that the largest collection was collection 3572, *Government of Canada Publications*. In 2015 it contained 123,647 seeds (as of July 12, 2021, it now contains 339,166 seeds). She noted how some collections have mementos created before Archive-It came into existence because those collections' archivists imported mementos from other web archives. They found that the top 5 domains for Archive-It seeds based on the number of collections they appear in were facebook.com, twitter.com, youtube.com, blogspot.com, and wordpress.com. Comparing the two, they discover that Storify and Archive-It do diverge quite a bit. Most URIs in Storify came from the .com TLD, whereas many URIs in Archive-It come from .gov and .edu domains. Archive-It is used to preserve the content of government agencies and educational institutions, often to meet some regulatory requirement. This study determined that links rendered as social cards are a common element used in storytelling. They also established that a good target number of elements for a summary story should be 28.

AlNoamany et al. [11] formalized the concepts needed to produce stories from Archive-It collections. They detail a taxonomy of four different types of stories possible using the dimensions of time and seed:

- **Fixed Page, Fixed Time (FPFT)** – In this case, the seed is the same for all elements in the story, and all mementos are from the same time period. This type of story can be used to view the same seed but captured from different physical locations, languages, user-agents, or other environment-specific changes.

- **Sliding Page, Sliding Time (SPST)** – For these stories, both the seed and the time may vary across the life of the collection.

- **Fixed Page, Sliding Time (FPST)** – This shows how a single seed changes over time. This type of story is useful for watching an unfolding news story.

- **Sliding Page, Fixed Time (SPFT) –** Here, the seed changes, but the time stays fixed. This story documents different sources reporting for the same time period.

They note that FPFT stories are not currently possible with the technology provided by web archives. AlNoamany visualized the DSA framework as a filter, or sieve, as seen in Figure 70. Each step removes mementos from consideration until only around 28 representative mementos remain. They use the number 28 because it is the number of elements found in their earlier study's most popular Storify stories.

Before we discuss her algorithm for storytelling, we must first cover the concept of **path depth**. The path depth for each seed URI consists of the number of items separated by slashes after the domain name. If the path consists of a query string, 1 is added to the path depth, similar to McCown et al. [218]. If the last item in the path consists of a known default page (e.g., index.html), then we subtract 1 from the path depth. Default pages are determined by a list of well-known default pages[17]. Typically, pages with a higher path depth discuss more specific topics than those with a lower path depth [57]. AlNoamany makes use of this concept to select mementos covering more specific topics from the collection.

Given a web archive collection, AlNoamany's Algorithm for producing these stories is as follows:

1. Mark mementos that are deemed off-topic from the collection. Exclude them from further consideration.
2. Mark mementos that are near-duplicates to each other. Exclude them from further consideration.
3. Mark mementos that do not use the English language. Exclude them and only keep the English language mementos.
4. Slice the collection by memento-datetime.
5. Cluster the mementos within each slice based on similarity of content.
6. Select the best representative memento from each cluster by using a quality equation.
7. Order the selected mementos by publication date or memento-datetime.

[17] https://support.tigertech.net/index-file-names

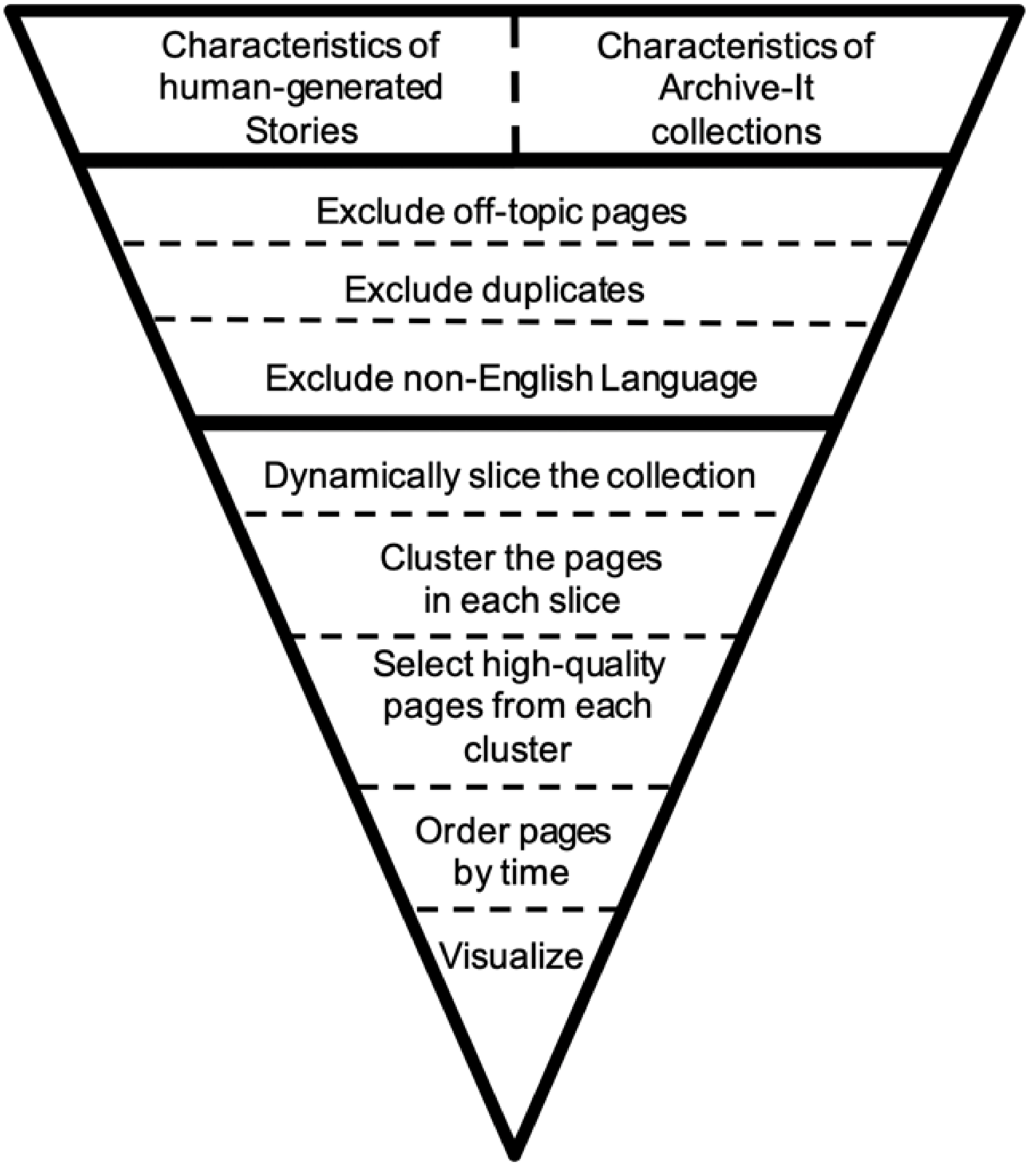

Characteristics of human-generated Stories
Characteristics of Archive-It collections
Exclude off-topic pages
Exclude duplicates
Exclude non-English Language
Dynamically slice the collection
Cluster the pages in each slice
Select high-quality pages from each cluster
Order pages by time
Visualize

Fig. 70. The Dark and Stormy Archives framework, visualized as a filter in AlNoamany's presentations [8].

8. Visualize the selected mementos with Storify.

AlNoamany built Step 1 on her earlier off-topic detection work. Step 2 uses Simhash to compute hashes and measure their distances from each other. Step 3 uses the Java language detection library developed by Shuyo [284]. Step 4 slices the collection into several slices depending on the number of mementos N in the collection, as shown in Equation 10:

$$S = \begin{cases} \lceil 28 + \log_{10}(N) \rceil & \text{if } N > 28 \\ N & \text{if } N \leq 28 \end{cases} \tag{10}$$

Clustering in Step 5 is performed by computing the Simhash distance between different mementos and executing DBSCAN [84].

Several factors inform the quality equation used for Step 6. AlNoamany noted that the surrogates produced by Storify for news articles had more topically relevant content than those produced by a news website's main page. Such articles have a greater URI path depth. They also noted that social media URIs produced poor surrogates with Storify. Ideally, surrogates of social media posts would contain content about that post. However, surrogates of social media sites often contain information about how to sign up for an account at the site rather than containing the post's content. AlNoamany uses a URI classification scheme from her previous work with Padia [257] to detect social media sites and give them a lower weight compared to news sites. Finally, not all mementos are of the same quality. Due to issues with crawling, some mementos are missing images, stylesheets, and scripts. Brunelle et al. [45] covered this extensively and even produced a method of evaluating the quality of mementos, and AlNoamany incorporates Brunelle et al.'s **Memento Damage** metric into this quality equation. The equation is as follows:

$$M_q = (1 - w_d \times D_m) + w_l \times M_l + w_c \times M_c \tag{11}$$

where D_m is memento damage, M_l is the path depth of the URI, and M_c is the weight of Padia's URI category. The other terms in the equation are weights whose values have been determined empirically as $w_l = 0.45$, $w_d = 0.40$, and $w_c = 0.15$.

For Step 7, they attempt to extract the publication date using natural language processing and fall back to memento-datetime otherwise. Finally, in Step 8, Storify renders the mementos to visualize the story. The Storify API provides some tools for augmenting or changing the behavior of surrogates. By default, Storify would use a favicon and domain from the web archive. AlNoamany overrode these values with those of the seed URI and appended the memento-datetime.

This dissertation refers to AlNoamany's Algorithm as **DSA1** because it is the first attempt at solving this problem. Figure 71 shows how her proof-of-concept system maps to our model. Her

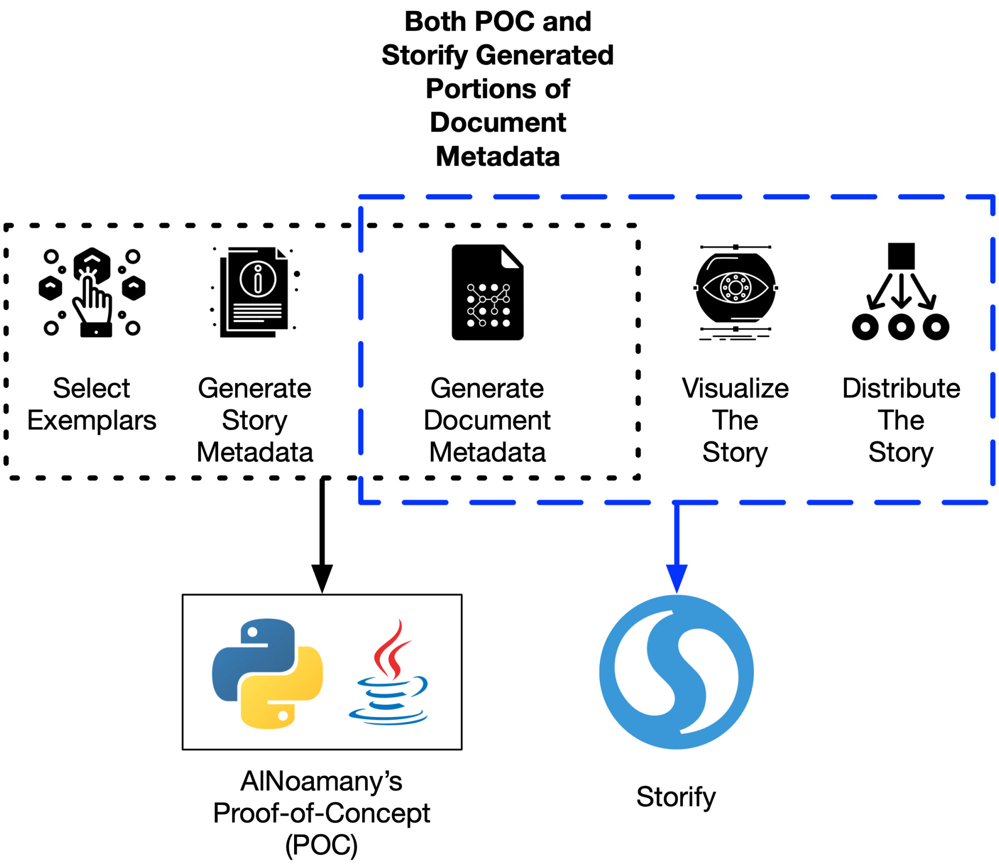

Both POC and Storify Generated Portions of Document Metadata
Select Exemplars
Generate Story Metadata
Generate Document Metadata
Visualize The Story
Distribute The Story
AlNoamany's Proof-of-Concept (POC)
Storify

Fig. 71. How AlNoamany's proof-of-concept system maps to our storytelling model.

Java and Python code took care of selecting exemplars through steps 1 - 7 of her algorithm. It also scraped Archive-It for the collection title to generate story metadata. In Step 8, she augmented Storify's surrogates by applying additional document metadata. Finally, Storify took care of visualizing the story and distributing it.

To evaluate the output of this algorithm, the authors first needed a gold standard against which to test. They asked expert archivists to select representative mementos from their Archive-It collections using the following guidelines:

- mementos must come from within the collection
- mementos must be on-topic
- mementos must be English
- the number of mementos must be around 28
- stories can be of one of the types mentioned above
- mementos must not be a near-duplicate of another memento in the story
- mementos of better quality are preferred
- stories can be from a specific time period in the collection if the collection spans many years

The archivists generated 23 stories from 10 different collections, hereafter referred to as *human-generated stories*. The authors then executed the DSA framework against the same stories to produce *automatically-generated stories*. In addition, the authors created *randomly-generated stories* for each collection. Finally, they selected a single memento and repeated it 28 times to produce a *poorly-generated story* for each collection.

With the stories now available, AlNoamany et al. turned to MT to recruit participants for 1,035 tasks. Each task presented a participant with two sets of two stories, side by side. They instructed the participant to consider each story as an overview of a topic and choose their preferred story. Unknown to the participant, AlNoamany presented one of the following sets of tasks to them:

- human vs. automatic, human vs. poor
- human vs. random, human vs. poor
- random vs. automatic, automatic vs. poor

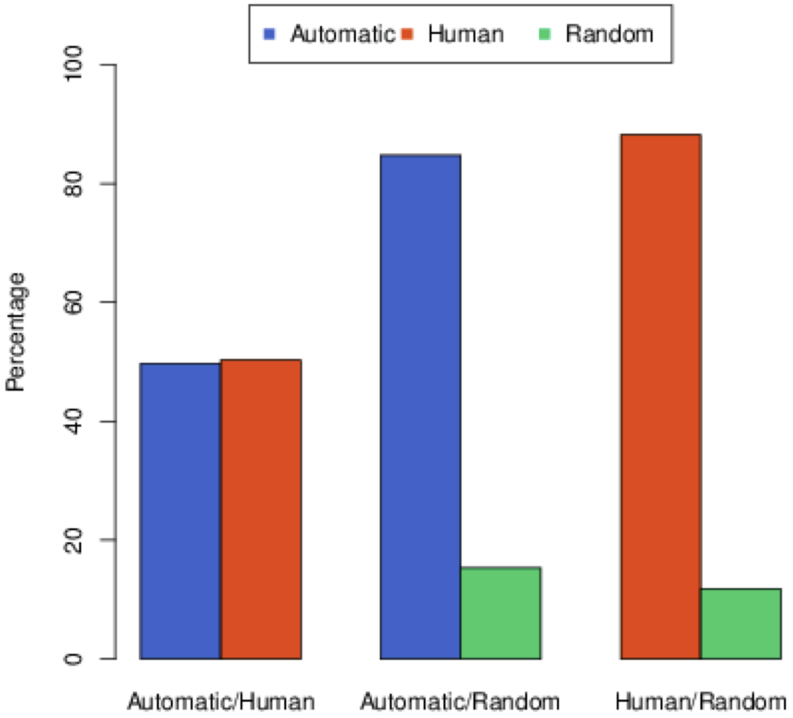


Fig. 72. The results of the user study conducted by AlNoamany et al. [9].

The task containing a poor story was an attention question. If a participant selected a poor story as their preferred story, then their response was discarded. Figure 72 displays the results for the other story types. There was some variation amongst stories from this study, but overall, the participants could not tell the difference between the stories generated by humans or the DSA framework.

AlNoamany's work is compelling. Her work provides structural and topical methods of reducing a collection of mementos to a representative sample. Unlike other visualizations, such as those by Padia, users consuming these stories require minimal training to understand them because the stories use social cards that the users have experienced elsewhere on the web. In addition, unlike stories created with live web curation platforms, AlNoamany's stories link to web archives, meaning they should never have broken links. Even better, users could not tell the difference between stories produced by the DSA or archivists.

There is more to be done. AlNoamany only evaluated collections based on events. In Chapter 4.1, we show that there are more types of collections. She tailored her solution to Storify, which is now gone. In Chapter 5.1 we show how existing platforms are not a suitable replacement for Storify. We introduce the tools MementoEmbed (Chapter 5.4) and Raintale (Chapter 6.1) to not only replace Storify but expand the possibilities of storytelling with web archives beyond Storify. AlNoamany proved that test participants could not tell the difference between stories created by a human or created via this algorithm. She did not, however, evaluate how well these stories worked as summaries, which we cover in Chapter 4.4.

3.6 CHAPTER SUMMARY

In this chapter, we reviewed the existing work related to the different processes in our storytelling model. For selecting exemplars, the existing algorithms that select exemplar sentences, images, and documents from collections inspire us. Their primitives are the concepts of filtering, clustering, and scoring. We address RQ1 in Chapter 4.1 where we explore web archive collections' structural features to understand better how to apply these primitives. We address RQ2 by evaluating different methods of applying these primitives to selecting exemplars. In Chapter 4.2 we demonstrate how to filter off-topic mementos from collections. In Chapter 4.3 we discuss a model that applies these concepts as primitives for creating more complex algorithms. We evaluate some of these complex algorithms in Chapter 4.4 and implement our model of primitives in the tool Hypercane introduced in Chapter 4.5.

Existing efforts at generating story metadata for collections rely upon manual review of each memento in the collection. Web archive collections consist of thousands of mementos, making it expensive to apply these solutions. Even if they only applied metadata to individual seeds, the scale is still in the thousands for some collections. In Chapter 4.1 we discuss structural features of collections that can help users understand the curatorial involvement with a collection, how it changed over time, the diversity of its seeds, and its lifespan. We will also cover how Hypercane can address the need to generate story metadata in Chapter 4.5.

We covered past studies that tried to evaluate which surrogates best worked for search results. These studies focused on whether or not the participant selected the relevant result. A number of these studies had inconclusive results. More importantly, they focused on search engine result relevance rather than collection understanding. RQ3 asks which surrogate is best for understanding groups of mementos. We address RQ3 in Chapter 5.2 where we perform our evaluation of which surrogate works best for providing an understanding of the underlying collection. In Chapter 5.3 we discuss how to generate surrogates, specifically social cards when metadata is scarce. This work helps us address RQ4, where we ask how to emulate human behavior while automating the creation of surrogates.

We highlighted the efforts by others to visualize web archive collections. These efforts attempted to summarize the collection as a whole, often generating story and document metadata for use in complex visualizations that allow users to explore facets of the collection. These implementations did not select exemplars as a summary. Instead, they tried to visualize all mementos at once. Their visualizations' complexities also increases the cognitive load on the user because they require that the user train with them to be able to apply them. In contrast, we are relying upon the visualization paradigm of social media storytelling that most web users are already familiar with.

Finally, we introduced AlNoamany's pioneering work on combining social media storytelling with web archives. She combined all of the storytelling processes in our model. She developed an algorithm, referred to in later sections as DSA1, that selects exemplars by filtering, clustering, and scoring the collection. She rendered these exemplars via Storify, which handled generating document metadata, visualizing the story, and distributing it. AlNoamany also evaluated existing stories to determine that popular stories had a median of 28 elements. This number serves as a good target for our stories. She also established that human participants could not tell the difference between stories generated via her method and those created by archivists from their own collections. The rest of this dissertation builds upon her work.

In the next chapter, we cover our work addressing the first steps to storytelling with web archives — selecting exemplars and generating story metadata.

CHAPTER 4

SELECTING EXEMPLARS AND GENERATING STORY METADATA

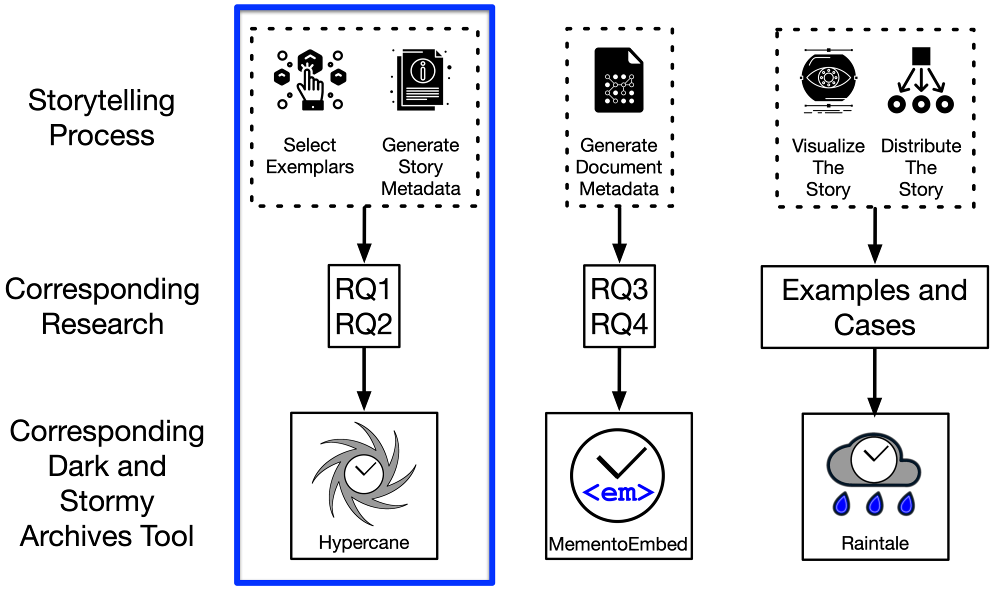


Fig. 73. The processes for storytelling that are covered in this chapter are indicated by the blue box.

In this chapter, as shown in Figure 73, we cover research questions related to selecting exemplars and generating story metadata. These are the starting points of storytelling with a corpus. We start with the semantic categories of collections that we discovered in Archive-It, indicating the intent behind creating a collection. We then discuss an Archive-It collection's structural features. These structural features help us understand the nature of collections beyond text analysis, thus indicating which features might be helpful when selecting exemplars or selecting collections for experiments. Both of these help us address RQ1. We also show how these structural features can predict the semantic category of a collection.

For RQ2, we evaluate different ways of selecting exemplars. The first step in selecting exemplars is to identify which mementos no longer match the overall collection topic. We analyze different methods of identifying off-topic mementos and establish how differences in word count perform best. Then we cover a model for how to build algorithms that select exemplars from a collection, which we then use to produce different exemplar selection algorithms. Considering that search engines are the existing method with which users might explore collections, we evaluate the exemplars chosen by each of these algorithms against a search engine to show that these algorithms select exemplars that have poor retrievability with a state-of-the-art web archive search engine and queries automatically generated by several methods. Finally, we introduce Hypercane, a tool inspired by our algorithm-building model that allows users to select exemplars.

4.1 STRUCTURAL FEATURES AND TYPES OF COLLECTIONS

Before we can select exemplars, we need to understand the nature of web archive collections. Collections contain structural features that help us understand them in ways that text analysis will not. As noted in Chapter 3.1 selecting exemplars can involve filtering, scoring, and clustering. Some structural features give us an idea of how to cluster, by what features to filter, and how to score candidate exemplars. In addition to helping us select exemplars, structural features help us compare collections. If we want to create a diverse sample of collections for an experiment, as we do in Chapter 4.4, we want to include collections that contain a variety of values for these features. Between October and December 2017, we evaluated 3,382 Archive-It collections to better understand these features and how we can use them to describe and even predict a collection's semantic category.

4.1.1 SEMANTIC CATEGORIES OF WEB ARCHIVE COLLECTIONS

We recognize that archivists create web archive collections for different reasons. A **semantic category** describes the intention behind creating a collection. In December 2017, we manually reviewed the metadata of 3,382 Archive-It collections and placed them into four semantic categories. Table 10 shows the distribution of these semantic categories across our sample. Below we describe these categories in more detail.

Self-Archiving — These collections consist of one or more domains either (1) belonging to the archiving organization or (2) being archived as part of some archiving initiative of which the collecting agency is part. Collections fitting into this category include the *University of Utah Web Archive* (ID 2278) archived by the University of Utah or the *City of Eagan Websites* (ID 2289) archived by the City of Eagan, Minnesota. In each case, the organization is archiving its own web presence. Less apparent are collections like *Governor of Tennessee, Phil Bredesen* (ID 391) archived by the Tennessee State Library and Archives. In these latter cases, the archiving organization, the name of the collection, or the ownership of the seeds do not match, but the Tennessee State Library and Archives specifically exists to archive the State of Tennessee government. Tennessee State Libraries has collections for many, if not all, Tennessee state agencies.

TABLE 10

Distribution of collections for each semantic category (as of December 2017)

Semantic Category	# of Collections	% of All Collections
Self-Archiving	1,828	54.1%
Subject-based Archiving	935	27.6%
Time Bounded - Expected	476	14.1%
Time Bounded - Spontaneous	143	4.2%
Total	3,382	100%

From this behavior, we can infer that they are tasked with archiving the web presence of all Tennessee state government agencies. Other organizations with collections that fit into this category are the Federal Depository Library Program Web Archive and the Region of Waterloo Archives. The Self-Archiving category dominated Archive-It with 54.1% of collections in 2017.

We can learn about different aspects of organizations through stories built from Self-archiving collections. With a *sliding page, sliding time* (SPST) story, we get to understand the collection and the organization it represents as a whole. Organizations often segment units into different URIs. For example, in *University of Utah Web Archives*, to learn about the university as a whole, we visit the mementos for its home page (`http://www.utah.edu/`). To visit the University Library, we visit `https://lib.utah.edu/`. With a *fixed page, sliding time* (FPST) story, we can take a specific unit of the organization and follow it through time. With a *sliding page, fixed time* (SPFT) story, we can view how the whole organization was responding to the events of the time period. We select exemplars differently depending on the aspect of the organization we want to explore.

Subject-based Archiving — Some collections consist of many seeds bound by a single topic. The topic may be evident, as with *Environmental Justice* (ID 7635), archived by Tufts University. The topic may also be vague, like *ISU Special Collections Department Manuscript Collections Web Sites* (ID 1501), archived by Iowa State University (ISU). In the former, the subject is “environmental justice.” In the latter, the subject is “the set of organizations that have shared physical items with the ISU Special Collections Department“’. Collection 1501 does not fit into the Self-Archiving category because these organizations are not part of ISU, nor is it apparent that ISU is specifically tasked via some broader archiving initiative to archive them. ISU is merely complementing its physical library collection by archiving additional information about the organizations that have contributed to it. Subject-based archiving made up 27.6% of collections in 2017.

Stories from Subject-based collections allow us to explore one or more subtopics. SPST stories provide an overview of the subject altogether, giving us enough information to further explore the topic (and the collection). A subject's subtopics could be hidden behind different URIs. For example, in *Environmental Justice*, different sources cover topics like community farming (`http://growingfoodandjustice.org/`) and sustainable water systems (`http://pacinst.org/`). The focus of these different sources allows us to use FPST stories to explore a subtopic over time. SPFT stories allow us to see the conversation around the subject at the time. If subtopics are not apparent through seeds, we may select exemplars from Subject-based collections through topic modeling, or we may cluster the collection to find subtopics and then filter for a specific subtopic to tailor the story to that aspect of the collection.

Time Bounded - Expected — These collections focus on an expected, planned event, such as *2010 Winter Olympics* (ID 5711) archived by the International Internet Preservation Consortium (IIPC). The collections may also be based on a specific time period, such as *Virginia's Political Landscape, 2007* (ID 663) archived by the Library of Virginia. Collections from institutions participating in the K-12 Archiving Initiative [191] also fit into this category, as their archivist plans to stop adding to them after a single semester or school year. These collections made up 14.1% of collections in 2017.

In Time Bounded - Expected collections, we can apply SPST stories to get an idea of the event as a whole. Time Bounded - Expected collections provide coverage of the event from different sources. In the case of *Virginia's Political Landscape*, each seed reflects a specific political candidate (e.g., `http://www.northam2007.com/`). An FPST story gives us an idea of how that candidate's message changed over time. SPFT stories give us an idea of how different candidates discussed the issues at the same time. In the case of *2010 Winter Olympics*, news sources (e.g., `http://jo-vancouver-2010.francetv.fr`) cover specific countries' interests in the games. Thus, FPST stories give us an idea of the feeling of anticipation by a country's athletes and their responses when they finally win or lose. Different SPFT stories let us see how everyone covered the games before it occurred, during, or after. We can also cluster the collections and filter on these clusters to create stories featuring a particular political opinion or sporting event.

Time Bounded - Spontaneous — These collections start after a spontaneous event. Collections fitting into this category include *Tucson Shootings* (ID 2305) archived by the Virginia Tech: Crisis, Tragedy, and Recovery Network [334, 335] and *Japan Earthquake* (ID 2438) archived by the University of Michigan, School of Information. They may also start after the beginning of a movement, such as *Black Lives Matter Movement* (ID 6396) archived by the San Jose State

University, School of Information. The key is that a curator creates these collections in response to this spontaneous event or movement, and the curator usually stops adding to them at some point. Perhaps due to the spontaneity of their creation, they are the smallest semantic category, only making up 4.2% of collections in 2017. They are also the only collection type evaluated by AlNoamany [11].

Like with Time Bounded - Expected collections, stories from Time Bounded - Spontaneous collections can give us an idea of the collection's whole topic or specific aspects. The difference is in the nature of the coverage. With FPST stories, we are trying to analyze the changing information in one source over time, not just following the events as if we were living during that time but also experiencing the misinformation and lack of information that the public experienced. For example, in *Japan Earthquake* we have 811 seeds covering the event. Some seeds (e.g., `http://11shokunin.com/keyholder/`) covered how the event affected individuals, including how the earthquake destroyed their workplaces. Other seeds focused on the damaged nuclear power plant and how activists were exploiting the contemporary public sentiment to keep it closed (e.g., `http://1120antinukes.tumblr.com/`). SPFT stories help us understand what the public and first responders knew while the event was resolving.

With these different semantic categories, we get some idea of the types of stories to tell and how to select the exemplars to tell them. These semantic categories are not the only way of looking at a collection, however.

4.1.2 SEED FEATURES

Seed features provide insight into the behavior of the curator concerning the collection. They also allow us to consider how a collection might be scored or clustered. For example, AlNoamany's Algorithm contains a scoring function that takes path depth into account. This section covers structural features related explicitly to seeds. The following section will discuss mementos.

Seed URI domain diversity — Seed URI domain diversity (also known as WSDL diversity [237] and source diversity by Nwala et al. [245]) quantifies the spread of the collection across different sources. A collection where all seeds are from the same domain would have a domain diversity of 0, and one where all seeds are from different domains would have a domain diversity of 1.

$$D = \frac{U}{C} \tag{12}$$

$$D' = \frac{CD - 1}{C - 1} = \frac{U - 1}{C - 1} \tag{13}$$

Equation 12 computes the diversity D as the number of unique domains U divided by the number of seeds C. In Equation 13, D' normalizes this diversity value D between 0 and 1. A collection with 1 seed, by definition, has a diversity of 0. A collection with a docmain diversity of 0 represents a collection where all of the content comes from a single domain, but this does not mean that such a collection only produces Fixed Page Sliding Time (FPST) stories (Chapters 1 and 3.5) because those require a single URI-R. A collection with a domain diversity of 0 only has a single source and may only present a single point of view, but it may also provide the opportunity to tell a story where that point of view changes over time.

Natasha, Elbert, and Ling want exemplars that summarize a collection. They would prefer that their exemplars have a similar domain diversity to their source collection. Natasha is considering the collection *South Louisiana Flood of 2016* which has 26 seeds. It has 10 unique domains out of 26 seeds for $D = \frac{10}{26} = 0.38$, and a normalized domain diversity score of $D' = \frac{10-1}{26-1} = \frac{9}{25} = 0.36$. Rustam wants a story that tracks a single source. His story's source collection may have a domain diversity of 1, but, because it only tracks a single source, his story should have a domain diversity of 0.

Seed URI path depth diversity — We acquire an idea of the spread of path depth across the collection by applying the above domain diversity equation to the seed path depth of every seed in the collection. This feature may indicate if the seed URIs consist solely of top-level pages or a mixture of top-level pages and more specific content [57]. Deeper pages tend to have information about more specific topics, and if the path depth diversity is high, then we have a mix of pages providing overviews and pages providing specific topics. Again, for a decent summary, the path depth diversity of the selected exemplars should be close to the path depth of the collection from their source collection. When comparing collections, path depth diversity helps us ensure that we have different collection types.

Natasha's collection has 5 unique path depths with values ranging from 0 to 4 and her collection has 26 seeds. Its path depth diversity $P = \frac{5-1}{26-1} = \frac{4}{25} = 0.16$. In Rustam's case, his story will have a path depth diversity of 0 because it only consists of one original resource.

Most frequent seed URI path depth — If a collection's most frequent seed URI path depth is 0, then it primarily consists of seeds of top-level pages. If the most frequent path depth is at a value of 2 or higher, then the collection primarily consists of seeds deeper in a website. The mode of the seed path depth is another feature that we can compare between the collection and exemplars. If this mode differs, it indicates that the exemplars only represent an aspect of the collection and not its overall structure. Because he is interested in a FPST story, representing an aspect concerning frequent path depth may work well for Rustam's case but not the others.

% Query string usage — Some collections consist primarily of URIs with query strings, whereas others consist of just paths. Path depth does not quantify the number of URIs with query strings. A collection with one seed has either 0% or 100% query string usage depending on the presence of a query string in this single seed URI. Because query string parameters can occur in any order, query strings can create issues for search engines trying to create a canonical URI for a given resource. Google created guidance in 2007 [251] instructing webmasters to create "clean URLs" with as few query string parameters as possible and, if they still need to create query strings, to register a canonical URL for each page on their website. Pages before this 2007 guidance applied query strings more liberally than those after. Certain websites (e.g., belonging to governments or banks) are built upon older technologies that heavily apply query strings. Because of these behaviors, we consider % query string usage as a feature that helps describe a web archive collection.

These features can help us evaluate the selected exemplars by comparing them with their source collections. A story with a lower domain diversity than its source collection covers fewer sources, but this may be desirable depending on the type of story we want to tell. A story with lower path depth diversity may cover more specific or general information than the collection as a whole. Comparing a collection's most frequent path depth with its exemplars gives us an idea of how the exemplars differ from their source collection, again indicating that the exemplars cover more specific or more general information. Finally, the % of query string usage gives us an idea of how well the exemplars match their source collection. Next, we cover features concerning mementos.

4.1.3 COLLECTION GROWTH CURVES

Collection growth curves help us selecting exemplars by describing the temporal spread of the collection. A collection growth curve provides insight into the seed curation and crawling behavior of an Archive-It collection, giving us an idea of how its mementos are clustered in terms of memento-datetime. Clustering by memento-datetime is a key part of different summarization algorithms, introduced in Chapters 3.1 and 3.5. Figure 74 shows an example growth curve for Archive-It Collection 366, *University of Southern California Website Archive*. The x-axis represents the **life of the collection** or the time between a collection's first memento and its last. We show the x-axis as a percentage of the collection's lifespan to normalize collections with different durations. The y-axis represents the percentage of URIs in the collection at a given time.

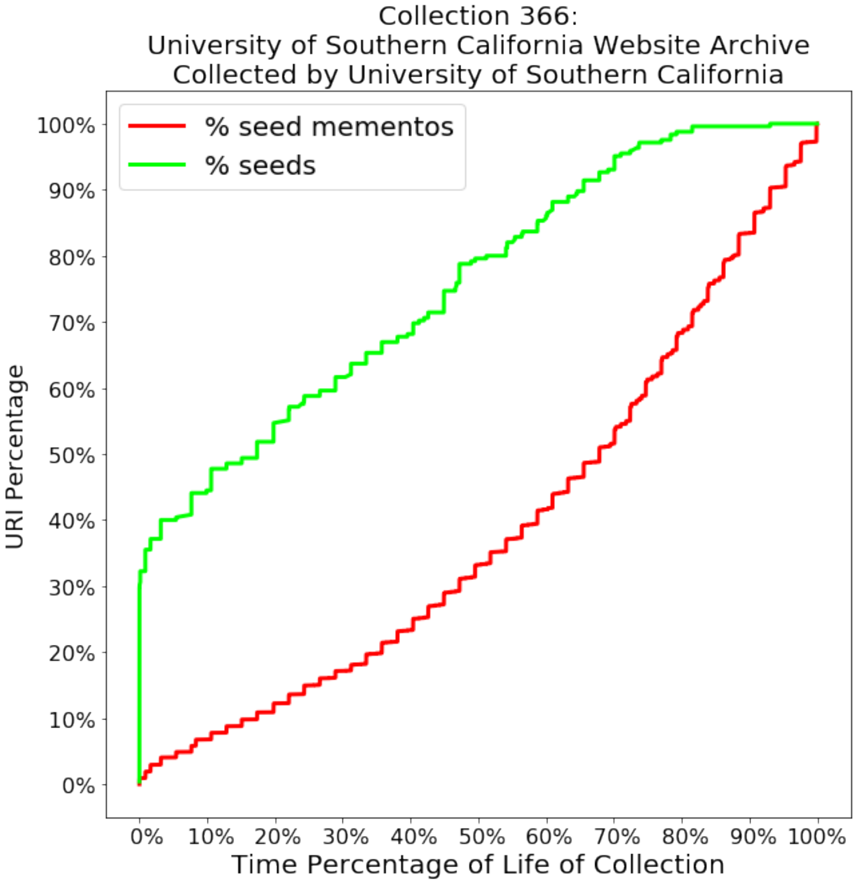


Fig. 74. The Growth Curve of Archive-It Collection 366

URIs come in two categories: seeds or seed mementos, represented by the green and red lines, respectively.

Growth curves for Archive-It collections consist of multiple parts. Figure 75 demonstrates how to interpret the information within a growth curve. An imaginary diagonal line shows a linear relationship between the growth of URIs over time. It divides each graph into two parts. If the seed line (green) is in the upper left corner, most of the seeds were added earlier to the collection, and if the seed line occupies the lower right corner, the archivist added most of them later. The seed line reflects an aspect of curatorial engagement with the collection, indicating when the curator first crawled a given seed. The closer the seed line is to the diagonal, the more often the curator added a new seed. We use the memento-datetime of the first memento for each seed to generate the seed line.

The meaning of the seed memento line is similar. Where the seed line indicates intent, the seed memento line (red) indicates the growth of the actual collection, which could just be the result of

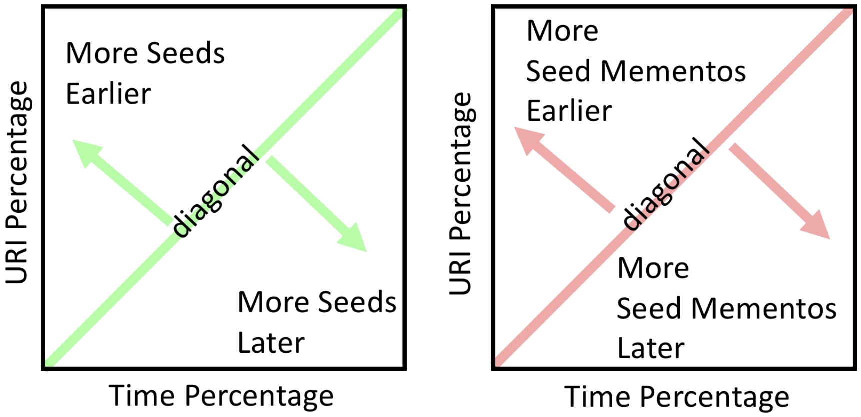


Fig. 75. The Anatomy of a Collection Growth Curve

setting up a periodic crawl. If the seed memento line mainly occupies the upper left corner, then most of the mementos were crawled earlier in the collection's life, meaning that the curator created most of the collection's holdings at that time, and thus its temporality is skewed earlier. If the seed memento line occupies the lower right corner, then the collection's temporality is skewed later. If the seed memento line runs along the diagonal, then the collection's temporality is spread more evenly across the collection. We use the memento-datetimes of all mementos to generate the seed memento line. More coverage of growth curve behaviors is available in Jones et al. [160].

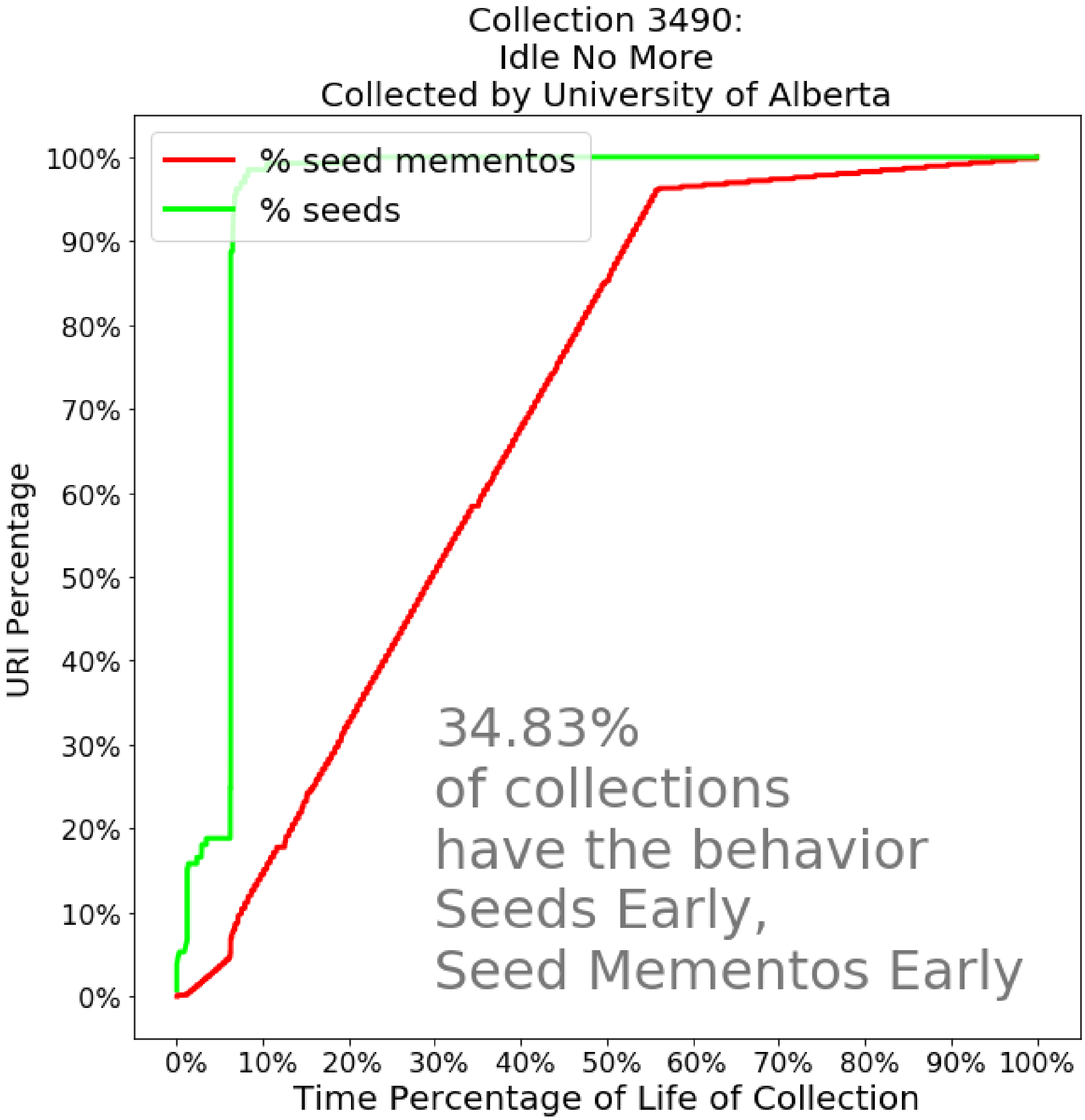
Growth Curve Behaviors

Using the area under the curve (AUC), we were able to identify different collection behaviors. If the AUC exceeds 0.5 — the area of the diagonal — then the growth was *early*. If the AUC is less than 0.5, then the growth was *late*. If the AUC is within 0.05 of the diagonal, then we considered it to be growing *continuously*.

Using these three primitives, we identified the behaviors shown in Figure 76.

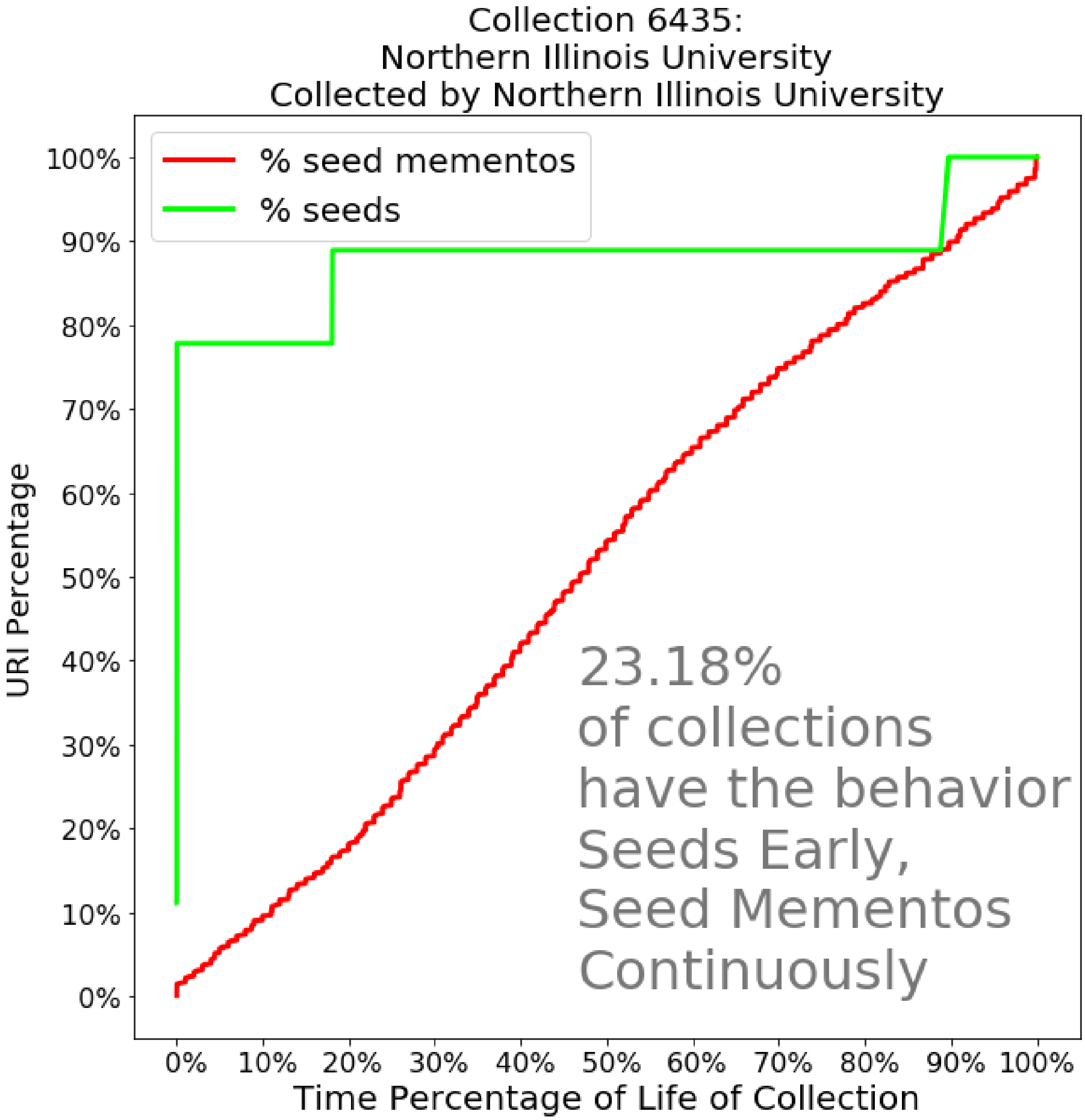
Seeds early, seed mementos early — Seen in Figure 76a with collection *Idle No More* (ID 3490), the growth curves with this behavior indicate that the curator made most curatorial decisions near the start of the life of the collection. The seed memento line skews early, indicating that most of the seed mementos in the collection come from that time period.

Seeds early, seed mementos continuously — Figure 76b shows collection *Northern Illinois University* (ID 6435), where the curator added more than 70% of the seeds near the beginning of the collection's life. The curator selected seeds early but then chose crawling strategies that added seed mementos steadily throughout its existence.



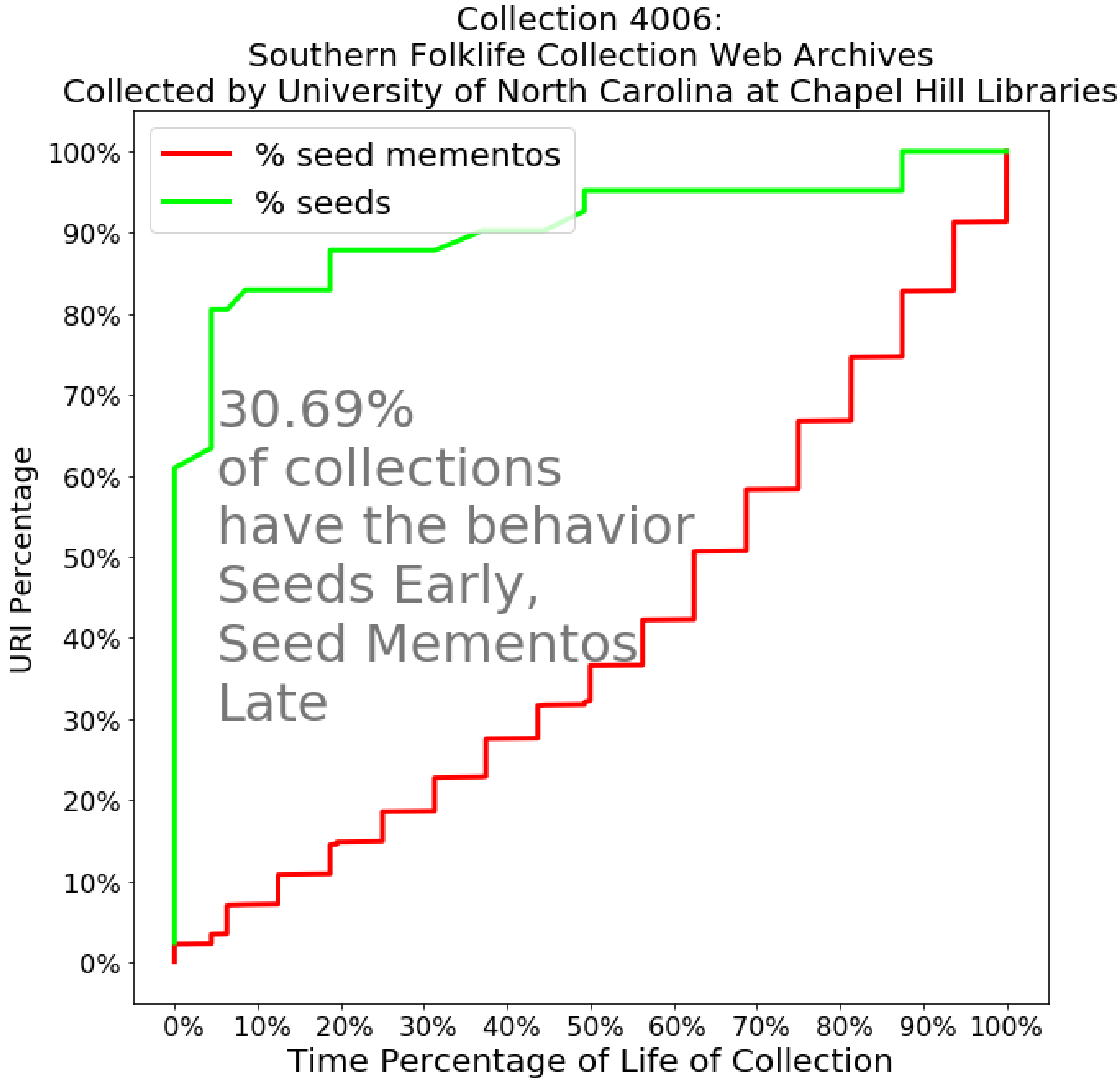
(a) Archive-It Collection 3490:
Seeds early, seed mementos early

Fig. 76. Examples from nine different growth curve behavior categories, grey inset text conveys the percentage of the 3,382 collections in this study that fit into each category.



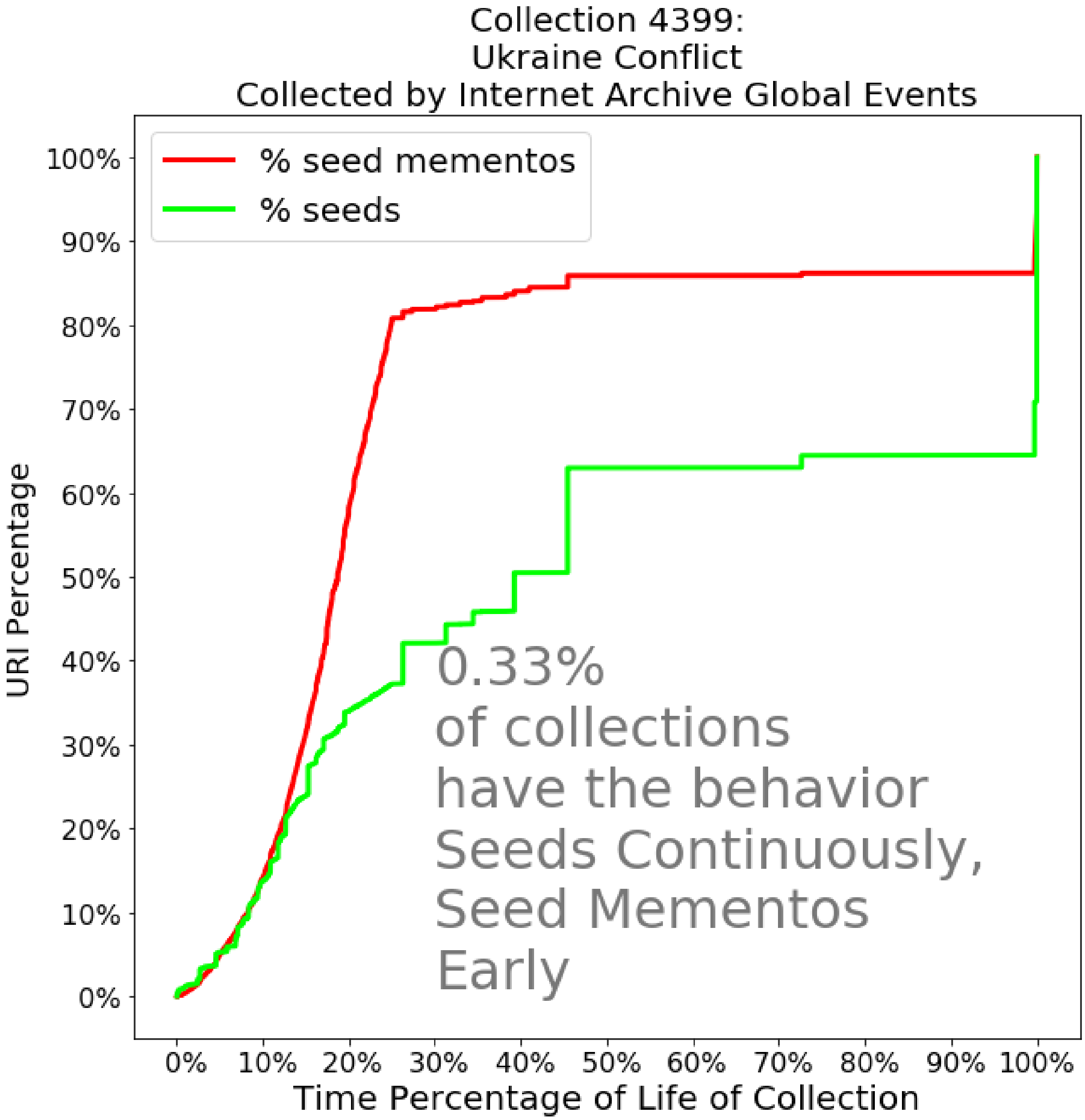
(b) Archive-It Collection 6435:
Seeds early, seed mementos continuously

Fig. 76. (Continued) Examples from nine different growth curve behavior categories, grey inset text conveys the percentage of the 3,382 collections in this study that fit into each category.



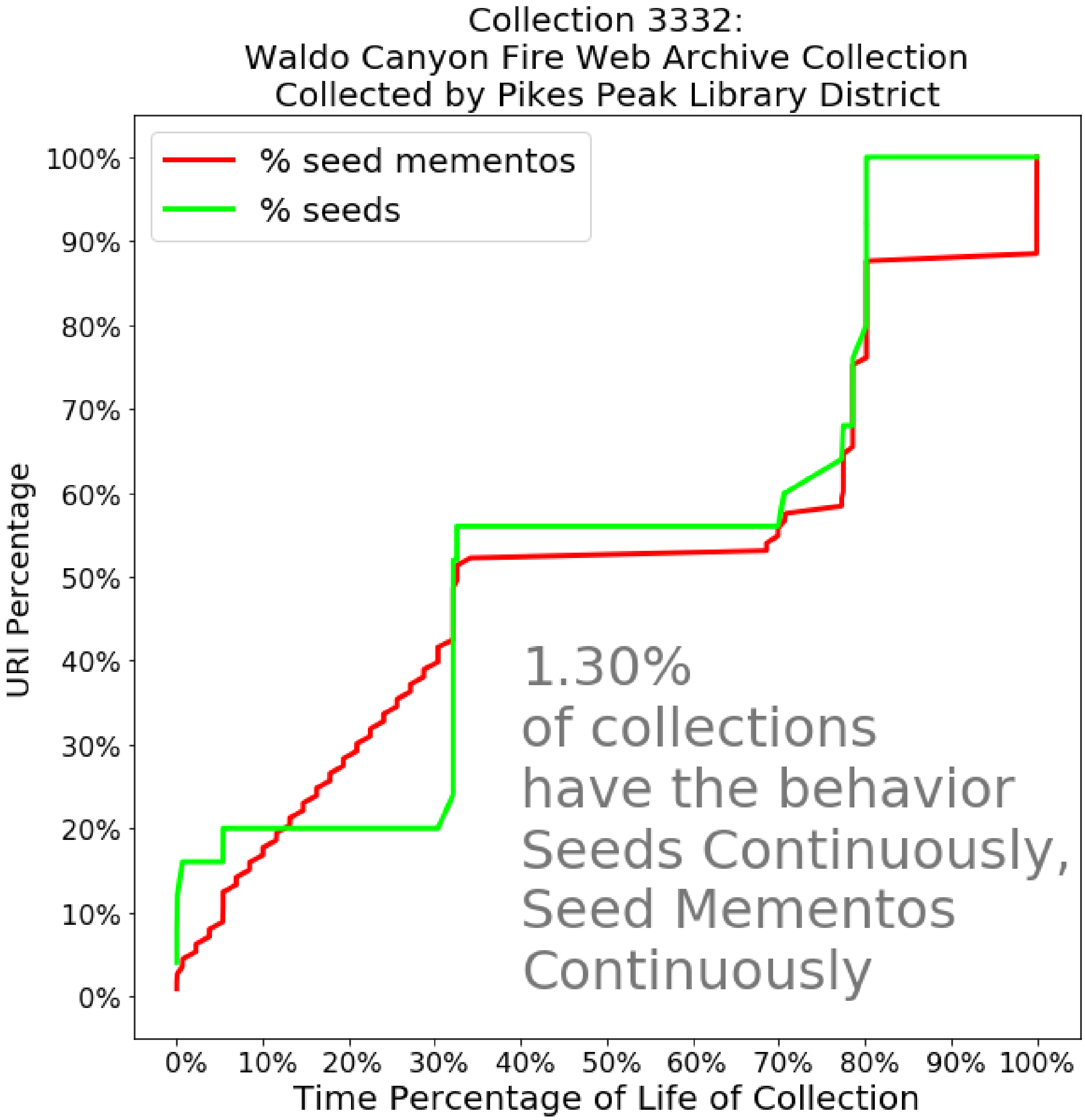
(c) Archive-It Collection 4006:
Seeds early, seed mementos late

Fig. 76. (Continued) Examples from nine different growth curve behavior categories, grey inset text conveys the percentage of the 3,382 collections in this study that fit into each category.



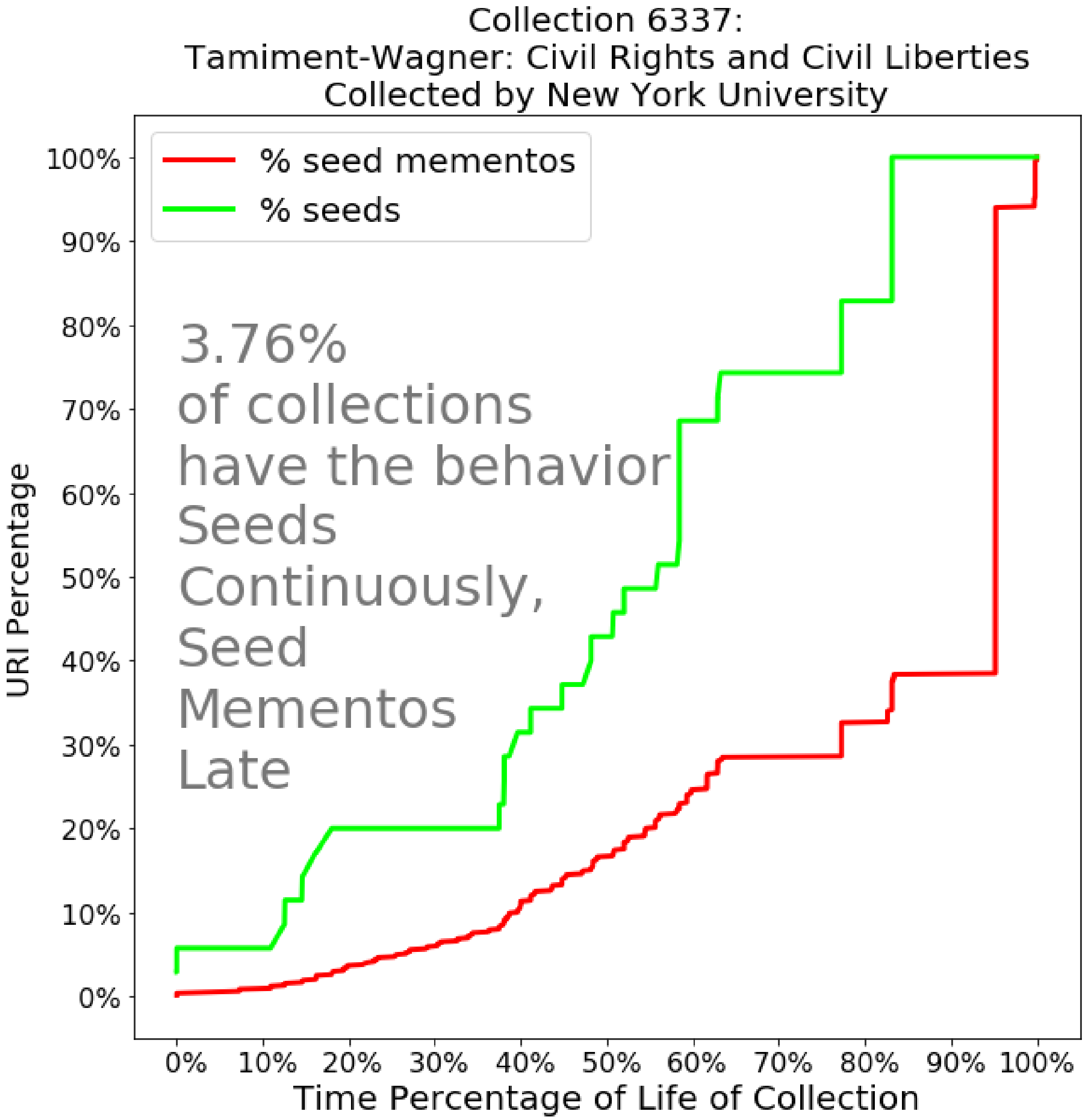
(d) Archive-It Collection 4399:
Seeds continuously, seed mementos early

Fig. 76. (Continued) Examples from nine different growth curve behavior categories, grey inset text conveys the percentage of the 3,382 collections in this study that fit into each category.



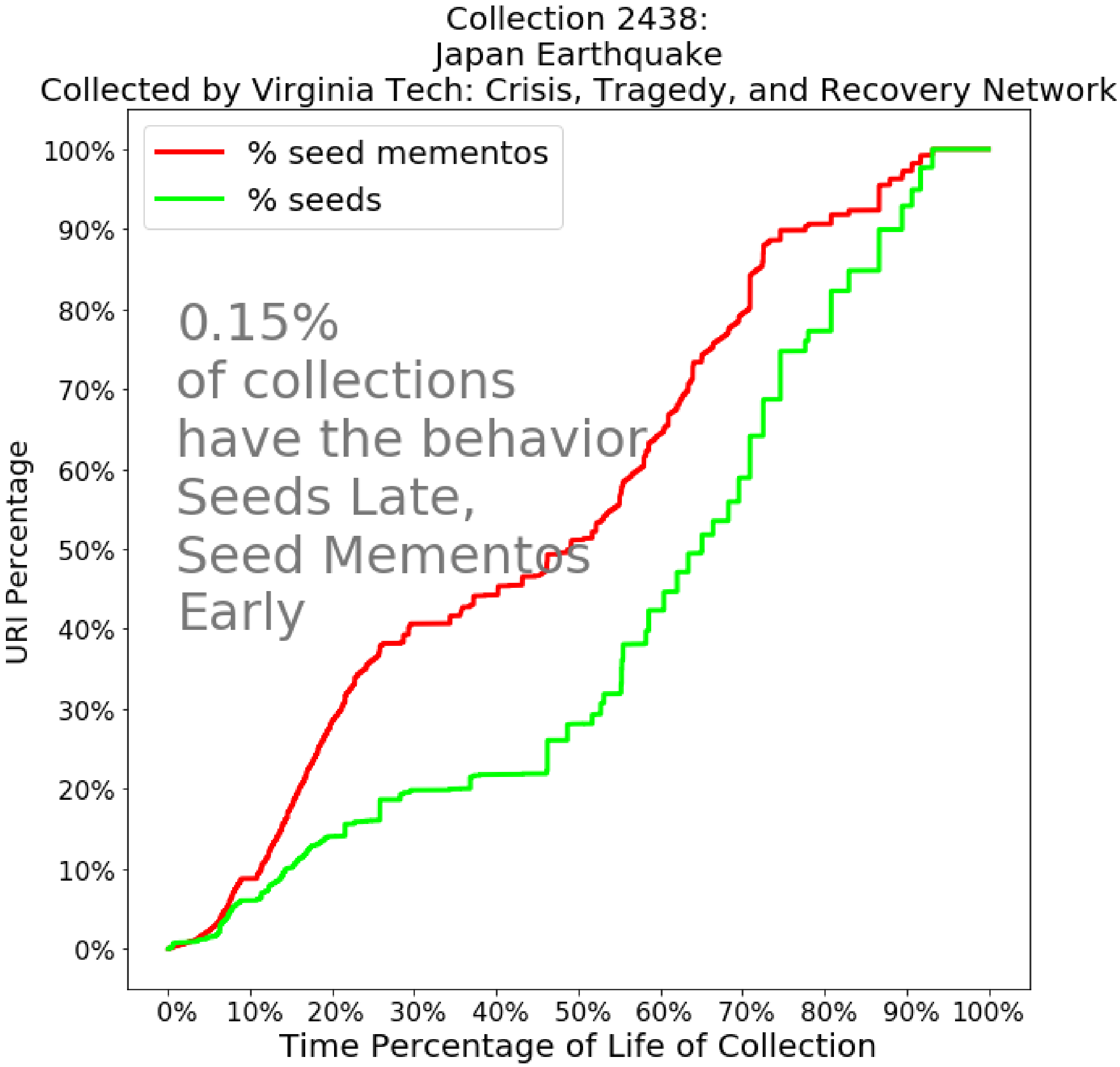
(e) Archive-It Collection 3332:
Seeds continuously, seed mementos continuously

Fig. 76. (Continued) Examples from nine different growth curve behavior categories, grey inset text conveys the percentage of the 3,382 collections in this study that fit into each category.



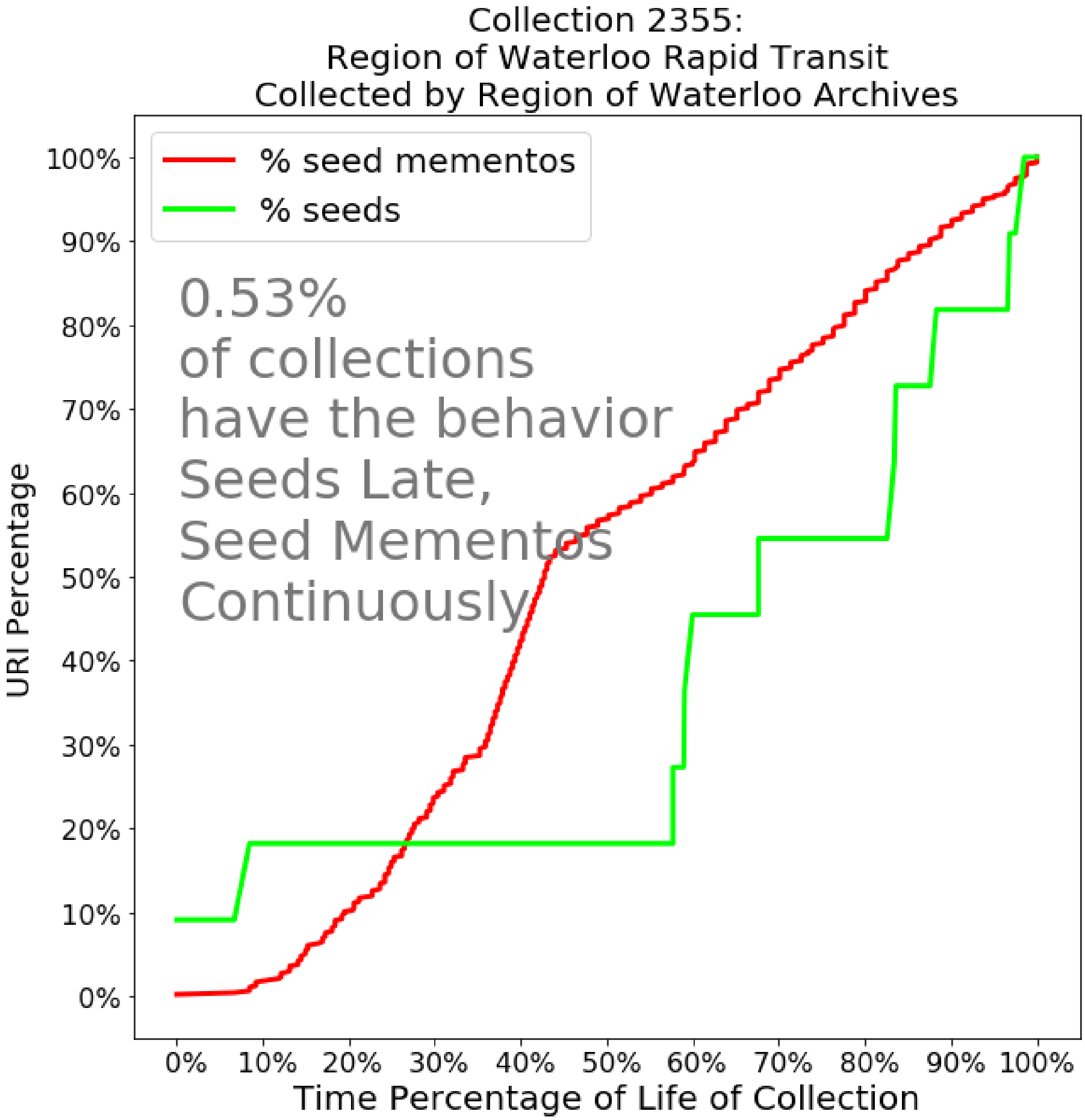
(f) Archive-It Collection 6337:
Seeds continuously, seed mementos later

Fig. 76. (Continued) Examples from nine different growth curve behavior categories, grey inset text conveys the percentage of the 3,382 collections in this study that fit into each category.



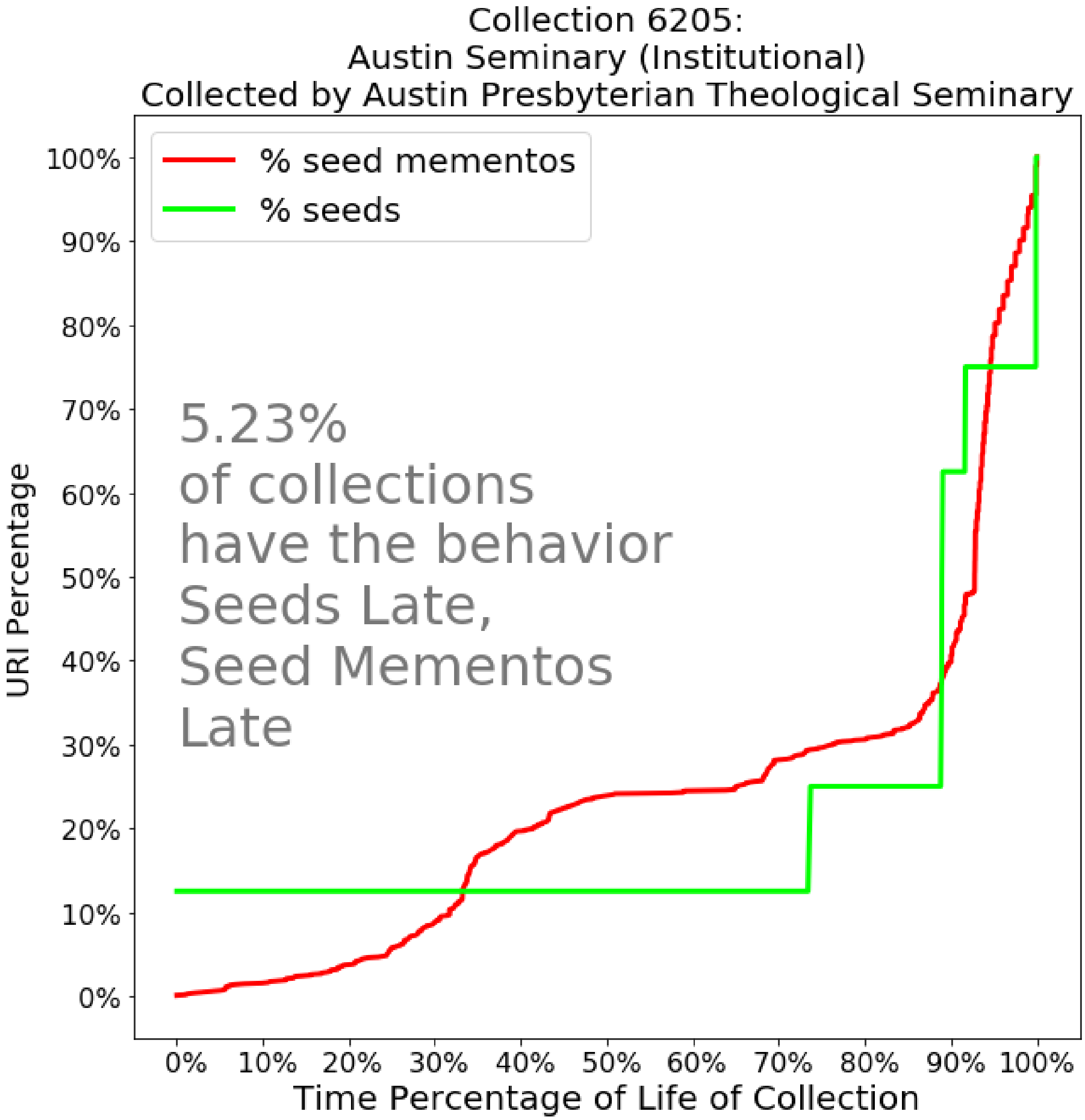
(g) Archive-It Collection 2438:
Seeds Late, seed mementos early

Fig. 76. (Continued) Examples from nine different growth curve behavior categories, grey inset text conveys the percentage of the 3,382 collections in this study that fit into each category.



(h) Archive-It Collection 2355:
Seeds late, seed mementos continuously

Fig. 76. (Continued) Examples from nine different growth curve behavior categories, grey inset text conveys the percentage of the 3,382 collections in this study that fit into each category.



(i) Archive-It Collection 6205:
Seeds late, seed mementos late

Fig. 76. (Continued) Examples from nine different growth curve behavior categories, grey inset text conveys the percentage of the 3,382 collections in this study that fit into each category.

Seeds early, seed mementos late — In all of these cases, the curator chose seeds early, but the crawling strategy produced seed mementos at a later time. In collection *Southern Folklife Collection Web Archives* (ID 4006), shown in Figure 76c, we see a case where the curator added 60% of the seeds earlier in the collection's life. In this case, the curator crawled 50% of all mementos by 65% of the collection's life.

Seeds early is the most frequent seed behavior, taking place in 88.7% of all collections studied.

Seeds continuously, seed mementos early — Figure 76d shows collection *Ukraine Conflict* (ID 4399), where the seed growth curve wraps around the diagonal, indicating that the curator added seeds more regularly, but most crawling happened earlier in the collection's life. In this case, there are more seed mementos from the earlier seeds.

Seeds continuously, seed mementos continuously — Collection *Waldo Canyon Fire Web Archive Collection* (ID 3332) is shown in Figure 76e. Collections with this behavior indicate continuous involvement both on the part of seed selection as well as crawling. Both lines wrap the diagonal as the collection grows steadily.

Seeds continuously, seed mementos late — Shown in Figure 76f with collection *Tamiment-Wagner: Civil Rights and Civil Liberties* (ID 6337), this behavior indicates that the curator was continuously engaged in adding seeds to the collection. However, most of the mementos were created later in the collection's life.

Seeds late, seed mementos early — This behavior is demonstrated by collection *Japan Earthquake* (ID 2438) in Figure 76g. In this case, the seed memento growth line exists farther left on the graph than the seed growth line. The early seeds added to the collection have more memento growth than the seeds that follow because the curator added more seeds later in the collection's life.

Seeds late, seed mementos continuously — Figure 76h shows collection *Region of Waterloo Rapid Transit* (ID 2355). In this case, the seed memento growth is steady, but something changed around 60% of its life span. Approximately 60% of the seed mementos belong to the first 20% of seeds.

Seeds late, seed mementos late — This behavior is exemplified by collection *Austin Seminary (Institutional)* (ID 6205), shown in Figure 76i. In this case, the curator added more seeds when the collection was already 70% old. Another dramatic shift happened when more seeds were added at the 90% mark. This pattern could indicate dramatically renewed interest in this collection.

Understanding these behaviors has implications for the type of temporal clustering needed for selecting exemplars. AlNoamany's Algorithm (Chapter 3.5) contains a step where mementos are sequentially placed into equally-sized *slices* by memento-datetime. This method works well when an archivist adds seed mementos continuously because every memento is equally temporally distributed. If a collection's growth is early or late, then mementos that are temporally close can end up in different clusters, not truly representing the collection in its resulting story. The DSA2, DSA3, and DSA4 algorithms, which we will cover in Chapter 4.3, apply k-means clustering to account for these different growth curve shapes.

Growth Curve Features

We have identified five growth curve features that provide insight into the behavior of a collection.

The **number of seeds** submitted to the collection varies, as does the **number of seed mementos**. We can count these by using the seed acquisition activities and TimeMaps. As the ratio of seeds equals the number of seed mementos, the growth curve for the collection is likely seeds continuously, seed mementos continuously. If we compare the number of seed mementos in the collection to the number of exemplars, we get an idea of the level of compression we have performed on the collection.

Difference between seed curve AUC and diagonal — The AUC of the seed curve indicates whether the seeds were added earlier or later to the collection. Subtracting this value from the AUC of the diagonal gives additional information helpful in understanding the nature of the seed curve. Negative values indicate that the curator first added seeds later. Positive values indicate that the curator added seeds earlier. Values very close to 0 indicate that the curator added seeds continuously.

Difference between seed memento curve AUC and diagonal – The area under the seed memento curve is helpful as well. Subtracting this value from the AUC of the diagonal provides similar information to the seed AUC feature mentioned above. This difference gives us a measure of the balanced size of temporal clusters in the collection. If the value is close to 0, then AlNoamany's time slices represent the temporal skew of the collection, but if the value is much higher or lower, then another clustering method would better represent the collection's temporal skew.

Difference between seed curve AUC and seed memento curve AUC — Subtracting the AUC of both curves indicates how close they are to each other. A value of 0 indicates that the curves are overlapping, likely meaning that there is one memento per seed. A positive value means that the curator added seeds earlier than the seed mementos. A negative value means that the seed memento growth has overtaken the seed growth.

Collection Lifespan — The collection lifespan is the difference in memento-datetime between the last memento and the first. For selecting exemplars, our best output provides representation across this entire temporal range. For SPST stories, the exemplars from a good summary should have a similar collection lifespan to that of the collection.

4.1.4 PREDICTING SEMANTIC CATEGORIES USING STRUCTURAL FEATURES

Each structural feature provides an aspect for a collection. We wanted to know how well the structural features mapped to these semantic categories. Our goal was to be able to predict some aspect of the descriptive information from the structural features introduced in the last section.

Based on testing 20 classifiers with Weka v. 3.8.2 [98] using 10-fold cross validation, we determined that Random Forest [41] is the best classifier for predicting a collection's semantic category (Table 11) based on the structural features we have covered.

We have four semantic categories, so these weighted average results do not provide a complete picture. Table 12 shows the results for Random Forest by semantic category. Random Forest correctly categorizes Self-archiving the best, with a weighted average $F_1 = 0.847$. This result is likely because 54.1% of Archive-It collections fall into this category. Other semantic categories do not fare so well. Random Forest performs worst with Time Bounded - Spontaneous, with $F_1 = 0.456$. This result is likely because it only makes up 4.2% of all collections, giving the classifier little with which to train. More surprising is that Random Forest does so well with Time Bounded - Expected at $F_1 = 0.621$, even though that semantic category only makes up 14.1% of all collections. Finally, Random Forest performs slightly worse with Subject-based Archiving with $F_1 = 0.562$ despite making up 27.6% of all collections.

4.1.5 SUMMARY

This section introduced four semantic categories of web archive collections: Self-Archiving, Subject-based Archiving, Time Bounded - Expected, and Time Bounded - Spontaneous. We discovered that Self-Archiving is the prevailing semantic category, making up 54.1% of the collections surveyed. Time Bounded - Spontaneous is the smallest category, only making up 4.2% of collections. We covered how we could select exemplars to tell different stories depending on the semantic category of the collection.

We introduced structural features for understanding an Archive-It collection, addressing RQ1. We have adapted collection growth curves to Archive-It collections, revealing their behaviors. Through these curves, we gain an understanding of the skew of the temporality of a collection. This skew can aid us in selecting exemplars because the temporal clustering we choose could depend on the collection's skew. We have also identified seed features that help us understand the curation strategy of a collection. Via domain diversity, we can tell if a collection consists of seeds from one domain or many, thus understanding that the collection comes from many different

TABLE 11

Weighted average results of 10 best classifiers for predicting semantic class evaluated using 10-fold cross validation runs while training on the complete data set

	Weighted Average			
Classifier	**TPR**	**FPR**	F_1	**ROC Area**
Random Forest	0.728	0.182	0.720	0.871
ForestPA	0.713	0.201	0.701	0.854
Decision Table	0.702	0.214	0.685	0.831
LMT	0.702	0.205	0.685	0.833
CDT	0.700	0.212	0.686	0.819
JRip	0.698	0.235	0.679	0.769
Simple Cart	0.694	0.199	0.683	0.811
FT	0.693	0.201	0.681	0.789
BFTree	0.689	0.214	0.676	0.766
Multilayer Perceptron	0.686	0.197	0.675	0.818

TABLE 12

Results by class of Random Forest classifier for predicting semantic category evaluated using 10-fold cross validation runs

	Weighted Average			
Semantic Category	**TPR**	**FPR**	F_1	**ROC Area**
Self-Archiving	0.891	0.250	0.847	0.899
Subject-based Archiving	0.538	0.144	0.562	0.794
Time Bounded - Expected	0.588	0.050	0.621	0.911
Time Bounded - Spontaneous	0.364	0.010	0.456	0.879
Weighted Average	0.728	0.182	0.720	0.871

sources. Using the most frequent URI path depth, we determine if most of the collection consists of top-level pages or specific deep URIs. Seeds with higher path depths tend to cover more specific topics, while those with lower path depths cover less specific ones. We understand the spread of path depths within a collection with seed path depth diversity, indicating if most of its seeds come from the same level of their websites. Understanding how much of a collection uses a query string in its URIs also provides information on the nature of its seeds. Through these features, we gain an understanding of the nature of what curators chose for archiving.

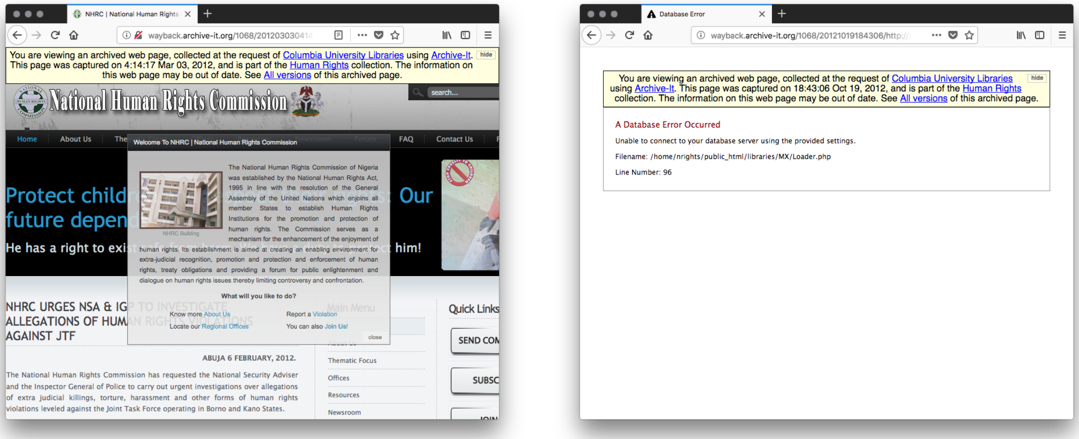
We also provided the results of training runs with classifiers and determined that the Random Forest classifier performs best at identifying the semantic category, with a weighted average F_1 score of 0.720.

In Chapter 4.3, we introduce a model for selecting exemplars from collections and detail several algorithms that take into account the features discussed in this section. In Chapter 4.4 we will revisit the structural features and growth curve behaviors as we discuss the collections we have selected for evaluating these algorithms against search engine results. In the next section, we tackle an important first step toward selecting exemplars from web archive collections: identifying off-topic mementos.

4.2 IDENTIFYING OFF-TOPIC MEMENTOS

The exemplars we select must represent one or more aspects of their collection, or they will not produce a good summary. Curators create collections for some purpose, and they choose seeds to support the collection's topic. In order to understand the history of an event or an organization, curators will often configure the platform to capture the same original resource multiple times, thus producing many mementos per seed. Many collections at Archive-It are like collection 2438, *Japan Earthquake*, with 81,014 seeds resulting in 486,227 seed mementos. For selecting exemplars, we want to optimize the amount of time spent evaluating mementos, and the sheer quantity of mementos to evaluate makes it imperative that we not spend time on mementos with low information value.

While a crawler automatically builds a web archive collection, the content of each original resource can drift from the content of its original capture. Occasionally this content drifts so far that the original resource is no longer on topic, but the crawler is still automatically capturing mementos of that page. Pages can go off-topic for a variety of reasons. Technical issues can befall a website, as shown in Figure 77. Sites may be taken down for maintenance, as seen in Figure 78. Hosting services may suspend a website, often due to nonpayment (Figure 79). Another company can purchase domains and replace the website (Figure 80). Hackers can deface websites (Figure


NHRC | National Human Rights
wayback.archive-it.org/1068/2012030304l4
National Human Rights Commission
search...
Home
About Us
FAQ
Contact Us
Quick Links
SEND COM
SUBSC
Database Error
wayback.archive-it.org/1068/20121019184306/http://

You are viewing an archived web page, collected at the request of Columbia University Libraries using Archive-It. This page was captured on 4:14:17 Mar 03, 2012, and is part of the Human Rights collection. The information on this web page may be out of date. See All versions of this archived page.

Welcome To NHRC | National Human Rights Commission

NHRC Building

The National Human Rights Commission of Nigeria was established by the National Human Rights Act, 1995 in line with the resolution of the General Assembly of the United Nations which enjoins all member States to establish Human Rights Institutions for the promotion and protection of human rights. The Commission serves as a mechanism for the enhancement of the enjoyment of human rights. Its establishment is aimed at creating an enabling environment for extra-judicial recognition, promotion and protection and enforcement of human rights, treaty obligations and providing a forum for public enlightenment and dialogue on human rights issues thereby limiting controversy and confrontation.

What will you like to do?

Know more About Us — Report a Violation
Locate our Regional Offices — You can also Join Us!

close

Protect child … Our future depen…

He has a right to exi… him!

NHRC URGES NSA & IG… ALLEGATIONS OF HUM… AGAINST JTF

ABUJA 6 FEBRUARY, 2012.

The National Human Rights Commission has requested the National Security Adviser and the Inspector General of Police to carry out urgent investigations over allegations of extra judicial killings, torture, harassment and other forms of human rights violations leveled against the Joint Task Force operating in Borno and Kano States.

Thematic Focus
Offices
Resources
Newsroom

You are viewing an archived web page, collected at the request of Columbia University Libraries using Archive-It. This page was captured on 18:43:06 Oct 19, 2012, and is part of the Human Rights collection. The information on this web page may be out of date. See All versions of this archived page.

A Database Error Occurred

Unable to connect to your database server using the provided settings.

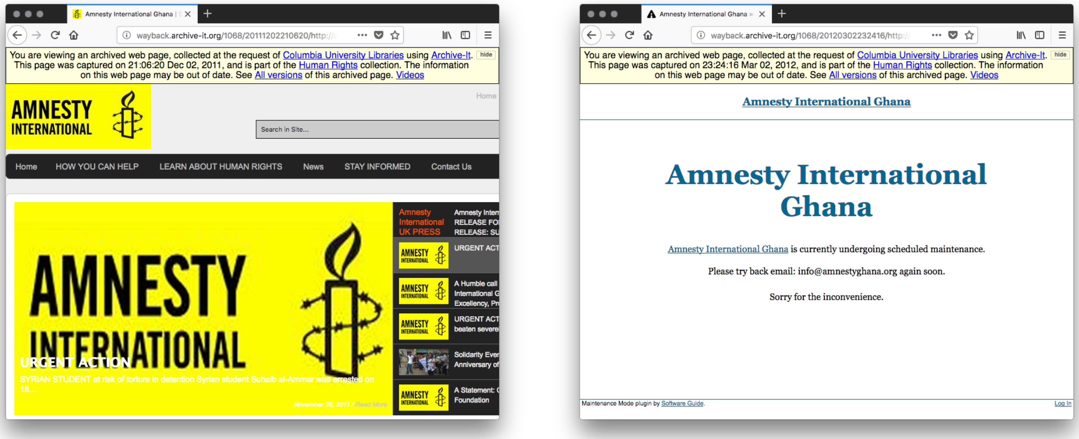
Filename: /home/nrights/public_html/libraries/MX/Loader.php

Line Number: 96

(a) On-topic
on 2012-03-10

(b) Off-topic due to technical issues
on 2012-10-19

Fig. 77. `http://www.nigeriarights.gov.ng/` preserved in Archive-It collection 1068: *Human Rights*.


Amnesty International Ghana |
wayback.archive-it.org/1068/20111202210620/http://
AMNESTY INTERNATIONAL
Home
Search in Site...
Home
HOW YOU CAN HELP
LEARN ABOUT HUMAN RIGHTS
News
STAY INFORMED
Contact Us
AMNESTY INTERNATIONAL
Amnesty International Ghana »
wayback.archive-it.org/1068/20120302232416/http://

You are viewing an archived web page, collected at the request of Columbia University Libraries using Archive-It. This page was captured on 21:06:20 Dec 02, 2011, and is part of the Human Rights collection. The information on this web page may be out of date. See All versions of this archived page. Videos

URGENT ACTION

SYRIAN STUDENT at risk of torture in detention Syrian student Suhaib al-Ammar was arrested on 18...

November 28, 2011 / Read More

Amnesty International UK PRESS — Amnesty Inter… RELEASE FO… RELEASE: SU…

URGENT ACT…

A Humble call … International G… Excellency, Pr…

URGENT ACT… beaten severe…

Solidarity Eve… Anniversary of…

A Statement: … Foundation

You are viewing an archived web page, collected at the request of Columbia University Libraries using Archive-It. This page was captured on 23:24:16 Mar 02, 2012, and is part of the Human Rights collection. The information on this web page may be out of date. See All versions of this archived page. Videos

Amnesty International Ghana

Amnesty International Ghana

Amnesty International Ghana is currently undergoing scheduled maintenance.

Please try back email: info@amnestyghana.org again soon.

Sorry for the inconvenience.

Maintenance Mode plugin by Software Guide. — Log In

(a) On-topic
on 2011-12-02

(b) Off-topic due to site maintenance
on 2012-03-02

Fig. 78. `http://amnestyghana.org/` preserved in Archive-It collection 1068: *Human Rights*.

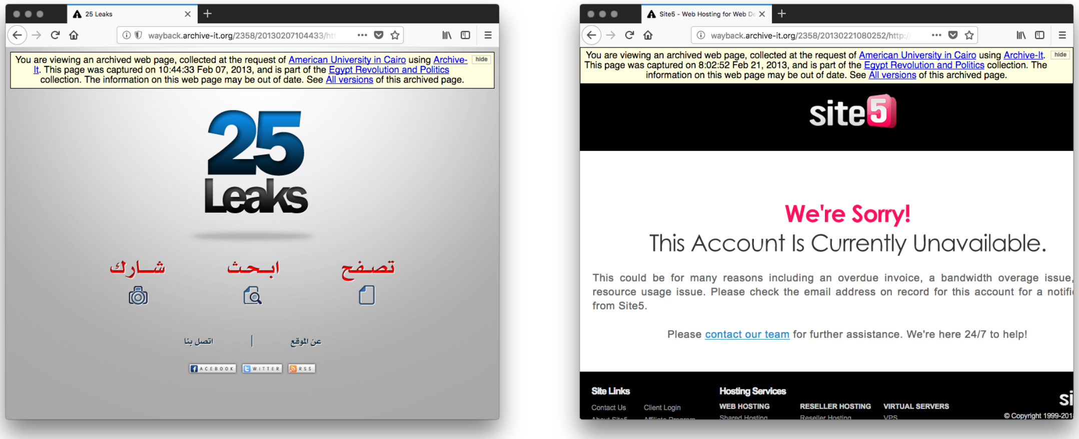
Algorithm 1 General algorithm used to determine which mementos in TimeMap are off-topic.

```
for $timemap \in collection$ do
    $f \leftarrow HTTP_{GET_{raw}}(memento_{first})$
    $f \leftarrow preprocess(f)$
    for $memento \in timemap$ do
        $m \leftarrow HTTP_{GET_{raw}}(memento)$
        $m \leftarrow preprocess(m)$
        $score \leftarrow computeSimilarity(f, m, measure)$
        $saveScore(score, memento, measure)$
    end for
end for
```

81). A web page can go off-topic due to website redesigns. Finally, as is often the case with news websites, resources can go off-topic because the given seed no longer publishes on-topic content. To perform research on the content of a web archive collection, it becomes imperative to exclude these **off-topic mementos** from downstream analysis.

We start by assuming that an original resource's first memento is on topic because it was captured after the curator judged and selected it [10]. We assume that subsequent mementos for that same resource are off-topic if their content drifts too far from the original. To make this work, we turn to TimeMaps to provide us with a list of mementos and their datetimes. Our work differs from Patel et al. [259] in that our method for identifying off-topic mementos is intra-TimeMap whereas Patel's is inter-TimeMap. We will discuss Patel's work further in Chapter 7.

Algorithm 1 displays the general algorithm used to determine which mementos in a TimeMap are off-topic. This algorithm iterates through all TimeMaps in the collection. For each TimeMap, it dereferences the raw version of the **first memento** (denoted by f) in for its content. After preprocessing (if necessary), the algorithm iterates through the preprocessed version of every other memento in the TimeMap, comparing the first memento to each additional **considered memento** (denoted by m) with the selected similarity measure (denoted by *measure*). We evaluated different textual similarity measures for off-topic detection. Table 13 lists these similarity measures, their score ranges, and whether we apply preprocessing, such as boilerplate removal, stemming, and tokenization, before leveraging them. Each similarity measure requires a threshold score for determining if a page is off-topic. In the next section we detail how we arrived at reasonable default values for each measure.



You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 10:44:33 Feb 07, 2013, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page.

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 8:02:52 Feb 21, 2013, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page.

We're Sorry!

This Account Is Currently Unavailable.

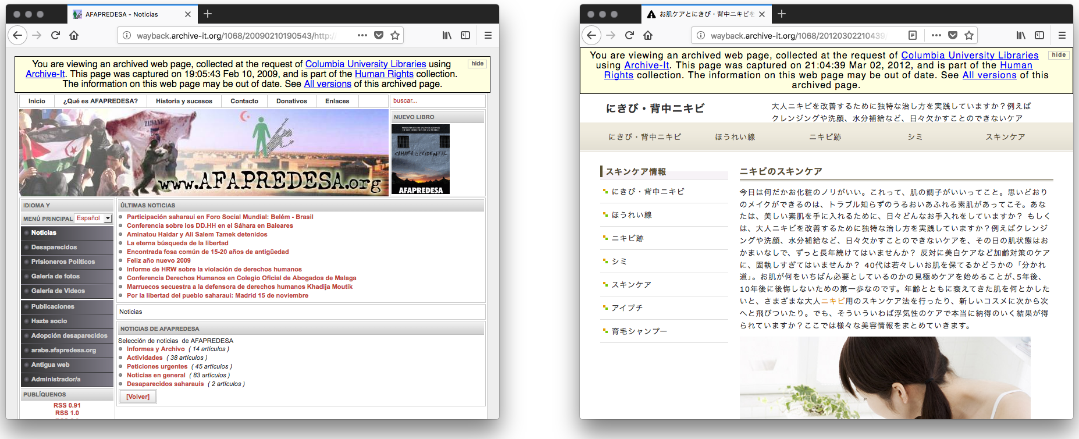
This could be for many reasons including an overdue invoice, a bandwidth overage issue, resource usage issue. Please check the email address on record for this account for a notifi from Site5.

Please contact our team for further assistance. We're here 24/7 to help!

(a) On-topic
on 2013-02-07

(b) Off-topic due to account suspension
on 2013-02-21

Fig. 79. `http://25leaks.com` preserved in Archive-It collection 2358: *Egypt Revolution and Politics*.



You are viewing an archived web page, collected at the request of Columbia University Libraries using Archive-It. This page was captured on 19:05:43 Feb 10, 2009, and is part of the Human Rights collection. The information on this web page may be out of date. See All versions of this archived page.

ÚLTIMAS NOTICIAS

- Participación saharaui en Foro Social Mundial: Belém - Brasil
- Conferencia sobre los DD.HH en el Sáhara en Baleares
- Aminatou Haidar y Ali Salem Tamek detenidos
- La eterna búsqueda de la libertad
- Encontrada fosa común de 15-20 años de antigüedad
- Feliz año nuevo 2009
- Informe de HRW sobre la violación de derechos humanos
- Conferencia Derechos Humanos en Colegio Oficial de Abogados de Malaga
- Marruecos secuestra a la defensora de derechos humanos Khadija Moutik
- Por la libertad del pueblo saharaui: Madrid 15 de noviembre

Noticias

NOTICIAS DE AFAPREDESA

Selección de noticias de AFAPREDESA

- Informes y Archivo (14 artículos)
- Actividades (38 artículos)
- Peticiones urgentes (45 artículos)
- Noticias en general (83 artículos)
- Desaparecidos saharauis (2 artículos)

You are viewing an archived web page, collected at the request of Columbia University Libraries using Archive-It. This page was captured on 21:04:39 Mar 02, 2012, and is part of the Human Rights collection. The information on this web page may be out of date. See All versions of this archived page.

にきび・背中ニキビ

大人ニキビを改善するために独特な治し方を実践していますか？例えば
クレンジングや洗顔、水分補給など、日々欠かすことのできないケア

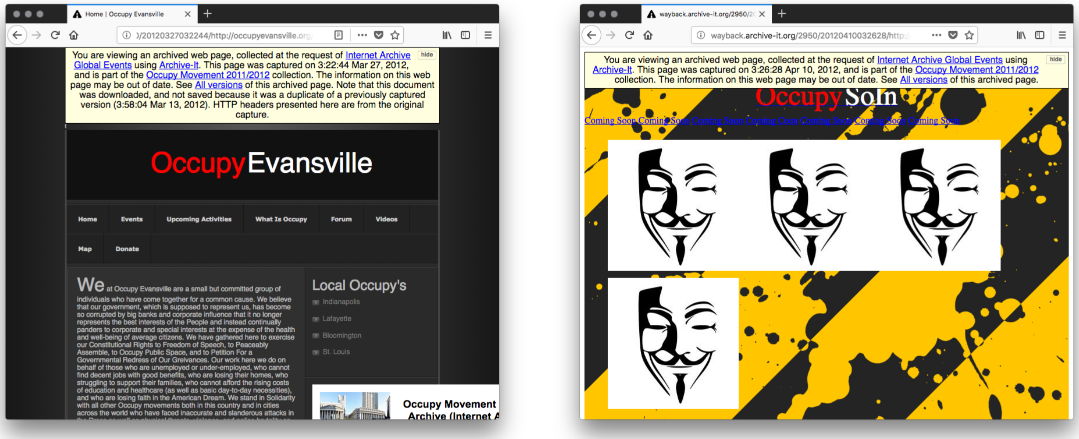
ニキビのスキンケア

今日は何だかお化粧のノリがいい。これって、肌の調子がいいってこと。思いどおりのメイクができるのは、トラブル知らずのうるおいあふれる素肌があってこそ。あなたは、美しい素肌を手に入れるために、日々どんなお手入れをしていますか？ もしくは、大人ニキビを改善するために独特な治し方を実践していますか？例えばクレンジングや洗顔、水分補給など、日々欠かすことのできないケアを、その日の肌状態はおかまいなしで、ずっと長年続けてはいませんか？ 反対に美白ケアなど加齢対策のケアに、固執しすぎてはいませんか？ 40代は若々しいお肌を保てるかどうかの「分かれ道」。お肌が何をいちばん必要としているのかの見極めケアを始めることが、5年後、10年後に後悔しないための第一歩なのです。年齢とともに衰えてきた肌を何とかしたいと、さまざまな大人ニキビ用のスキンケア法を行ったり、新しいコスメに次から次へと飛びついたり。でも、そういういわば浮気性のケアで本当に納得のいく結果が得られていますか？ここでは様々な美容情報をまとめていきます。

(a) On-topic
on 2009-02-10

(b) Off-topic due to change in ownership
on 2012-03-02

Fig. 80. `http://www.afapredesa.org` preserved in Archive-It collection 1068: *Human Rights*.



You are viewing an archived web page, collected at the request of Internet Archive Global Events using Archive-It. This page was captured on 3:22:44 Mar 27, 2012, and is part of the Occupy Movement 2011/2012 collection. The information on this web page may be out of date. See All versions of this archived page. Note that this document was downloaded, and not saved because it was a duplicate of a previously captured version (3:58:04 Mar 13, 2012). HTTP headers presented here are from the original capture.

We at Occupy Evansville are a small but committed group of individuals who have come together for a common cause. We believe that our government, which is supposed to represent us, has become so corrupted by big banks and corporate influence that it no longer represents the best interests of the People and instead continually panders to corporate and special interests at the expense of the health and well-being of average citizens. We have gathered here to exercise our Constitutional Rights to Freedom of Speech, to Peaceably Assemble, to Occupy Public Space, and to Petition For a Governmental Redress of Our Greivances. Our work here we do on behalf of those who are unemployed or under-employed, who cannot find decent jobs with good benefits, who are losing their homes, who struggling to support their families, who cannot afford the rising costs of education and healthcare (as well as basic day-to-day necessities), and who are losing faith in the American Dream. We stand in Solidarity with all other Occupy movements both in this country and in cities across the world who have faced inaccurate and slanderous attacks in

You are viewing an archived web page, collected at the request of Internet Archive Global Events using Archive-It. This page was captured on 3:26:28 Apr 10, 2012, and is part of the Occupy Movement 2011/2012 collection. The information on this web page may be out of date. See All versions of this archived page.

(a) On-topic on 2012-03-27

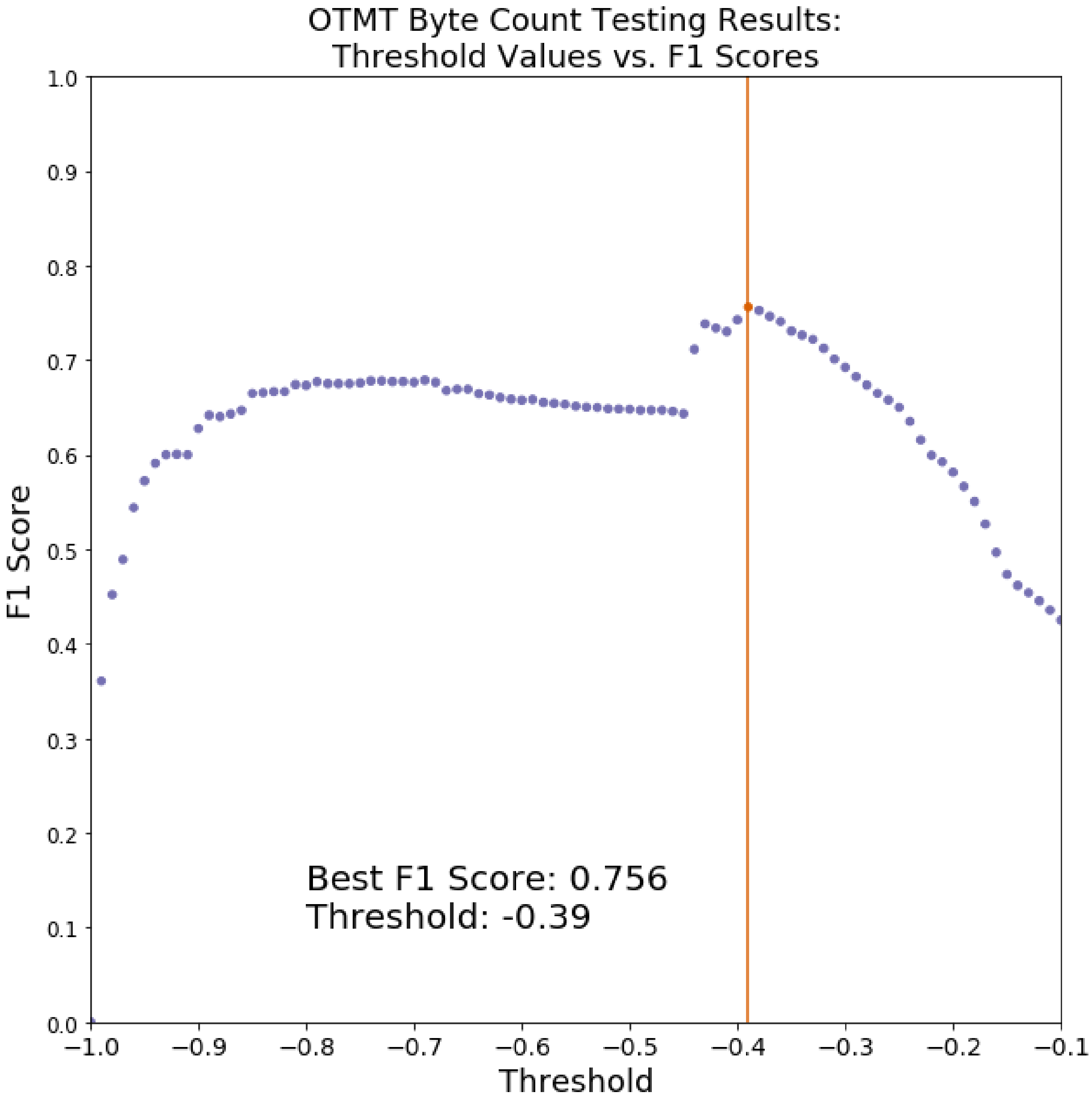
(b) Off-topic due to hacking on 2012-04-10

Fig. 81. `http://occupyevansville.org/` preserved in Archive-It collection 2950: *Occupy Movement 2011/2012*.

TABLE 13

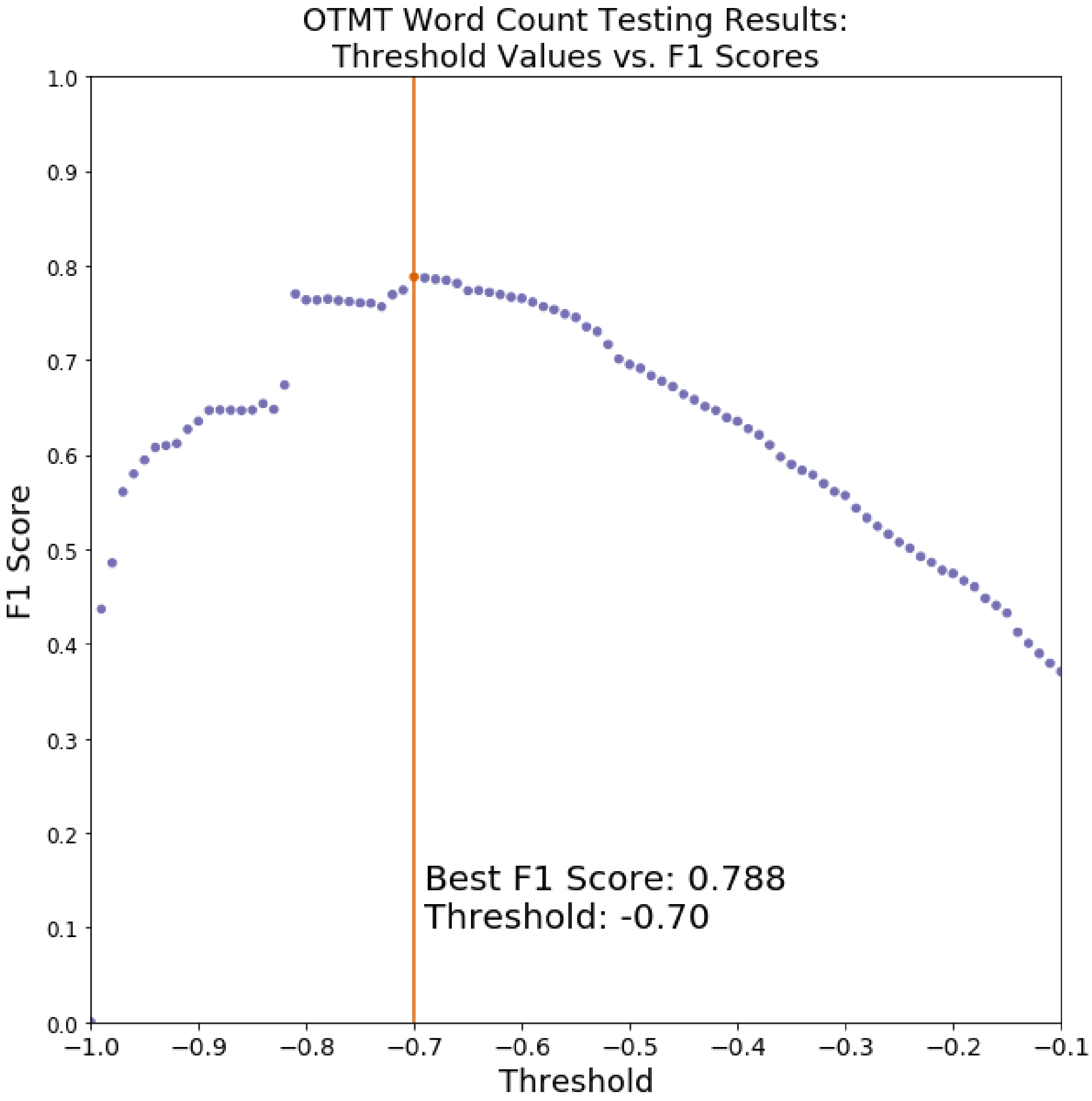
Similarity measures evaluated in our off-topic detection experiment.

Measure	**Fully Equivalent Score**	**Fully Dissimilar Score**	**Preprocessing Performed**
Byte Count	0.0	-1.0	No
Word Count	0.0	-1.0	Yes
Jaccard Distance	0.0	1.0	Yes
Sørensen-Dice Distance	0.0	1.0	Yes
Simhash of Term Frequencies	0	64	Yes
Simhash of Raw Content	0	64	No
Cosine Similarity of TF-IDF Vectors	1.0	0.0	Yes
Cosine Similarity of LSI Vectors	1.0	0.0	Yes



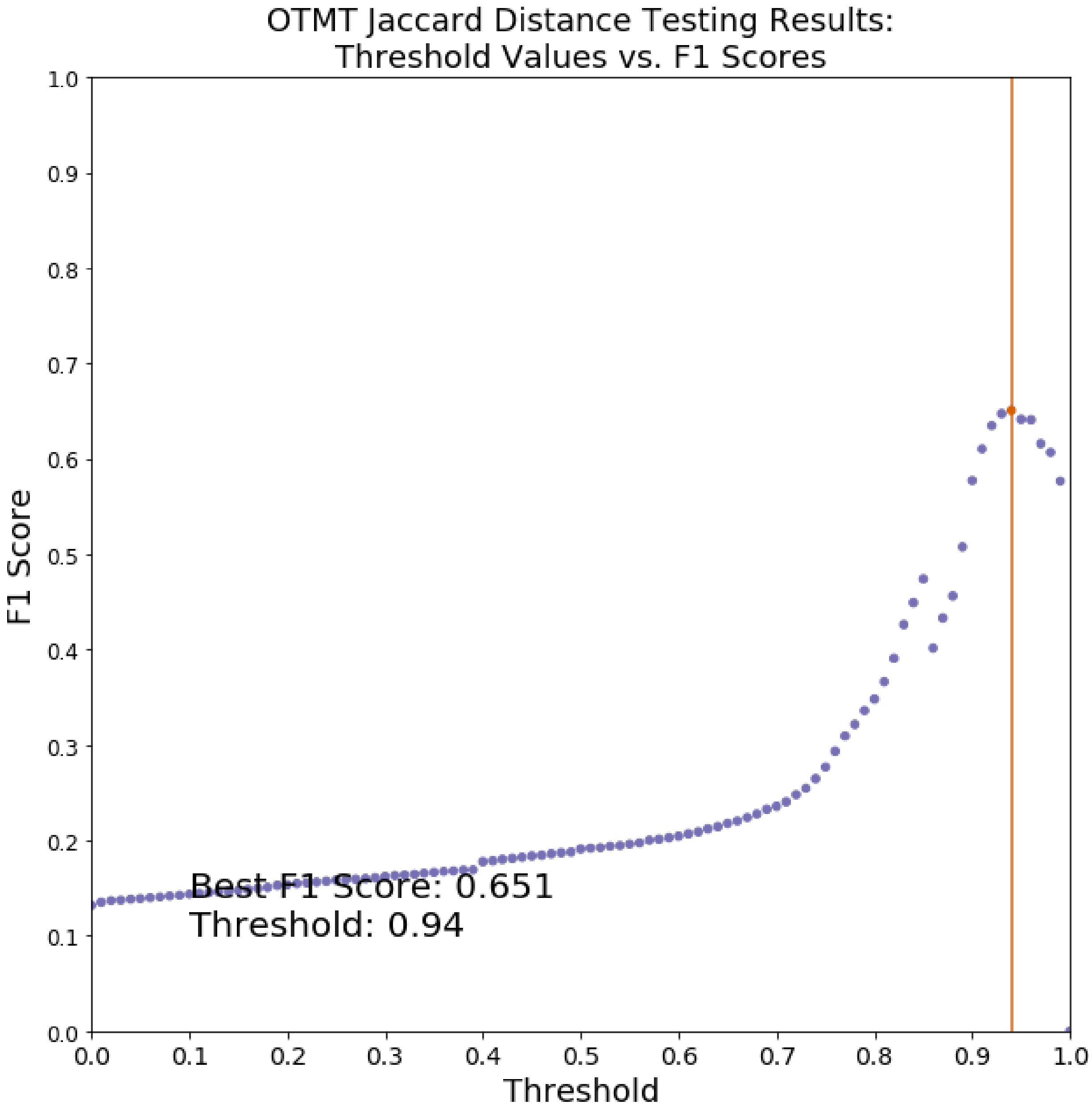
(a) Byte Count

Fig. 82. Scatter plots of threshold F_1 testing results for different similarity measures.



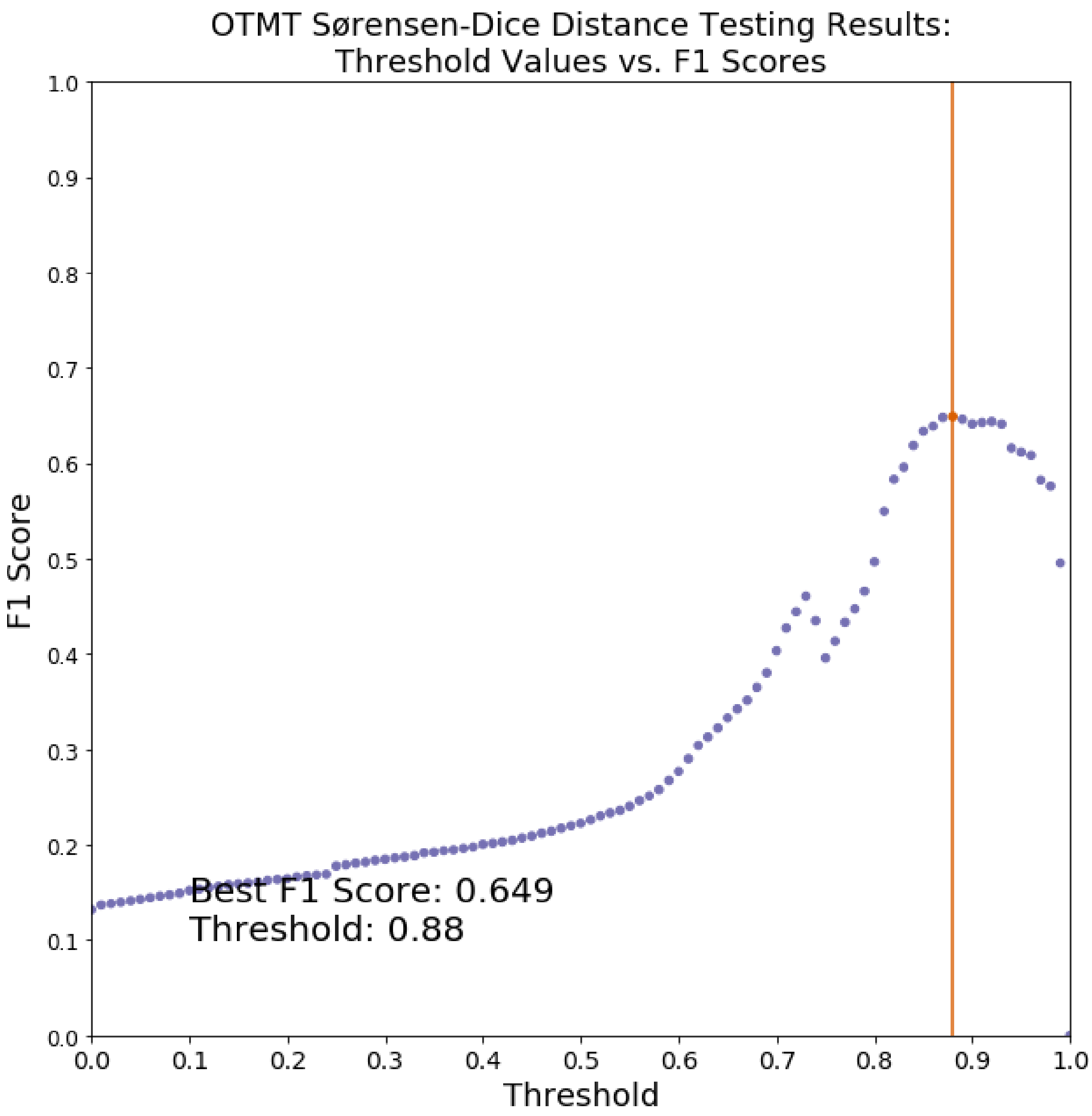
(b) Word Count

Fig. 82. (Continued) Scatter plots of threshold F_1 testing results for different similarity measures.



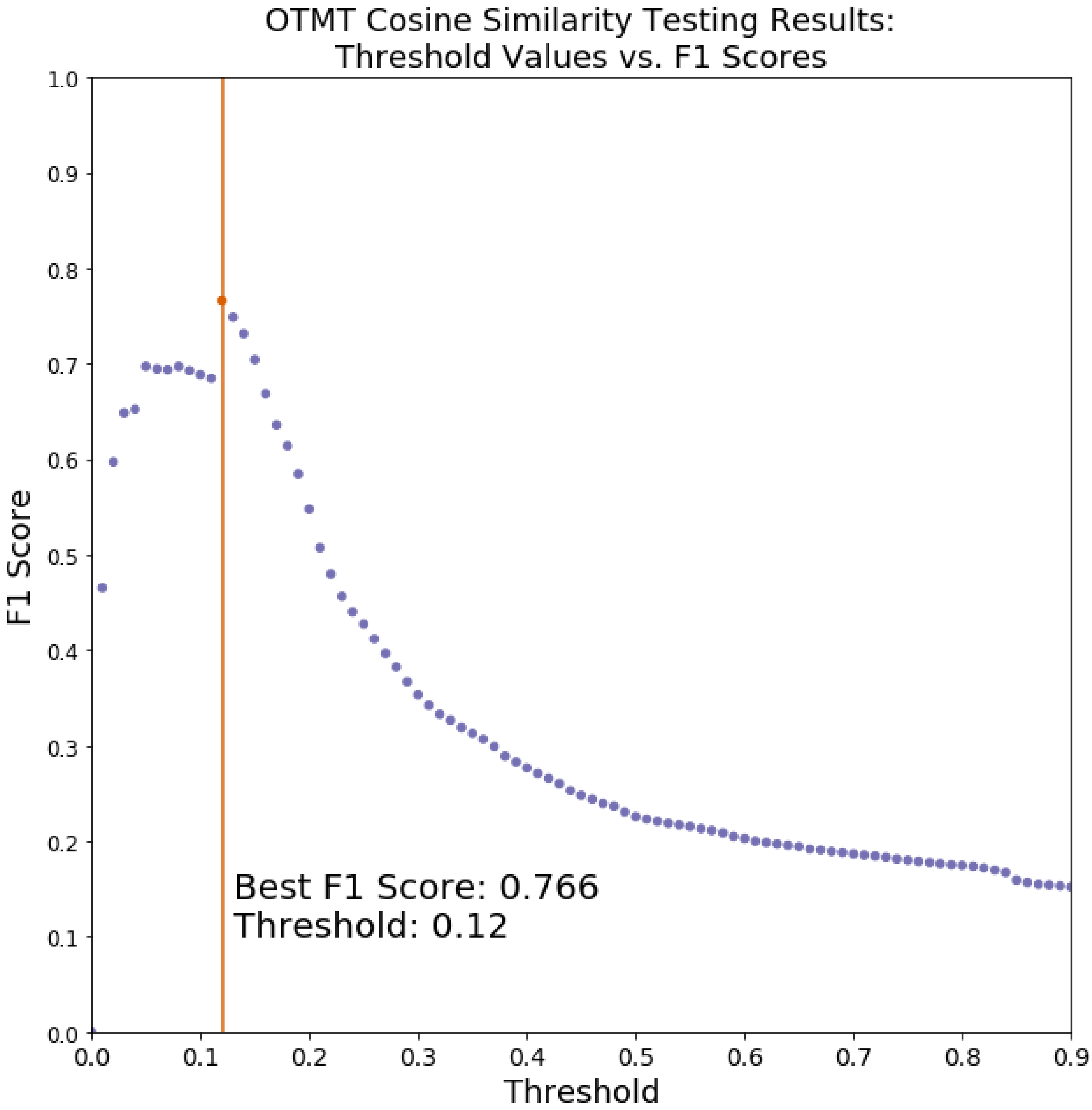
(c) Jaccard Distance

Fig. 82. (Continued) Scatter plots of threshold F_1 testing results for different similarity measures.



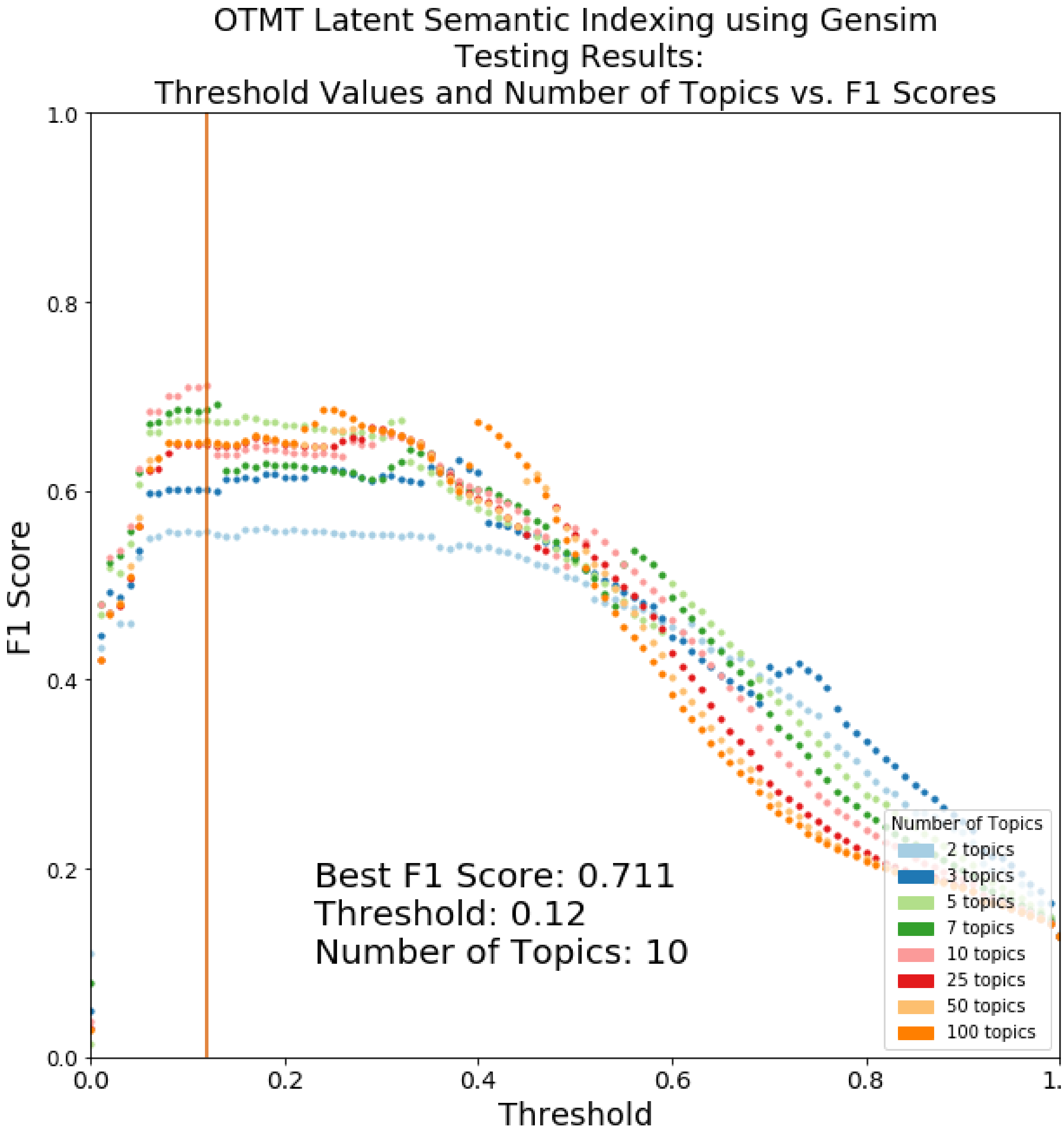
(d) Sørensen-Dice
Distance

Fig. 82. (Continued) Scatter plots of threshold F_1 testing results for different similarity measures.



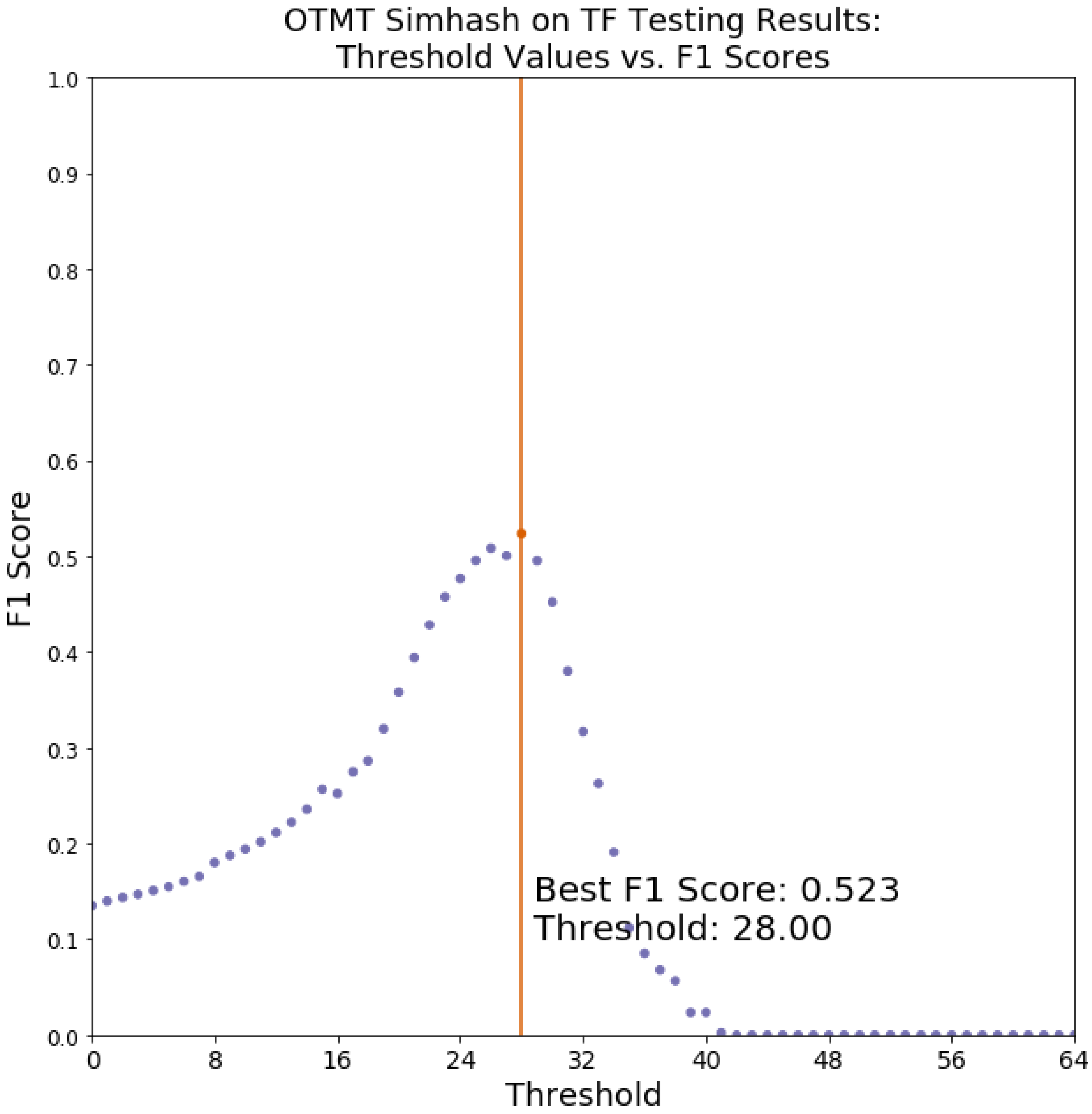
(e) Cosine Similarity
of TF-IDF Vectors

Fig. 82. (Continued) Scatter plots of threshold F_1 testing results for different similarity measures.



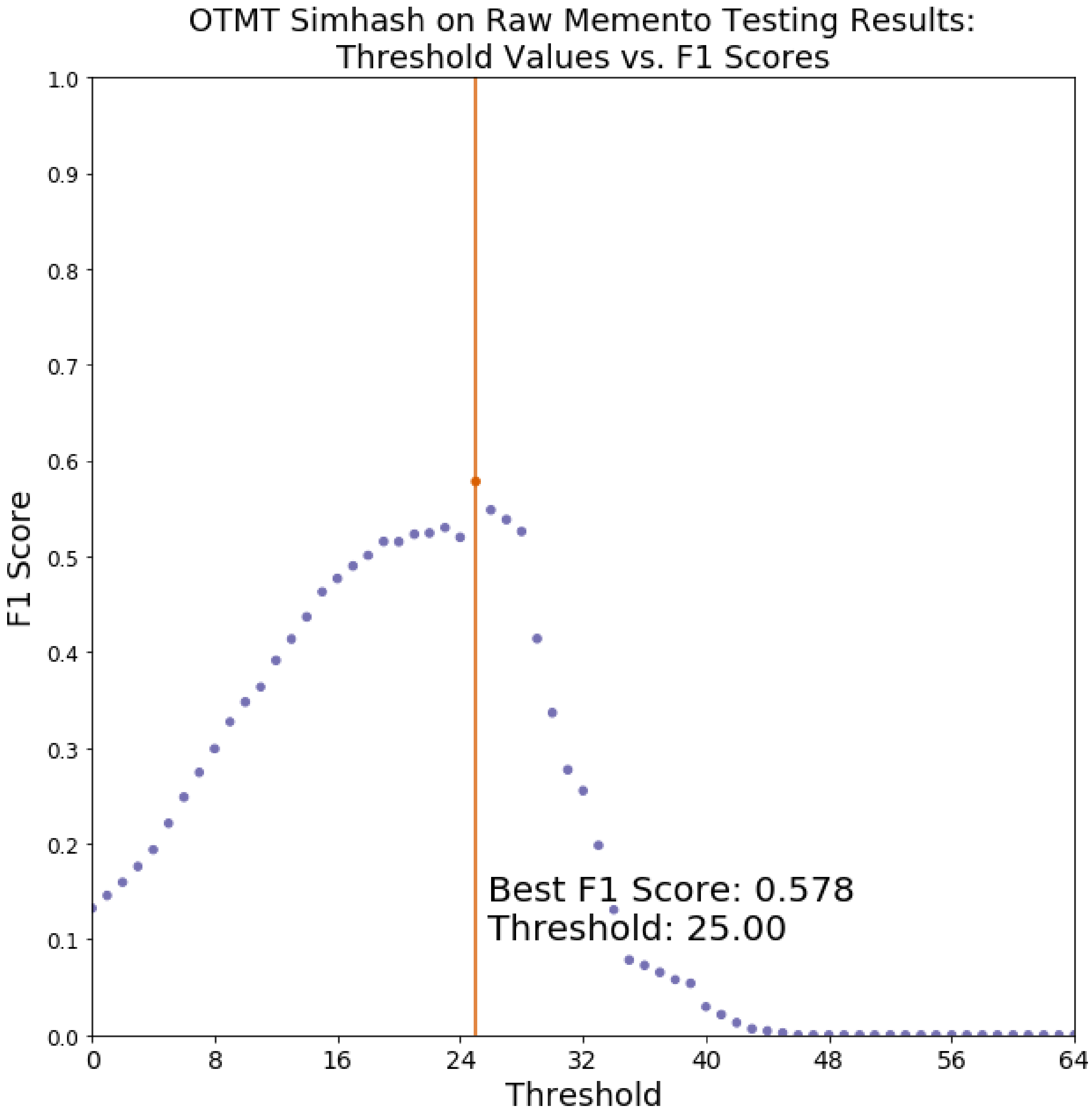
(f) Cosine Similarity
of LSI Vectors

Fig. 82. (Continued) Scatter plots of threshold F_1 testing results for different similarity measures.



(g) Simhash
of Term Frequencies

Fig. 82. (Continued) Scatter plots of threshold F_1 testing results for different similarity measures.



(h) Simhash
of raw memento content

Fig. 82. (Continued) Scatter plots of threshold F_1 testing results for different similarity measures.

TABLE 14

Distribution of the Gold Standard dataset.

Collection ID	Collection Name	# seeds	# mementos	# off-topic
1068	Human Rights	199	2302	95 (4%)
2358	Egypt Revolution and Politics	136	6886	384 (6%)
2950	Occupy Movement 2011/2012	255	6569	458 (7%)

4.2.1 EVALUATION OF OFF-TOPIC DETECTION

We conducted a study [164] to acquire reasonable default thresholds for each measure and determine which measures performed best for identifying off-topic mementos. We reused the dataset [152] from AlNoamany's 2016 off-topic work [10]. Table 14 displays information about the gold standard dataset used in this study. We executed our study against Archive-It in December 2017.

For evaluating these thresholds, we applied the F_1 measure (see Chapter 2.6.2). We also include Accuracy, also defined in Chapter 2.6.2, so we can compare our scores against AlNoamany's work. AlNoamany tested each similarity measure with 21 thresholds [10]. For more precision, we evaluated 100 thresholds for every measure but Simhash. For Simhash, we evaluated 64 thresholds because Simhash measures differences between documents by differences in bytes between hash values, and each document has a hash of 64 bytes.

We chose a measure and started with the lowest threshold for that measure. Our process then executed Algorithm 1 against a TimeMap with this threshold and measure. We recorded the scores and which mementos were declared off-topic by this combination. We then incremented the threshold and compared all mementos again. We repeated this process until we reached a designated upper limit. Then we selected another measure and repeated the process for that measure. Figure 82 displays visualizations of the results for each measure with each threshold on the x-axis and the resulting F_1 score on the y-axis.

For example, with byte count as the measure, we executed Algorithm 1 and saved the scores for each memento. We declared a memento off-topic if its score was less than -0.99. We then took each memento's byte count score and declared it off-topic if its score was less than -0.98. We then repeated for each memento with a threshold set at -0.97. We repeated this process in increments of -0.01 until we reached -1.

We then compared the off-topic determinations per threshold with the gold-standard data. From there, we were able to generate a corresponding F_1 score for each threshold value. We had assumed that testing the thresholds this way would help us discover threshold values close to those found by AlNoamany, and we did get close in some cases. The F_1 scores, however, are often worse, and we identified some differences between our environment and hers.

We used the justext library [269] for boilerplate removal, whereas AlNoamany used boilerpipe [185]. We used NLTK [35] for tokenization and stemming whereas AlNoamany used scikit-learn [260]. These different preprocessing techniques are not the only potential cause of these score differences. We experienced different download errors than AlNoamany did when she conducted the study in 2015. We ran our experiment on Archive-It 4.9 [273] in 2017, and Archive-It had received improvements to its playback engine since 2015 [21] which alters its behavior. These changes in behavior among web archives have been researched by Ainsworth et al. [3] and Aturban et al. [23].

We tried byte count threshold scores between 0 and 1 at increments of 0.01. A threshold score of -0.39 produces the best F_1 score. This score indicates that the off-topic memento is 39% smaller than the first memento in the TimeMap. AlNoamany's findings suggested a threshold value of -0.65, making the our results more strict. This result is one case where our F_1 score of 0.756 is higher than AlNoamany's finding of 0.584.

We used the same range and increments to test word count. A threshold score of -0.70 produces the best F_1 score. This score means that the off-topic memento has 70% fewer words than the first memento in the TimeMap. AlNoamany's findings suggested a threshold value of -0.85, again making our results of -0.70 more strict. This makes sense for off-topic mementos that result from technical issues, defacements, or web site redesigns because error messages (or hackers' statements) often contain far less text than the article they replace.

For Jaccard and Sørensen-Dice, we used the same range of threshold values from 0 to 1, at increments of 0.01. A threshold score of 0.94 has the best F_1 value for Jaccard. AlNoamany discovered that a value of 0.05 was best, but her work used the pure Jaccard Index rather than the Jaccard distance. This threshold is consistent with her findings as the Jaccard distance is $1-$ Jaccard index. However, the F_1 score of 0.651 is higher than her F_1 result of 0.538 for Jaccard. AlNoamany did not attempt Sørensen-Dice. Its best threshold value is close to that of Jaccard at 0.88. With both Sørensen dice and Jaccard, the considered memento must be very dissimilar from the first memento for our algorithm to mark it as off-topic.

Simhash scores range from 0 to 64, based on the number of bits different between simhashes.

TABLE 15

Off-Topic detection TimeMap measures sorted by best F_1 score and compared with AlNoamany's results.

	AlNoamany's results			Results of this study		
Similarity Measure	**Best F_1 Score**	**Corresponding Accuracy**	**Corresponding Threshold**	**Best F_1 Score**	**Corresponding Accuracy**	**Corresponding Threshold**
Word Count	0.806	0.982	-0.85	0.788	0.971	-0.70
Cosine Similarity of TF-IDF Vectors	0.881	0.983	0.15	0.766	0.965	0.12
Byte Count	0.584	0.962	-0.65	0.756	0.965	-0.43
Cosine Similarity of LSI Vectors		Not tested		0.711	0.965	0.10 with 10 topics
Jaccard Distance	0.538	0.962	0.95*	0.651	0.953	0.94
Sørensen-Dice Distance		Not tested		0.649	0.953	0.88
Simhash on raw memento content		Not tested		0.578	0.934	38
Simhash on TF		Not tested		0.523	0.942	34

* A derived value is shown for easier comparison. AlNoamany's threshold was 0.05, but she used Jaccard Index rather than Distance.

TABLE 16

The top 4 scoring measures combined in groups of 2.

Measure	**Best F_1 Score**	**Corresponding Thresholds**	**Corresponding Accuracy**
Cosine of LSI, Word Count	0.789	(0.01, -0.70)	0.971
Cosine of TF-IDF, Word Count	0.788	(0, -0.70)	0.971
Word Count, Byte Count	0.788	(-0.70, -0.94)	0.971
Cosine of TF-IDF, Byte Count	0.766	(0.12, -0.95)	0.965
Cosine of LSI, Cosine of TF-IDF	0.766	(0.12, 0.12)	0.965
Cosine of LSI, Byte Count	0.759	(0.01, -0.39)	0.965

AlNoamany did not test Simhash. For Simhash based on term frequencies, we achieve the best F_1 score of 0.523 with a threshold value of 28 bits. If a user chooses to feed just the raw content into the function, we achieve the best F_1 score of 0.578 at 25 bits.

As noted above, cosine similarity scores range from 1 to 0. We achieve very close to AlNoamany's threshold result of 0.15 for the cosine of TF-IDF vectors, with the highest F_1 score at a threshold of 0.12. This score is very close to the score of complete dissimilarity between documents. AlNoamany achieved an F_1 score of 0.881 compared to our 0.766.

Latent Semantic Indexing (LSI) assigns each of a TimeMap's mementos to a topic. If a memento has a poor correlation to its assigned topic, then it is likely off-topic. Applying the cosine similarity of LSI vectors produces an F_1 score of 0.711. The LSI algorithm requires that one specify the number of topics. We tried values of 2, 3, 5, 7, 10, 25, 50, and 100 topics. The value of 10 worked best for testing with our gold-standard data. Unfortunately, randomness exists LSI's operation in Gensim [250]. Five different runs with LSI produced F_1 scores near or at 0.711, but their corresponding thresholds ranged from 0.08 to 0.12. We took the mean of the scores and set the default threshold at 0.10.

Table 15 shows the best test scores for each measure. We show AlNoamany's results for comparison. Word count has the best F_1 score, followed by cosine similarity of TF-IDF Vectors. Byte count and cosine similarity of LSI Vectors score at third and fourth place, respectively.

We considered the possibility that we could do better with multiple measures. Structural measures require less time to execute than more semantic measures like cosine similarity. If we can short-circuit the process with a structural measure, we can eliminate those off-topic mementos before reviewing with a more time-intensive measure. AlNoamany did so and found that word count and cosine similarity worked best. We evaluate multiple measures as a *logical OR* – only one of the measures has to identify a memento as off-topic for it to be considered off-topic. We tried different combinations of thresholds just as before and recorded the corresponding F_1 and accuracy scores. Table 16 displays these combinations. The cosine of LSI vectors scores a slightly higher F_1 and the same accuracy as word count. The word count score exerts more influence than its partner measures in all cases where it is present, making both TF-IDF and LSI cosine measures require stricter thresholds to be successful. Combining these measures does not improve the F_1 score.

4.2.2 SUMMARY

As part of the storytelling process of selecting exemplars to represent a collection, filtering off-topic mementos is an important first step, and thus this work addresses RQ2. We evaluated eight different similarity measures with respect to this filter and found that word count has the best F_1 score. This is likely because most off-topic content consists of status messages from technical issues rather than an article drifting from the content of the collection. Next we discuss how filtering off-topic mementos is only part of the solution for selecting exemplars.

4.3 APPLYING PRIMITIVES TO SELECT EXEMPLARS

Removing off-topic mementos is just one type of filtering that we can apply to a web archive collection. As noted in Chapter 3.1, many have tried to automatically select exemplar sentences and images from an individual document as well as select exemplars documents from a corpus. With more analysis of this work, a model of primitives becomes evident. In Chapter 3 we mentioned the concepts of *filtering*, *scoring*, and *clustering* as potential primitives. Here, we expand this list to include a few other primitives that address the many other types of possible exemplar selection.

4.3.1 PRIMITIVES

Our first primitive is **sample**. Perhaps the simplest method of creating a small sample from a collection is randomly choosing k mementos from the N that exist in the collection such that $k << N$. While this is easiest if one already has a list of URI-Ms, we may need a very large value of k to capture a sample that takes into account the structural features of the collection, nor does random sampling evaluate the quality of the mementos it selects. For example, it is possible, though unlikely, that our k mementos will be off-topic. At this point we are relying upon **probabilistic sampling**, hoping that our k mementos will make a good story.

If probabilistic sampling has the potential to yield lower quality results, how do we improve our chances? Our second primitive is **filter**. With filter, we try to remove items of low information value from our sample. In the last section we covered identifying off-topic mementos, but this is not the only filtering that we can perform. Recall that web archives are built by automatic crawling; therefore, there are many mementos that contain the exact same content. To improve our chances of getting a good sample, we can apply Simhash [54] to filter out these near duplicates as well. Now we have combined our sample primitive with two filter primitives to create a better sample. We refer to this new sampling algorithm as **filtered random**. Figure 83 provides a simplified diagram of how this works: (1) we filter out off-topic mementos, (2) we filter out near duplicates, (3) we randomly sample k mementos from the remainder.

Filter and sample are similar, but we differentiate them by intent. Filter explicitly reduces a collection by some intelligent method, like excluding off-topic mementos. The sample primitive

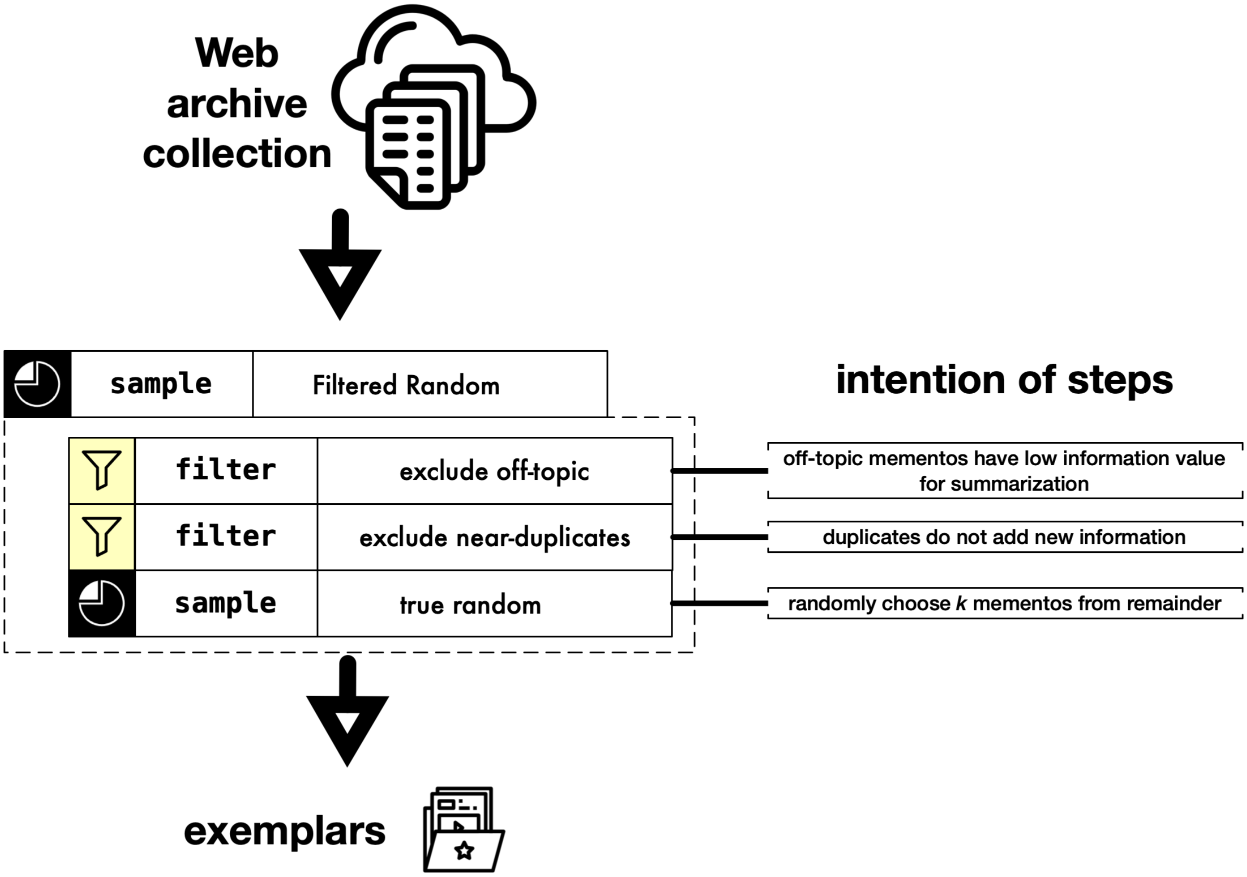

Web archive collection
sample
Filtered Random
filter
exclude off-topic
filter
exclude near-duplicates
sample
true random
intention of steps
off-topic mementos have low information value for summarization
duplicates do not add new information
randomly choose k mementos from remainder
exemplars

Fig. 83. We can select exemplars from a web archive collection by first removing off-topic and near-duplicates before randomly choosing from the remainder.

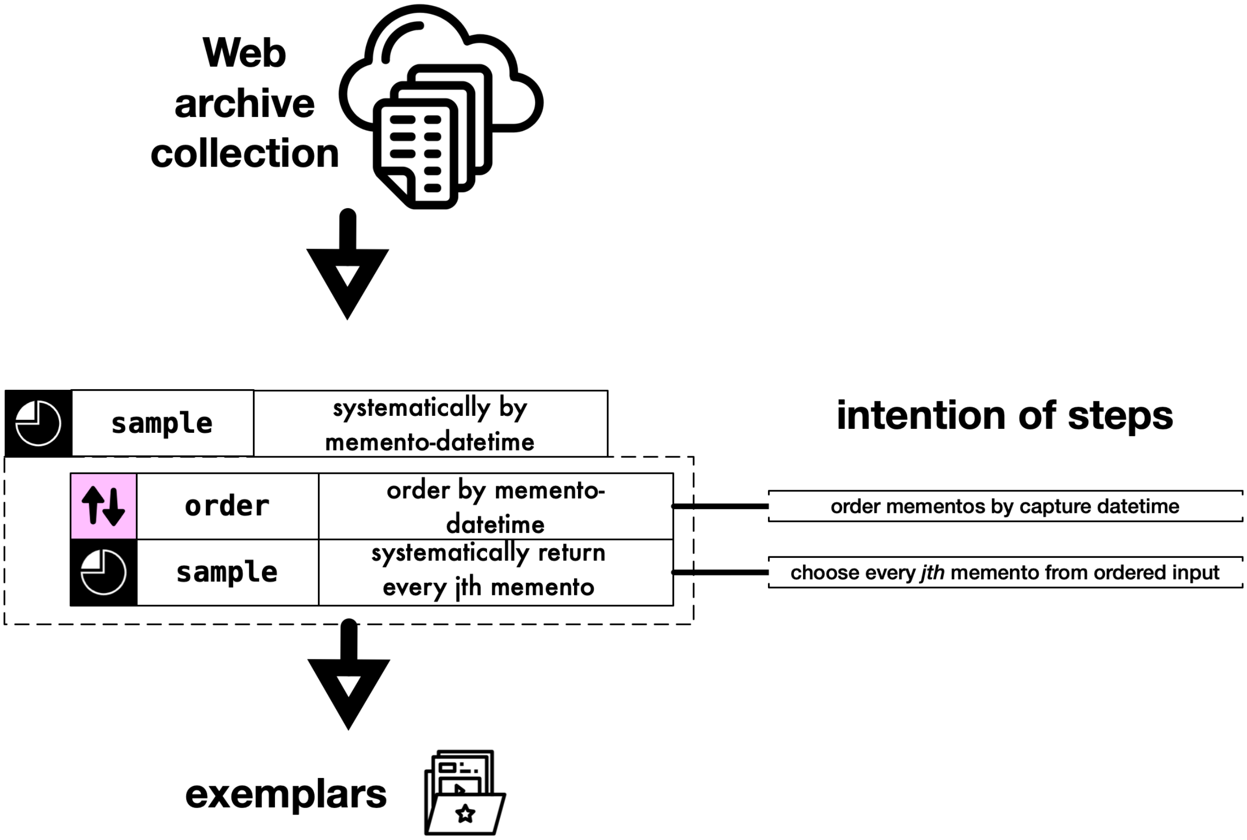

Web archive collection
sample
systematically by memento-datetime
intention of steps
order
order by memento-datetime
order mementos by capture datetime
sample
systematically return every jth memento
choose every *jth* memento from ordered input
exemplars

Fig. 84. We can select exemplars from a web archive collection by first ordering a collection, and then systematically sampling every j^{th} memento from the remainder.

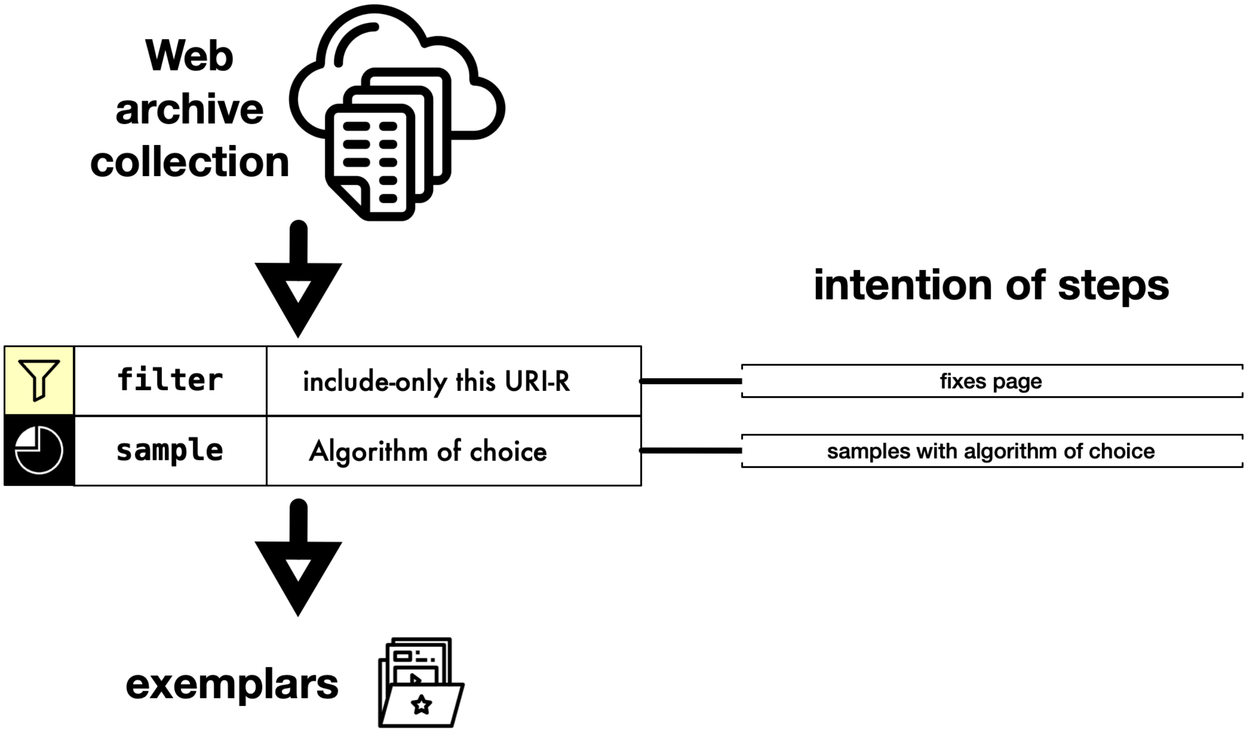


Fig. 85. Creating a *fixed page, sliding time* story (FPST) requires first filtering the collection by URI-R prior to sampling.

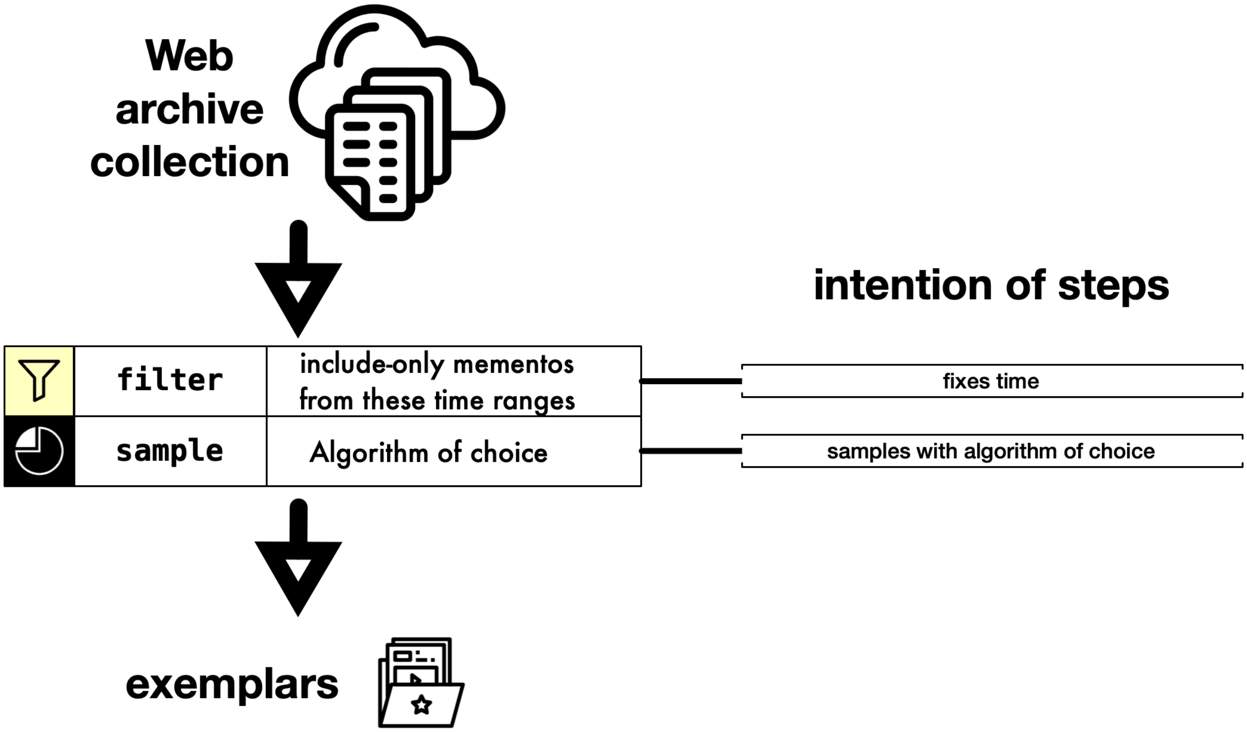


Fig. 86. Creating a *sliding page, fixed time* story (SPFT) requires first filtering the collection by a range of memento-datetimes prior to sampling.

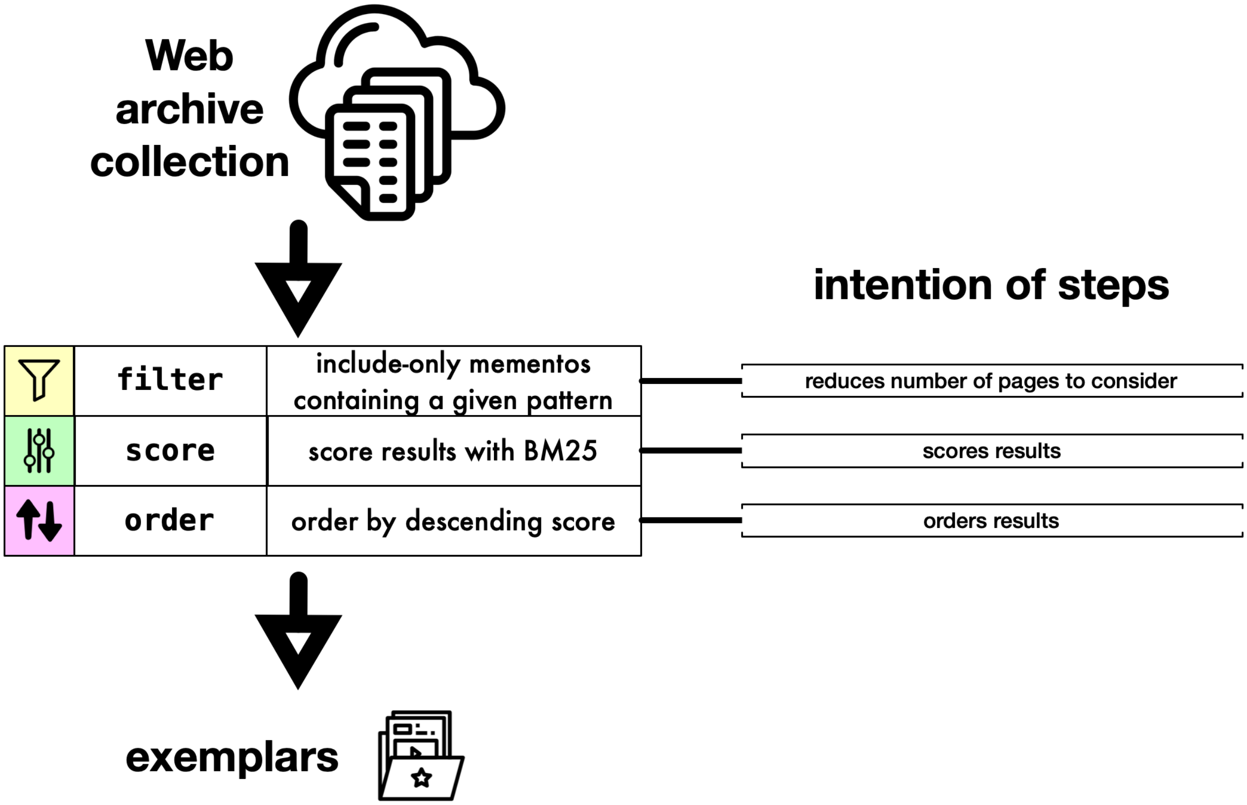


Fig. 87. One could create a story by filtering for a specific pattern, like a keyword and then they could score that output with a function like BM25 and order the results by descending score.

is either a probabilistic primitive, like true random or systematic sampling, or it is a combination of other primitives to make a new algorithm for selecting exemplars.

Instead of randomly sampling mementos, we can systemtically choose every j^{th} memento from the input until we reach our k mementos. Unfortunately, we do not know how the input is organized, so we have to organize it in some way so that our systematic sampling makes sense. Thus, we introduce a new primitive, **order**, that allows us to help the systematic sampling produce more meaningful results. In the case of our algorithm from Figure 84 we apply order to sort the collection by memento-datetime and then systematically sample every j^{th} memento from this sorted list.

At this point, we are no longer probabilistically sampling, but instead we are performing **intelligent sampling**. Recall from AlNoamany's work [12] that study participants could tell the difference between randomly generated stories and those produced by her algorithm. The more intelligence we introduce into our selection, the better our results, but the less chance we have of creating a statisically representative sample of the collection. For summarization, we need to balance these concepts so that our resulting exemplars are representative while also being of high quality. For storytelling, high quality equates to having good information scent. For other work, high quality may equate to representing a minority topic in the collection. High quality could also equate to producing mementos of low information value, possibly to analyze how many pages in the collection suffered web site defacement. Thus, not only is there more than one way to select exemplars, but there are different use cases for many different summaries and story types.

Consider Rustam's use case from Chapter 1. He wanted to view how a given web page changed over time to produce a *fixed page, sliding time* (FPST) story. Unless his chosen web archive collection only collected a single original resource, it is likely his story cannot be satisfied through random sampling alone. He must filter the collection before sampling, as shown in Figure 85. In this diagram, we do not prescribe a sampling algorithm because that, too, depends on Rustam's use case for the output of that filter primitive. He may just filter the collection and produce such a small sample that he can manually review them afterward.

Also recall Olayinka's use case from Chapter 1. She wanted to view what the news covered during a given time period to produce a *sliding page, fixed time* (SPFT) story. If her web archive collection does not match her desired time period, then she must apply a filter first before continuing, as shown in Figure 86.

Someone might also be interested in pages that only contain a given pattern of n-grams. Instead of filtering by URI-R or memento-datetime, we could filter to only include pages that contain that pattern. If this filter action is not sufficient and we still need to make sense of these documents, we need an additional primitive, **score**. With score, we elevate the importance of some mementos over others, allowing us to make additional decisions on presentation or future filtering. We could score each document from our filtered list by how well it matches this pattern with an algorithm like BM25 [280]. Then we could order those results by descending score. Our filter step reduced our set of mementos to only those that meet our definition of high information value. Our score step imbues each memento with a value. Our order step provides additional information by elevating the highest scoring memento to the top of the list. At this point we have created a primitive search engine (Figure 87) .

In some use cases the user knows what aspect they want to explore in a collection. Rustam knows the page that he wants to study. Olayinka has decided on her time period. In our primitive search engine example, the user must have an idea of the n-gram pattern for which they want to filter. For other use cases, the user starts with no prior knowledge. Natasha, Elbert, and Ling only have a web archive collection as input. They just want to create a summary of it. They can work with a *sliding page, sliding time* (SPST) story. To create SPST stories, we introduce our final primitive, **cluster**, in the next section.

4.3.2 PRIMITIVES COMBINE TO CREATE DSA1

AlNomany's algorithm, introduced in Chapter 3.5, creates SPST stories. We refer to this algorithm as **DSA1**. We can apply our primitives to create her algorithm. She first filters to exclude off-topic and near-duplicate mementos. She then filters again to include only English language mementos. For a large collection, this intelligent sample is still too large to process. We must give additional meaning to the remaining mementos so we can reduce the set further. The cluster primitive uses a given algorithm to group mementos by some common feature. DSA1 performs two clustering steps, once in the dimension of time and a second in the dimension of aboutness. The first clustering step surfaces temporal aspects of the collection. The second clustering step ensures that novel mementos have exposure in future steps. DSA1 then scores the resulting mementos with a function that places a high value on mementos with long paths, belonging to specific page categories, and having low memento-damage. After this, she filters one final time so that only the highest scoring memento from each cluster becomes an exemplar. The final ordering step ensures that the exemplars will make a good story because they will flow by publication date. Figure 88 shows how our primitives model of DSA1 maps to AlNoamany's original algorithm diagram (Figure 70 on page 130.)

The DSA1 algorithm applies two filters from our filtered random algorithm. Rather than randomly sampling mementos from the remainder, it tries to make sense of the collection through clustering. Clustering breaks the collection into subsets and it also infuses the mementos in those subsets with new meaning. In step 4 of Figure 88, we give meaning to each set of mementos by placing it into a different cluster based on time. This is different from just extracting the memento-datetime. It instead creates several sub-collections from which we can perform further actions. In step 5, we cluster again to instill topical information among each of the temporal clusters. By clustering, we create several subsets that represent different aspects of the collection and also ensure that those aspects are all scored separately. In step 6, DSA1 scores all of the mementos in these clusters so we can select the highest scoring ones for our story in step 7. Without clustering, we would still only have one group from which to choose the highest scoring memento. With that one group, scoring by path depth, page category, and memento-damage would not help us choose exemplars that represent these collection aspects.

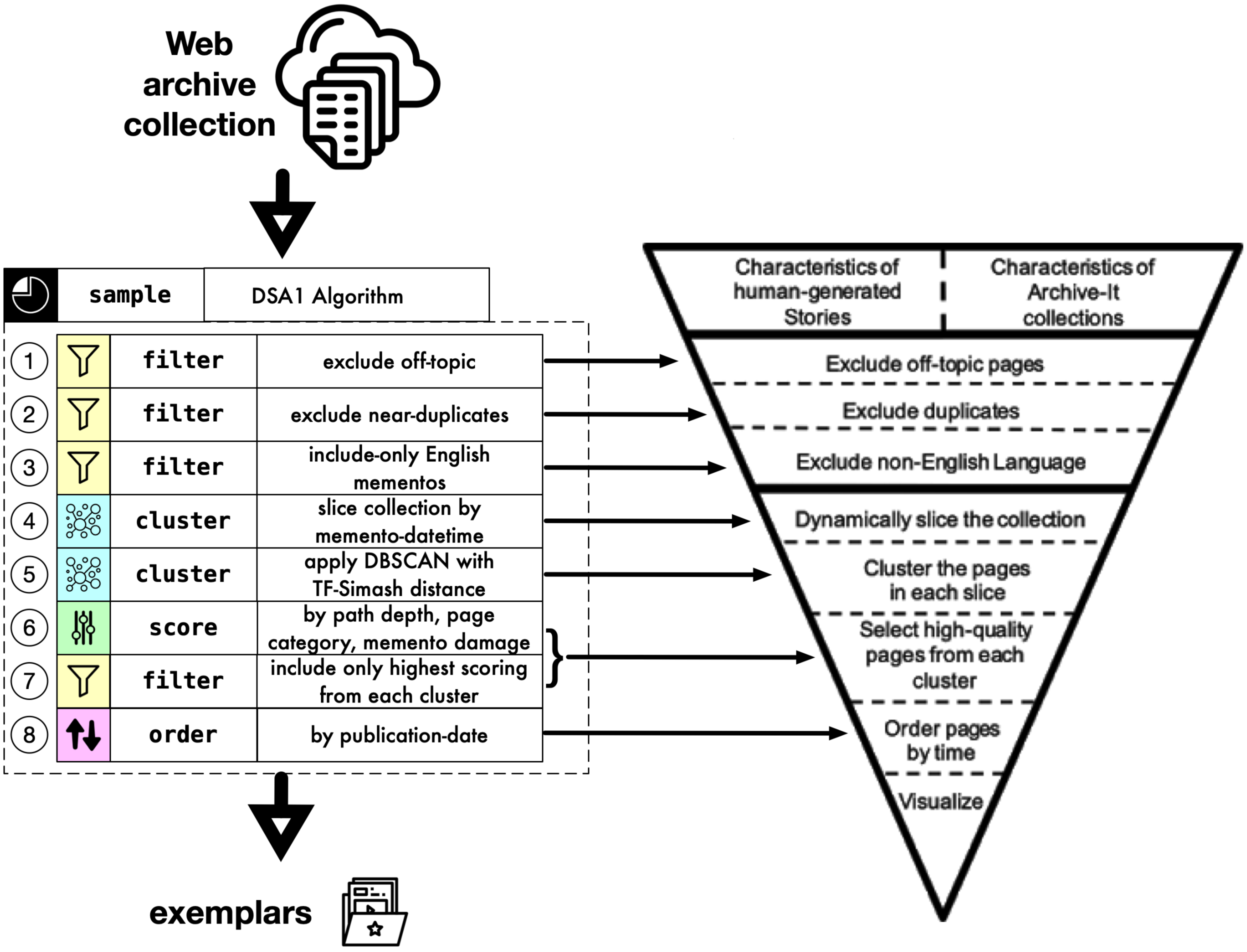
4.3.3 A FOCUS ON GROWTH CURVES AND SURROGATE METADATA IN DSA2

DSA1 is not the only possible algorithm. One of the points of this section is that these primitives can be combined in many ways to produce algorithms with different effects. DSA2 attempts to address several shortcomings of DSA1.

The first shortcoming is that DSA1 temporally clusters the collection by creating $k \approx 28$ slices. At this step, DSA1 has the potential to produce more than 28 mementos because DSA1 step 5 (Figure 88) clusters a second time. This is typically not a problem for small collections because, in practice this second clustering does not yield many secondary clusters, but for a large collection, it has the potential inflate the size of the story well beyond $k \approx 28$ mementos.

The second shortcoming has to do with the mechanism of creating time slices. After DSA1 step 3, there are j mementos. As part of step 4, DSA1 creates k slices. It then sorts the j mementos by memento-datetime. Starting from the first memento in this list, DSA1 fills each slice until it reaches a size of $\lceil \frac{j}{k} \rceil$ mementos, and then starts filling the next slice. This works well for collections whose growth curve has its mementos growing continuously, but if the memento growth curve is early or late, then it is possible for DSA1 to place mementos with very similar datetimes in separate slices. This has the net effect of heavily representing the portions of the collection that are temporally dense.

The third shortcoming deals with certain assumptions in the DSA1 scoring function. This function takes into account the type of page that the memento represents. News is scored more highly than image sharing services, which are scored more highly than social media posts, which are scored more highly than blogs. The function also tries to find quality mementos by favoring those with low memento-damage [46]. These two decisions were made to favor mementos that



		sample	DSA1 Algorithm
1	▽	filter	exclude off-topic
2	▽	filter	exclude near-duplicates
3	▽	filter	include-only English mementos
4		cluster	slice collection by memento-datetime
5		cluster	apply DBSCAN with TF-Simash distance
6		score	by path depth, page category, memento damage
7	▽	filter	include only highest scoring from each cluster
8	⇅	order	by publication-date

Web archive collection
exemplars
Characteristics of human-generated Stories
Characteristics of Archive-It collections
Exclude off-topic pages
Exclude duplicates
Exclude non-English Language
Dynamically slice the collection
Cluster the pages in each slice
Select high-quality pages from each cluster
Order pages by time
Visualize

Fig. 88. AlNoamany's Algorithm, referred to in this dissertation as DSA1, combines our primitives to select exemplars for SPST stories.

would produce better surrogates for our stories.

DSA2 (Figure 89) tries to address these shortcomings with the additional knowledge that we have gained in the years since DSA1's development. Per Chapter 4.1 we now have a better understanding of the growth curves of web archive collections. Rather than applying the time slices, DSA2 step 3 uses k-means clustering to ensure all memento-datetime regions of the collection have better representation, regardless of growth curve. In step 4, DSA2 applies LDA topic modeling rather than DBSCAN clustering with TF-Simhash distance. DBSCAN does not accept a number of clusters as an argument, discovering the clusters on its own, if any. This is why DSA1's use of DBSCAN sometimes produces no additional clusters. With LDA topic modeling, we can specify the number of topics desired, giving us better control over how we get to $k \approx 28$ mementos.

$$S = \begin{cases} \lceil 28 + \log_{10}(N) \rceil & \text{if } N > 767 \\ \lceil \sqrt(N) \rceil & \text{if } N \leq 767 \end{cases} \tag{14}$$

$$S_t = S_l = \lceil \sqrt(S) \rceil \tag{15}$$

Equation 10 from Chapter 3.5 is AlNoamany's method for determining the number of clusters for time slicing in DSA1. To determine the number of clusters for each clustering step in DSA2, we update Equation 10 with Equation 14 based on experience with generating stories from AlNoamany's method. Here N indicates the number of mementos under consideration. Our boundary of 767 was determined empirically because stories without the second case would often contain poor quality mementos because the resulting clusters would be too small.

DSA2 allows us to specify the size of each cluster. In Equation 15 we take the square root of S to produce an input of S_t for the number of temporal clusters for DSA2 step 3. We do the same to arrive at S_l for the number of lexical clusters (number of topics) for DSA2 step 4. Because $S = S_t \cdot S_l$, we have a chance of reaching our target of $S \approx 28$ for our story.

In step 5 of DSA2, we can take advantage of the knowledge of how surrogates are created. In Chapter 5.3, we will cover how to create surrogates even though they lack the metadata covered in Chapter 2.4. With this knowledge, the scoring function scores highly those mementos that contain all of this metadata. If that fails, then it favors a memento with more text and a higher number of images so that automatic summarization algorithms have more candidates from which to select.

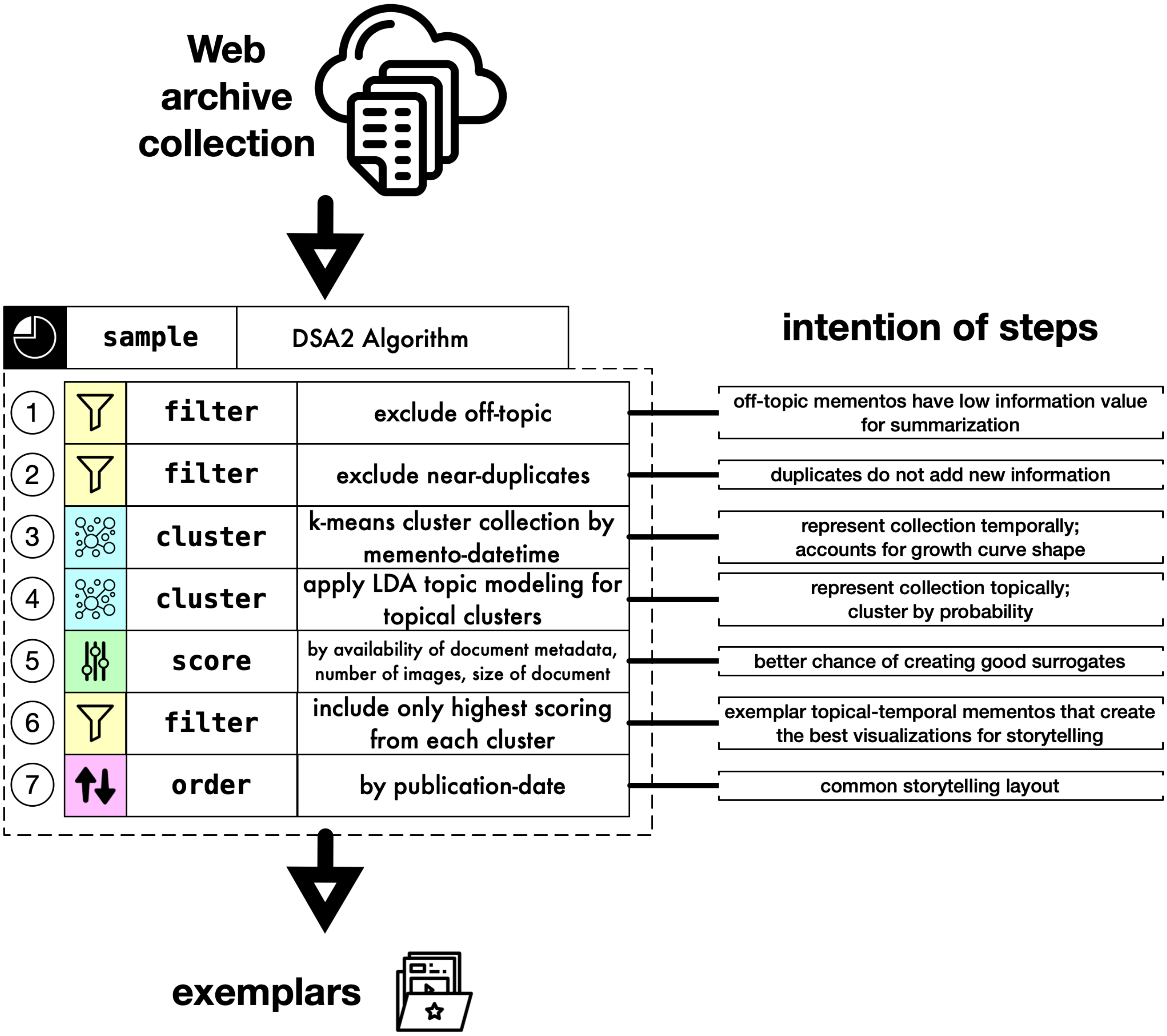

Web archive collection
sample
DSA2 Algorithm
intention of steps
1
filter
exclude off-topic
off-topic mementos have low information value for summarization
2
filter
exclude near-duplicates
duplicates do not add new information
3
cluster
k-means cluster collection by memento-datetime
represent collection temporally; accounts for growth curve shape
4
cluster
apply LDA topic modeling for topical clusters
represent collection topically; cluster by probability
5
score
by availability of document metadata, number of images, size of document
better chance of creating good surrogates
6
filter
include only highest scoring from each cluster
exemplar topical-temporal mementos that create the best visualizations for storytelling
7
order
by publication-date
common storytelling layout
exemplars

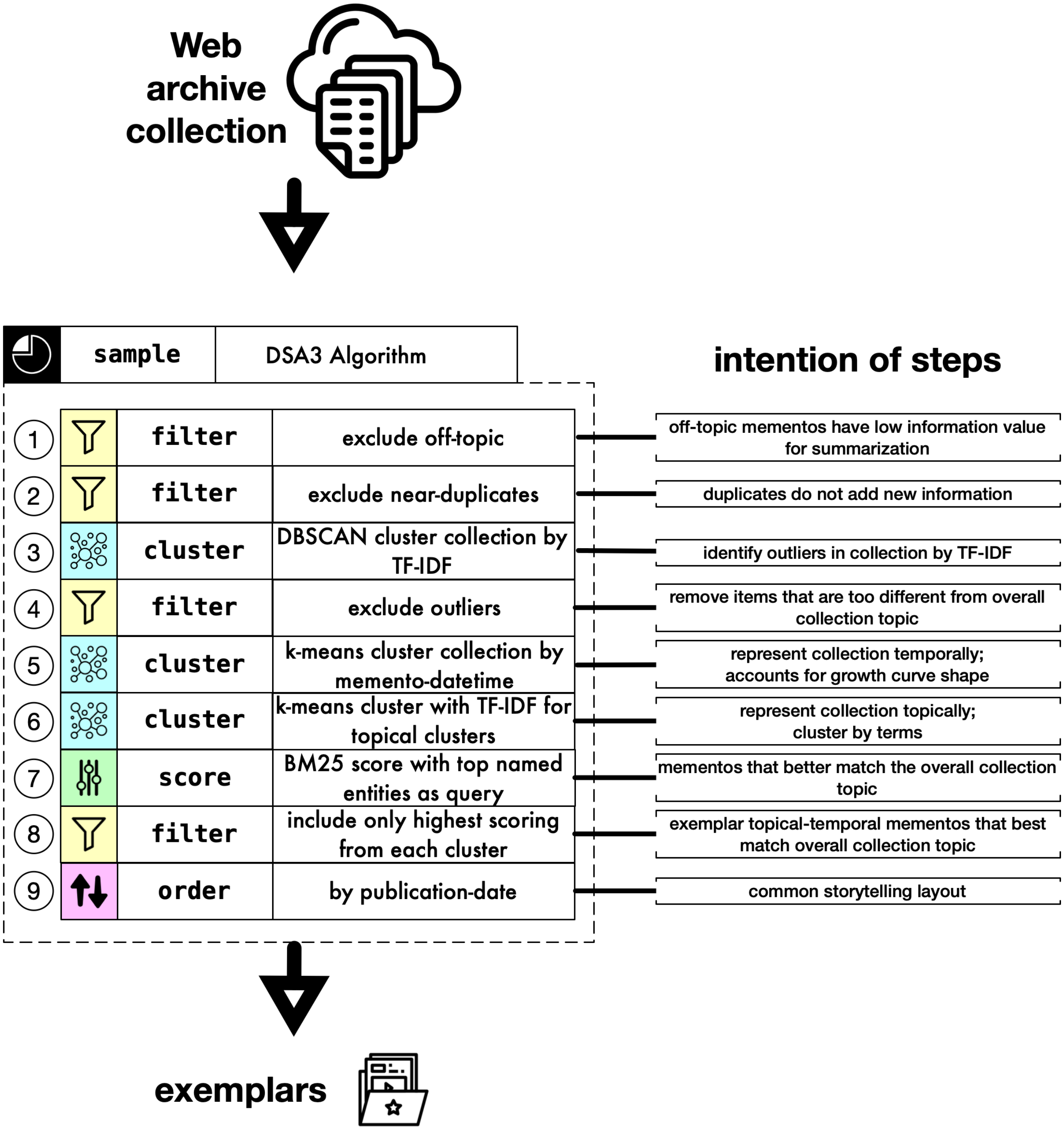
Fig. 89. The DSA2 algorithm selects mementos that should produce better surrogates.

$$
\begin{aligned}
&C_s = w_c \cdot f_c + w_t \cdot f_t + w_d \cdot f_d + w_i \cdot f_i \\
&f_c = \begin{cases} 1 & \text{if } \texttt{twitter:card} \text{ present and not empty} \\ 0 & \text{otherwise} \end{cases} \\
&f_t = \begin{cases} 1 & \text{if } \texttt{twitter:title} \text{ or } \texttt{og:title} \text{ present and not empty} \\ 0 & \text{otherwise} \end{cases} \\
&f_d = \begin{cases} 1 & \text{if } \texttt{twitter:description} \text{ or } \texttt{og:description} \text{ present and not empty} \\ 0 & \text{otherwise} \end{cases} \\
&f_i = \begin{cases} 1 & \text{if } \texttt{twitter:image} \text{ or } \texttt{og:image} \text{ present and not empty} \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{16}
$$

The DSA2 scoring function first applies the **simple card score** shown in Equation 16. This function takes into account existing social media social card metadata standards. As shown, f_c has a value of 1 if the memento lists a non-empty value for the `twitter:card` field and 0 otherwise. The f_t variable has a value of 1 if the memento has a non-empty `twitter:title` or `og:title` field. We apply the same idea for arriving at values for f_d for the description fields and f_i for striking images. The card score C_s should range from 0 for a memento with no metadata at all to 1 for a memento where all metadata fields are full. To ensure this is the case, weights w_c, w_t, w_d, and w_i should all sum to 1.

$$M_q = w_k \cdot C_s + w_t \cdot M_t + w_m \cdot M_i \tag{17}$$

The rest of the DSA2 scoring function, shown in Equation 17, provides values to differentiate one memento from others if C_s is 0. Recall that Facebook introduced social card metadata standards in 2010 and there are many mementos who were captured prior to 2010. M_t returns the length of the text normalized by its Z-score. M_i returns the number of images in the page normalized by its Z-score. The Z-scores are calculated based on the mean and standard deviation of the input, so they are only relative to the input and not comparable to results for other groups of mementos. The weights w_k, w_t, and w_m can be applied to increase or decrease the importance of these terms for certain types of stories. For example, some stories may favor pages with more images, so a high w_m is best in that case.



	sample	DSA3 Algorithm	intention of steps
1	filter	exclude off-topic	off-topic mementos have low information value for summarization
2	filter	exclude near-duplicates	duplicates do not add new information
3	cluster	DBSCAN cluster collection by TF-IDF	identify outliers in collection by TF-IDF
4	filter	exclude outliers	remove items that are too different from overall collection topic
5	cluster	k-means cluster collection by memento-datetime	represent collection temporally; accounts for growth curve shape
6	cluster	k-means cluster with TF-IDF for topical clusters	represent collection topically; cluster by terms
7	score	BM25 score with top named entities as query	mementos that better match the overall collection topic
8	filter	include only highest scoring from each cluster	exemplar topical-temporal mementos that best match overall collection topic
9	order	by publication-date	common storytelling layout

Fig. 90. The DSA3 algorithm selects mementos that better match the overall collection topic.

4.3.4 EXEMPLARS MATCHING THE OVERALL COLLECTION TOPIC WITH DSA3

We created DSA3 not to improve upon issues with other algorithms, but instead to select exemplars that best match the overall collection topic. DSA3, shown in Figure 90, contains more steps than DSA2 or DSA1. The goal is to still represent the collection temporally while emphasizing its topical representation.

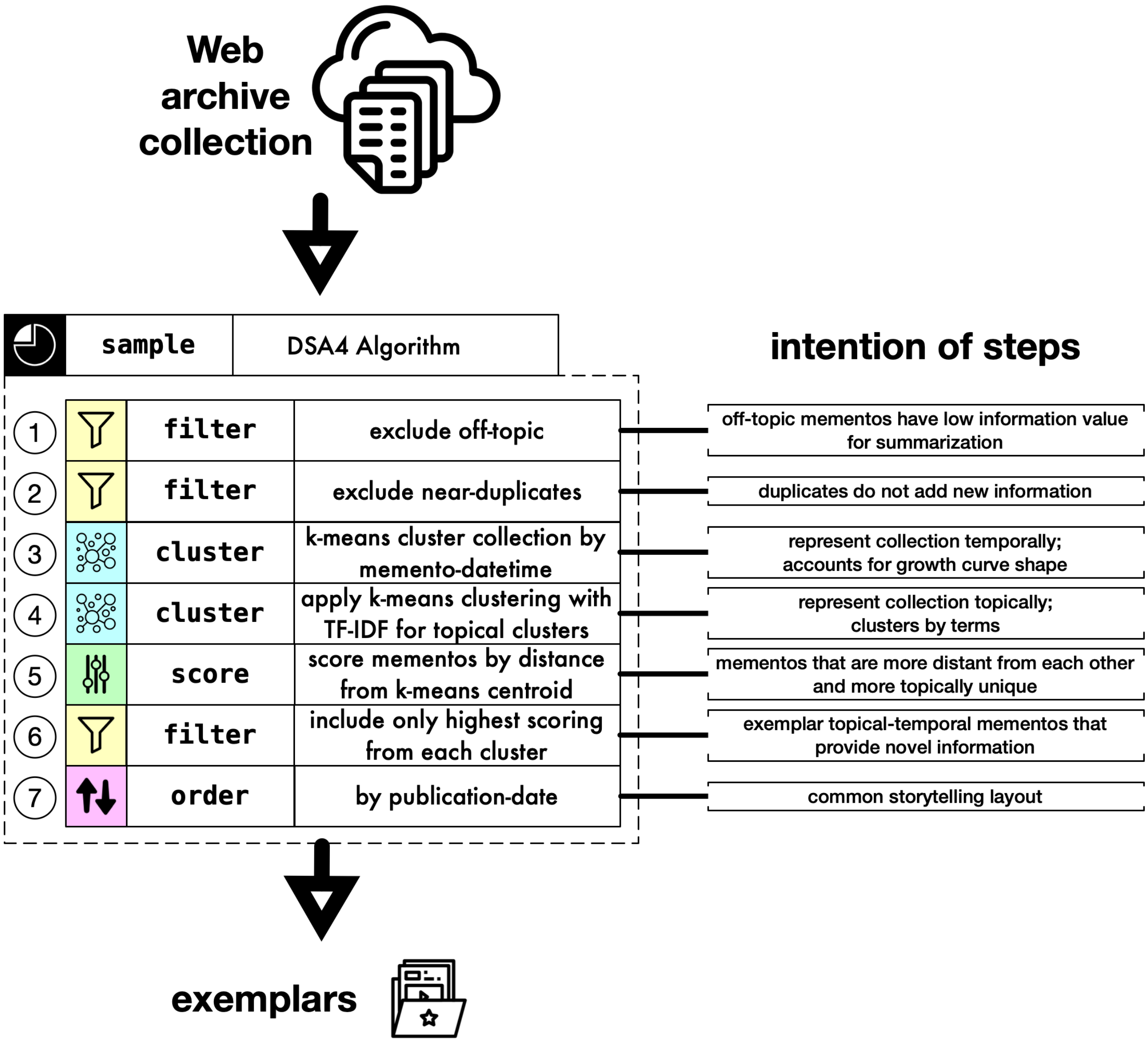
DSA3 step 3 contains an addtional clustering step that allows us to further identify topical outliers. Here we cluster by DBSCAN with TF-IDF as the feature. As noted in Chapter 2.7.1, DBSCAN establishes core points and those within a distance of ε are considered to lie within the same cluster as the core point. We empirically determined that $\varepsilon = 1.3$ gives the best balance of outliers and non-outliers. Those outside of a cluster are outliers. For small collections, values of ε lower than 1.3 cause DBSCAN to consider approximately 90% of the collection to be outliers. DSA3 step 4 filters these outliers from consideration, further ensuring that resulting steps only process documents that are central to the overall collection topic. In Steps 5 and 6, DSA3 performs temporal and topical clustering with K-means. First, we cluster by memento-datetime to ensure that the collection growth curve is considered. Second, we cluster by TF-IDF to ensure that we represent different topical aspects of the collection. We apply K-means clustering for these steps so we can control the number of clusters and the size of the story, using Equations 14 and 15 to calcluate the number of clusters.

Finally, instead of the card scoring performed in step 5, we score by first extracting the most frequent named entities from the collection. We then apply those entities as search terms to every memento in the collection and record their BM25 scores. This way the mementos that contain the most frequent named entities score highest, ensuring that we tell a story with the mementos that represent the overall collection's 5W questions (Chapter 2.7.1).

Because we do not score based on surrogate creation capability, it is possible that DSA3 exemplars may not produce good surrogates.

4.3.5 NOVEL EXEMPLARS WITH DSA4

DSA4, in Figure 91, creates a story of the more novel mementos in a collection. Step 3 clusters each memento by memento-datetime with k-means just like in DSA2 and DSA3. Step 4 clusters each memento by TF-IDF with k-means just in like DSA3. The scoring function computes the distance of each memento from its cluster centroid. A higher distance from its cluster centroid means that the memento has a lower similarity to the other mementos in its cluster, thus making it more novel.



				intention of steps
1	filter	exclude off-topic		off-topic mementos have low information value for summarization
2	filter	exclude near-duplicates		duplicates do not add new information
3	cluster	k-means cluster collection by memento-datetime		represent collection temporally; accounts for growth curve shape
4	cluster	apply k-means clustering with TF-IDF for topical clusters		represent collection topically; clusters by terms
5	score	score mementos by distance from k-means centroid		mementos that are more distant from each other and more topically unique
6	filter	include only highest scoring from each cluster		exemplar topical-temporal mementos that provide novel information
7	order	by publication-date		common storytelling layout

Fig. 91. The DSA4 algorithm selects mementos that surface more novel content.

Because we do not score based on surrogate creation capability, it is possible that DSA4 exemplars may not produce good surrogates.

4.3.6 SUMMARY

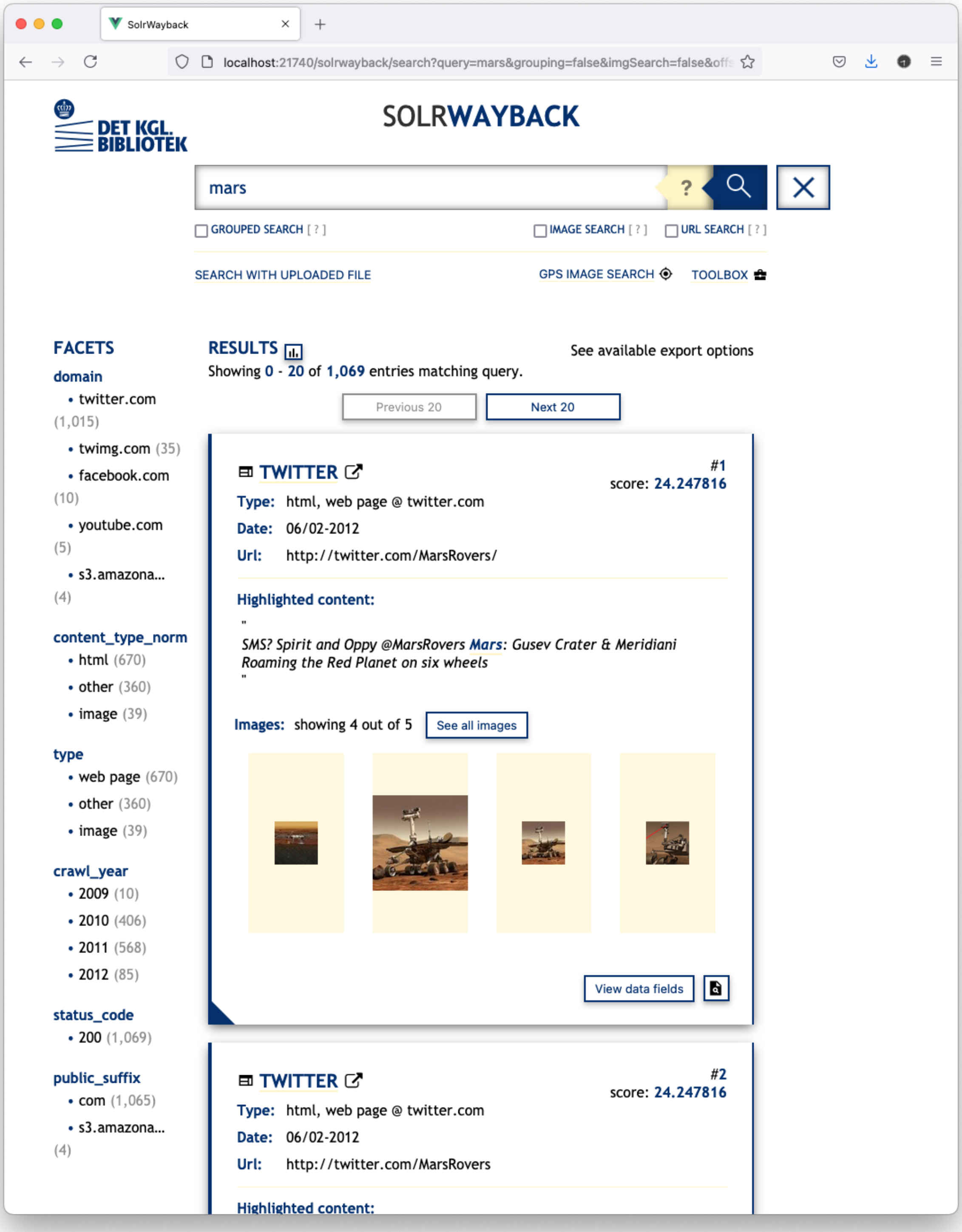
With these algorithm primitives, there are many ways to select exemplars for our stories. The sample primitive allows us to probabilistically sample from a collection, or it may be a function for a much more complex intelligent sampling algoirthm. The filter primitive reduces the input by some procedure, such as excluding off-topic mementos. The order primitive arranges the mementos in order by some feature. The score primitive runs a scoring function across all of the mementos and produces a result. The cluster primitive divides the collection into subsets based on a clustering algorithm and a given feature. Through these primitives, we can create many different algorithms for different types of stories.

Here are some examples. The filtered random algorithm filters off-topic and non-duplicate mementos from the collection and then randomly samples from the remainder. We can also order the collection by memento-datetime and then systematically sample every j^{th} memento. The DSA algorithms all combine filter, cluster, score, and order to intelligently sample collections for different purposes. The combinations are not limited to these examples. If necessary, it is even possible to create a primitive search engine from filter, score, and order.

Because search engine indices can take up as much as space as the original collection, not all web archive platforms have search engines. If present, search engines are the de facto method of exploring web archive collections. Now that we have introduced these primitives and some algorithms for selecting exemplars, how well do our algorithms produce novel mementos compared to a search engine?

4.4 EVALUATING THE DSA ALGORITHMS

The DSA algorithms select exemplars from web archive collections by processing each memento's content and metadata. A user could choose a different path. Some web archive collections, like those at Archive-It, contain search engines. People like Natasha could apply a search engine to explore a collection with queries like *when was the shooting?* or *shooting victims*. How well could someone recreate the list of exemplars produced by a DSA algorithm with a search engine? To answer this question, we conducted an experiment comparing search engine results (SERs) based on queries with the mementos selected by DSA1, DSA2, DSA3, and DSA4. Our assumption is that once the off-topic, near-duplicate, and other noise has been removed from the collection, then what remains is of higher quality and our DSA algorithms can focus on extracting



SolrWayback

localhost:21740/solrwayback/search?query=mars&grouping=false&imgSearch=false&off

DET KGL. BIBLIOTEK

SOLRWAYBACK

mars

□ GROUPED SEARCH [?]

□ IMAGE SEARCH [?]

□ URL SEARCH [?]

SEARCH WITH UPLOADED FILE

GPS IMAGE SEARCH

TOOLBOX

FACETS

domain
- twitter.com (1,015)
- twimg.com (35)
- facebook.com (10)
- youtube.com (5)
- s3.amazona... (4)

content_type_norm
- html (670)
- other (360)
- image (39)

type
- web page (670)
- other (360)
- image (39)

crawl_year
- 2009 (10)
- 2010 (406)
- 2011 (568)
- 2012 (85)

status_code
- 200 (1,069)

public_suffix
- com (1,065)
- s3.amazona... (4)

RESULTS

See available export options

Showing 0 - 20 of 1,069 entries matching query.

Previous 20 Next 20

TWITTER

#1
score: 24.247816

Type: html, web page @ twitter.com
Date: 06/02-2012
Url: http://twitter.com/MarsRovers/

Highlighted content:

"
*SMS? Spirit and Oppy @MarsRovers **Mars**: Gusev Crater & Meridiani Roaming the Red Planet on six wheels*
"

Images: showing 4 out of 5 See all images

View data fields

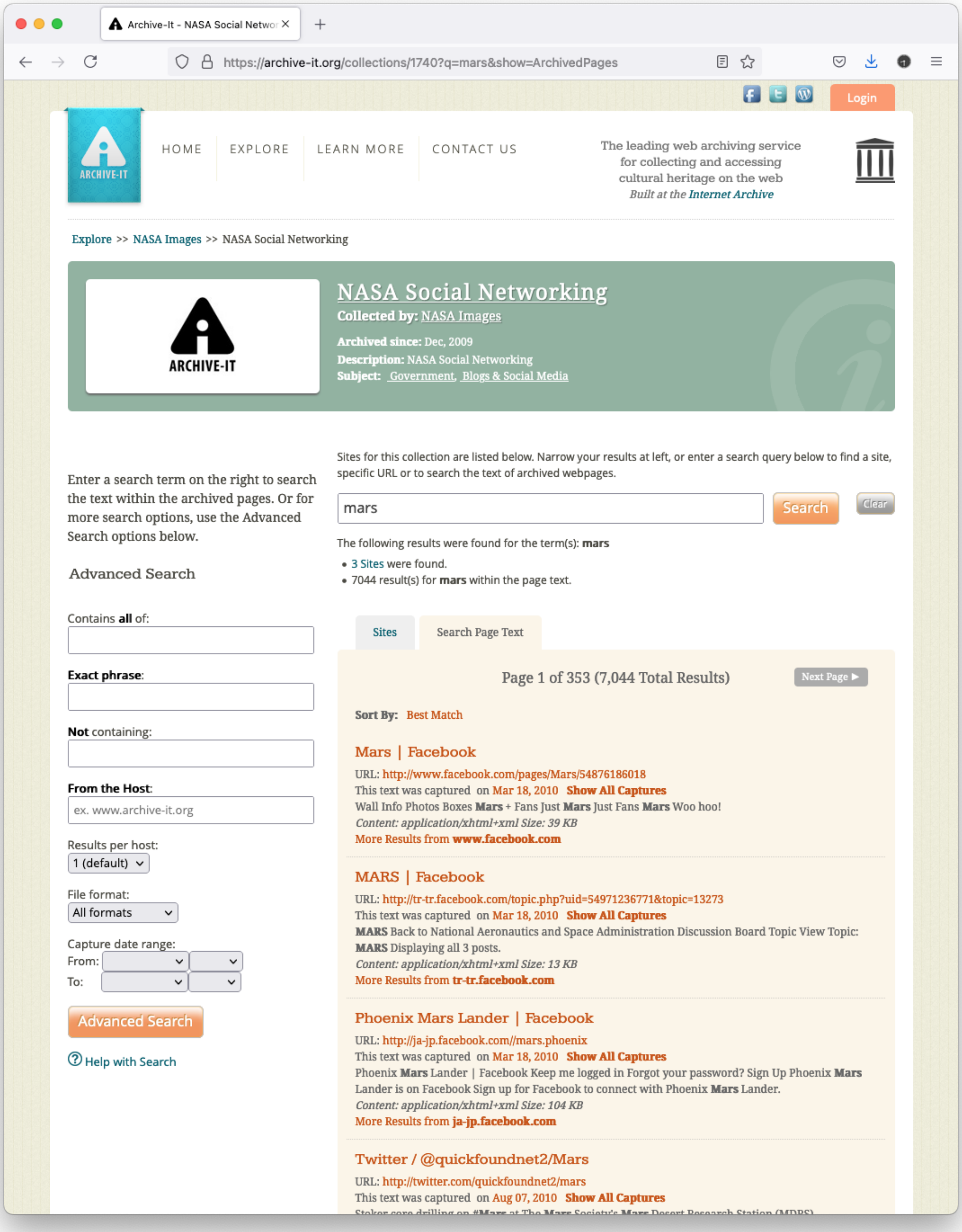
TWITTER

#2
score: 24.247816

Type: html, web page @ twitter.com
Date: 06/02-2012
Url: http://twitter.com/MarsRovers

Highlighted content:

Fig. 92. SolrWayback allows users to search a set of WARCs for terms, images, locations, and URI-Rs.


Archive-It - NASA Social Networ
https://archive-it.org/collections/1740?q=mars&show=ArchivedPages
Login
ARCHIVE-IT
HOME
EXPLORE
LEARN MORE
CONTACT US

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

Explore >> NASA Images >> NASA Social Networking

NASA Social Networking
Collected by: NASA Images
Archived since: Dec, 2009
Description: NASA Social Networking
Subject: Government, Blogs & Social Media

Enter a search term on the right to search the text within the archived pages. Or for more search options, use the Advanced Search options below.

Advanced Search

Contains **all** of:

Exact phrase:

Not containing:

From the Host: ex. www.archive-it.org

Results per host: 1 (default)

File format: All formats

Capture date range:
From:
To:

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.

mars

The following results were found for the term(s): **mars**

- 3 Sites were found.
- 7044 result(s) for **mars** within the page text.

Page 1 of 353 (7,044 Total Results)

Sort By: Best Match

Mars | Facebook
URL: http://www.facebook.com/pages/Mars/54876186018
This text was captured on Mar 18, 2010 **Show All Captures**
Wall Info Photos Boxes **Mars** + Fans Just **Mars** Just Fans **Mars** Woo hoo!
Content: application/xhtml+xml Size: 39 KB
More Results from **www.facebook.com**

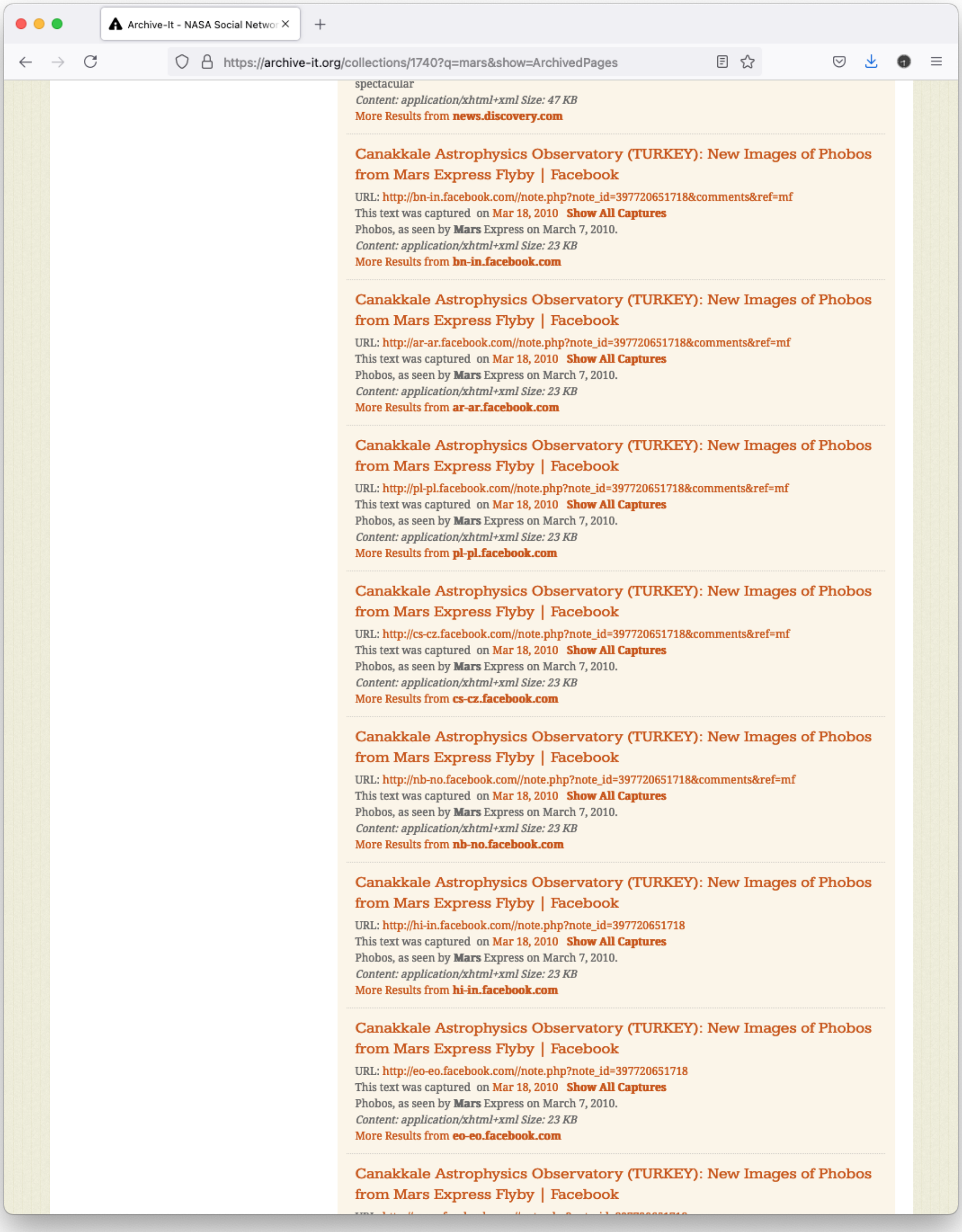
MARS | Facebook
URL: http://tr-tr.facebook.com/topic.php?uid=54971236771&topic=13273
This text was captured on Mar 18, 2010 **Show All Captures**
MARS Back to National Aeronautics and Space Administration Discussion Board Topic View Topic: **MARS** Displaying all 3 posts.
Content: application/xhtml+xml Size: 13 KB
More Results from **tr-tr.facebook.com**

Phoenix Mars Lander | Facebook
URL: http://ja-jp.facebook.com//mars.phoenix
This text was captured on Mar 18, 2010 **Show All Captures**
Phoenix **Mars** Lander | Facebook Keep me logged in Forgot your password? Sign Up Phoenix **Mars** Lander is on Facebook Sign up for Facebook to connect with Phoenix **Mars** Lander.
Content: application/xhtml+xml Size: 104 KB
More Results from **ja-jp.facebook.com**

Twitter / @quickfoundnet2/Mars
URL: http://twitter.com/quickfoundnet2/mars
This text was captured on Aug 07, 2010 **Show All Captures**
Stoker core drilling on #**Mars** at The **Mars** Society's **Mars** Desert Research Station (MDRS)

(a) Archive-It search engine

Fig. 93. Archive-It allows users to search individual collections by term or URI-R.


Archive-It - NASA Social Networ
https://archive-it.org/collections/1740?q=mars&show=ArchivedPages

spectacular
Content: application/xhtml+xml Size: 47 KB
More Results from **news.discovery.com**

Canakkale Astrophysics Observatory (TURKEY): New Images of Phobos from Mars Express Flyby | Facebook
URL: http://bn-in.facebook.com//note.php?note_id=397720651718&comments&ref=mf
This text was captured on Mar 18, 2010 **Show All Captures**
Phobos, as seen by **Mars** Express on March 7, 2010.
Content: application/xhtml+xml Size: 23 KB
More Results from **bn-in.facebook.com**

Canakkale Astrophysics Observatory (TURKEY): New Images of Phobos from Mars Express Flyby | Facebook
URL: http://ar-ar.facebook.com//note.php?note_id=397720651718&comments&ref=mf
This text was captured on Mar 18, 2010 **Show All Captures**
Phobos, as seen by **Mars** Express on March 7, 2010.
Content: application/xhtml+xml Size: 23 KB
More Results from **ar-ar.facebook.com**

Canakkale Astrophysics Observatory (TURKEY): New Images of Phobos from Mars Express Flyby | Facebook
URL: http://pl-pl.facebook.com//note.php?note_id=397720651718&comments&ref=mf
This text was captured on Mar 18, 2010 **Show All Captures**
Phobos, as seen by **Mars** Express on March 7, 2010.
Content: application/xhtml+xml Size: 23 KB
More Results from **pl-pl.facebook.com**

Canakkale Astrophysics Observatory (TURKEY): New Images of Phobos from Mars Express Flyby | Facebook
URL: http://cs-cz.facebook.com//note.php?note_id=397720651718&comments&ref=mf
This text was captured on Mar 18, 2010 **Show All Captures**
Phobos, as seen by **Mars** Express on March 7, 2010.
Content: application/xhtml+xml Size: 23 KB
More Results from **cs-cz.facebook.com**

Canakkale Astrophysics Observatory (TURKEY): New Images of Phobos from Mars Express Flyby | Facebook
URL: http://nb-no.facebook.com//note.php?note_id=397720651718&comments&ref=mf
This text was captured on Mar 18, 2010 **Show All Captures**
Phobos, as seen by **Mars** Express on March 7, 2010.
Content: application/xhtml+xml Size: 23 KB
More Results from **nb-no.facebook.com**

Canakkale Astrophysics Observatory (TURKEY): New Images of Phobos from Mars Express Flyby | Facebook
URL: http://hi-in.facebook.com//note.php?note_id=397720651718
This text was captured on Mar 18, 2010 **Show All Captures**
Phobos, as seen by **Mars** Express on March 7, 2010.
Content: application/xhtml+xml Size: 23 KB
More Results from **hi-in.facebook.com**

Canakkale Astrophysics Observatory (TURKEY): New Images of Phobos from Mars Express Flyby | Facebook
URL: http://eo-eo.facebook.com//note.php?note_id=397720651718
This text was captured on Mar 18, 2010 **Show All Captures**
Phobos, as seen by **Mars** Express on March 7, 2010.
Content: application/xhtml+xml Size: 23 KB
More Results from **eo-eo.facebook.com**

Canakkale Astrophysics Observatory (TURKEY): New Images of Phobos from Mars Express Flyby | Facebook

(a) Repeated results for the same URI-R in the Archive-It search engine

Fig. 94. (Continued) Archive-It allows users to search individual collections by term or URI-R.

novelty from this higher quality sample. We hypothesized that the DSA algorithms would surface mementos not easily discovered if a user explored the collection via a search engine. We also hypothesized that DSA3, designed to surface content more in line with the general collection topic, should have the highest retrievability and DSA4, designed to surface more novel content, should have the lowest retrievability.

SolrWayback [80] is a search engine created for web archives. It requires that the user indexes WARCs from their collection with warc-indexer [132]. With the index in place, a user can supply a search query and review the SERs, as shown in Figure 92. From here, the user has many options. They can click on the results and view the pages in SolrWayback's playback engine. They can narrow these results via facets or review the images in individual results. If necessary, they can export the search results in CSV or WARC to analyze them with another tool.

SolrWayback has not solved some of the longstanding issues with applying search engines to web archives. If multiple mementos match the search term, they are all returned, so the first n results returned often contain different mementos for the same original resource, and the next $n+m$ results contain mementos for a second original resource, and so on. This behavior is standard for search engines with web archives because we have not found a better paradigm for presenting memento results. Archive-It's search interface (Figure 93a) also has this issue (Figure 94a) with repeated results for the same URI-R.

We chose SolrWayback as our comparison search engine because it incorporates Solr, the de facto standard search engine for document collections. We did not use Archive-It's search engine for several reasons. We wanted our collection behavior to be predictable. Archive-It collections grow over time, as seen in Chapter 4.1, and we did not want a crawling event in one of Archive-It's collections to introduce unexpected behavior during our experiment. We also do not control Archive-It. It can experience technical problems and undergo upgrades, further reducing the predictability of our results. We are also only aware of the seed mementos for each collection. Searches with Archive-It can produce deep mementos that DSA is unaware of without a complete crawl. With SolrWayback, we can provide a stable environment for a fixed set of mementos.

4.4.1 METHODS

Table 17 lists the collections that we have selected for this effort. We chose collections from different semantic types, values for domain diversity, values for path depth diversity, collection lifespan, numbers of seeds and mementos, and shapes of growth curves. We also needed collections with sufficient metadata – at least a filled description field – so we could use that metadata to generate queries.

TABLE 17

Archive-It collections selected for the study comparing samples to search engines.

ID	Name	Semantic Type	Domain Diversity	Path Depth Diversity	% Query string Usage	Collection Lifespan	Number of seeds	Number of mementos	Growth Curve Shape
694	April 16 Archive	Time bounded - spontaneous	0.8391	0.0805	0.1705	48.6 weeks	88	374	seeds late, seed mementos late
1305	Elected Officials (2009-2016)	Time bounded - expected	0.0833	0.1667	0.1538	7.76 years	13	788	seeds early, seed mementos continuously
1740	NASA Social Networking	Self-archiving	0.0185	0.0277	0.1284	2.1 years	109	3211	seeds early, seed mementos continuously
2778	U.S. LGBTQ Web Collection	Subject-based	0.5136	0.0104	0.0667	7.5 years	480	4397	seeds early, seed mementos late
3470	Seattle Public Schools	Self-archiving	0.2986	0.0109	0	2.3 years	366	11491	seeds early, seed mementos late
4619	New York Climate Change Science Web Archive	Subject-based	0.1071	0.0205	0.1341	5.04 years	469	2378	seeds early, seed mementos continuously seed mementos continuously
5711	2010 Winter Olympics	Time bounded - expected	0.8272	0.0314	0.03125	22.9 weeks	192	3920	seeds late, seed mementos late
6811	Flint Water Crisis Websites Archive	Time bounded - spontaneous	0.1991	0.4425	0.2247	4.5 years	227	977	seeds early, seed mementos early

We synthesized WARCs from the seed mementos accessible from each of these Archive-It collections. We generated these WARCs so they would resemble the content of the original crawls performed by Archive-It. We ensured that the memento-datetimes and URI-Rs present in the WARCs matched those from their mementos. We also used the raw memento content for each WARC so that Archive-It branding and link rewriting would not be indexed by SolrWayback, tainting our results.

To ensure reproducibility, we were careful not to alter the default behavior of SolrWayback. Even though Solr can index multiple collections, by default SolrWayback only indexes a single set of WARCs. To isolate collections from each other, we created eight instances of SolrWayback, one for each collection.

As noted in Chapter 2.7.1, we can generate queries based on documents. We automatically generated queries for each exemplar memento selected by DSA1, DSA2, DSA3, and DSA4. We selected query generation techniques that would allow us to demonstrate the spectrum of searching user behavior.

Our first query generation technique, *doc2query-T5* [236], processes the text in a document and generates questions based on the text to simulate a naïve user searching for the topic of the document. With doc2query-T5, the queries emulate the behavior of a user asking questions of a search engine. Example queries generated by doc2query-T5 include “what is the closest asteroid to the earth” and “how many known exoplanets are there?” To use doc2query-T5, we supplied a memento stripped of its boilerplate and indicated how many queries we wanted it to generate from that content. Occasionally, doc2query-T5 generates nonsensical queries, like “what area of earth is rhea on” because it does not have background knowledge of the subject of the document. This query is nonsensical because Rhea is a moon of Saturn and hence cannot be “on” Earth. These nonsensical queries are in the minority and may simulate a user with little education on the collection’s topic. To ensure that one nonsensical query does not heavily influence the results, we use each memento’s top five queries.

We next generated queries from the top five named entities from each memento. Recall from Chapter 2.7.1 that named entities answer the 5W questions like *who?*, *when?*, and *where?* We applied spaCy [121] to extract these named entities from a memento. Then we calculated the frequency at which each occurred and selected the top five. This method simulates a more adept search user looking for information by applying pre-existing topical knowledge. We applied named entity query sizes of 1, 2, and 5 terms. A single term simulates a user looking for a particular topic. Two terms is the standard query size found in the search engine logs used by Traub et al. [300] in an experiment on retrievability. A query of five terms simulates a power user

TABLE 18

Example queries generated from the document at URI-M `https://wayback.archive-it.org/694/20070523182123/http://www.washingtonpost.com/wp-srv/metro/vatechshootings/`

Query Type	**Example**
doc2queryT5	“virginia tech killer ammo” “what kind of ammo did seung hui cho have” “how many rounds was cho ammo?” “who killed at virginia tech” “who was the gunman who killed so many people?”
top entities	virginia tech va. tim craig seung hui cho norris hall
lexical signature	killer officials disclosed craig conclude
title	Virginia Tech Shootings (washingtonpost.com)

who has a particular topical interest.

We then used lexical signatures as a query method because they are designed to return specific documents. Lexical signatures contain the terms with the highest TF-IDF scores from the document [265]. We calculated IDF scores based on term usage from the entire collection. As with named entities, we applied lexical signature sizes of 1, 2, and 5 terms. A query with 1 or 2 terms simulates a user looking for a particular topic, much like with named entities. A lexical signature of 5 terms simulates a user who has no identifier for a document but is seeking a document for which they recall details. Phelps and Wilensky [265] referred to lexical signatures as "robust hyperlinks" because they would allow live web search engine users to find the same document at a different URI. A 5-term lexical signature is an optimum size identified by Phelps and Wilensky, and later Park et al. [258, 265] to successfully retrieve specific documents from the live web.

Finally, we employed titles as a fourth query type because they work so well for finding specific documents on the live web [182]. These queries simulate users searching for a known item [211]. In this case, the user is aware of the document by its name and uses that as an identifier. Titles are very good at returning the same document on the live web but at a different URI.

For each collection, we generated a story of k mementos with each DSA algorithm. From each of these k mementos, we generated a query using one of the query methods highlighted above. We generated an additional query by using the collection metadata as input to the same query method. This last query simulates a user who was inspired to form a query after reading the metadata. For all query types but doc2query-T5, we have $\mathfrak{Q} = k + 1$ queries per story. For doc2query-T5, we chose the 5 top queries for each memento, so we have $\mathfrak{Q} = 5 \cdot (k + 1)$ queries per story. Table 18 shows a set of example queries generated from a memento found in the collection 694 about the Virginia Tech Massacre.

Then we evaluated how well these queries returned SERs containing the same mementos as each story. To measure their success, we applied Azzopardi's retrievability metric introduced in Chapter 2.7.1. This metric counts the number of queries that successfully return the expected document using the search engine. It requires a cutoff value c to indicate how many search results the user viewed. Thus, we discover if a user found a story memento within c search results for each set of queries representing that story.

A DSA algorithm that achieves a retrievability score close to 1.0 is selecting the same exemplars users could expect to find by querying a search engine. A DSA algorithm with a low retrievability score demonstrates that it is capable of surfacing mementos that a user cannot easily find through an existing search approach.

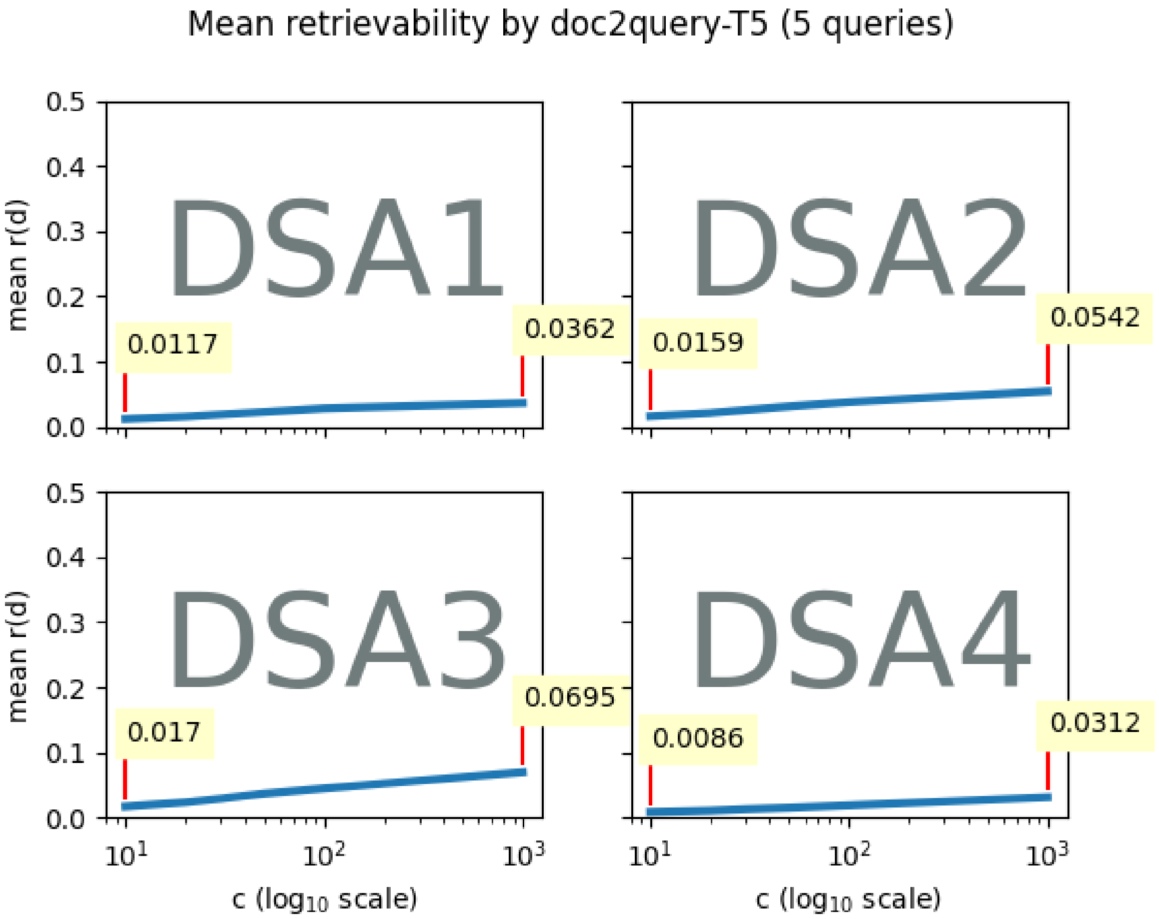


Fig. 95. Mean retrievability per cutoff value by query method: doc2query-T5 (5 queries).

TABLE 19

p-values at $c = 10$ ($p < 0.05$ in bold) by query method: doc2query-T5 (5 queries).

	DSA1	DSA2	DSA3	DSA4
DSA1	1.00000	0.19280	0.13230	0.23783
DSA2	0.19280	1.00000	0.74984	**0.01077**
DSA3	0.13230	0.74984	1.00000	**0.00676**
DSA4	0.23783	**0.01077**	**0.00676**	1.00000

4.4.2 RETRIEVABILITY DISTRIBUTIONS OF DIFFERENT DSA ALGORITHMS

Figures 95 through 118 demonstrate different ways of looking at the retrievability scores for each query method.

Our doc2query-T5 queries simulate a casual user. In Figure 95, we see very little change in the mean retrievability of documents from each DSA algorithm with queries generated by doc2query-T5. The x-axis is scaled by $\log_{10}$, so we can see the trend as we accept more search results moving from left to right. At $c = 10^1 = 10$, we have the standard user experience of evaluating "ten blue links," or the first page of search results. At $c = 10^3 = 1000$, we simulate a user performing an intensive search, potentially to build their own collection of documents. The y-axis contains the mean normalized retrievability for a memento ($r(d)$). The mean retrievability starts at a very low value for all algorithms. The algorithm with the lowest retrievability is DSA4 with a mean retrievability of 0.00086 for $c = 10$ and 0.0312 for $c = 10^3$. As mentioned in Chapter 4.3.5 we expect DSA4 to return more novel content. We expect DSA3 to return content describing the main topic of the collection. The scores are so poor that it is hard to determine by this data alone that DSA3 and DSA4 succeed in their intentions.

Figure 19 contains a table listing the p-values for each algorithm at $c = 10$, based on SciPy's [316] implementation of Student's t-test. Those in bold are values where $p < 0.05$, indicating that the differences in retrievability between these DSA algorithms for this query method are statistically significant. We see that our evaluations of DSA2 vs DSA4 and DSA3 vs DSA4 are statistically significant for doc2query-T5. We conclude that DSA2 and DSA3 have higher retrievability values with doc2query-T5 than DSA4 at $c = 10$, and thus DSA4 produces more novel content than either of these algorithms. These results are encouraging because they match our goals for DSA4, but we have to remember that the results represent only one of many query methods.

Figure 96 shows a histogram of each algorithm's retrievability at $c = 10$. Here we note that none of the DSA algorithms break past a maximum retrievability of 0.2. DSA4's 147 documents across all stories only reach 0.1277. At $c = 1000$ (Figure 97), the maximum retrievability scores only reach as high as 0.862 with DSA3 and as low as 0.5319 with DSA4. A user with these casual queries may be able to reassemble a significant portion of a story build by DSA3, but there are still items out of their reach.

Our entity queries simulate users trying to understand a topic through the 5W questions. Figure 98 shows better mean retrievability scores than for doc2query, with DSA3 reaching a mean of 0.2226 retrievability by $c = 1000$. Comparisons applying the top entity as a search term are not

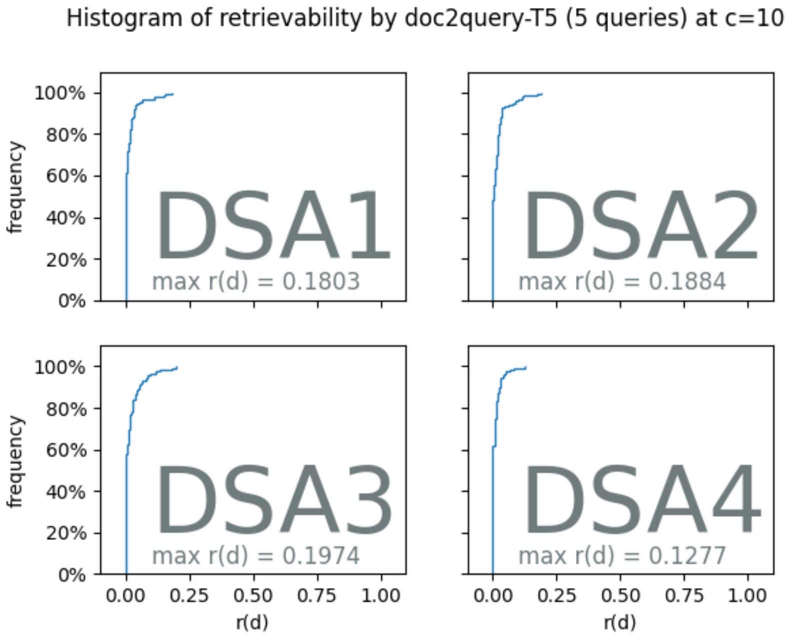


Fig. 96. Histogram at $c = 10$ by query method: doc2query-T5 (5 queries).

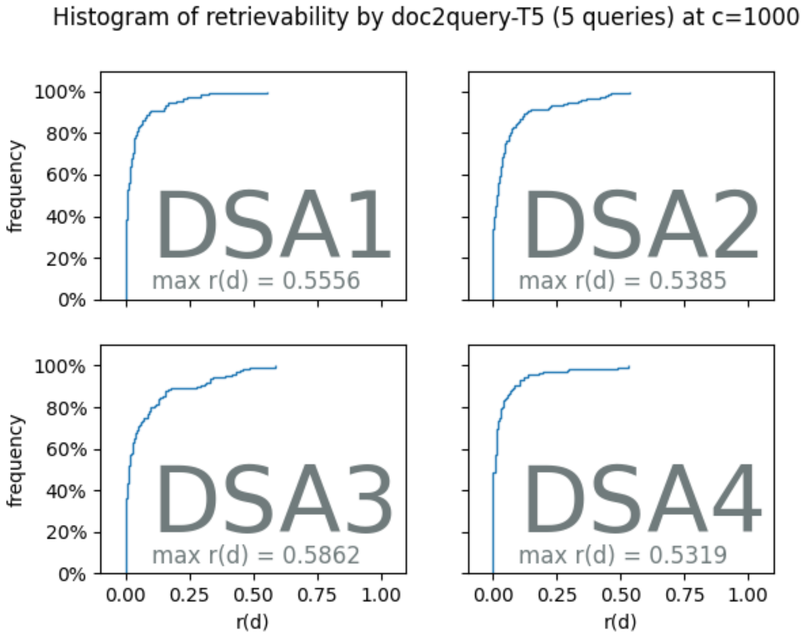


Fig. 97. Histogram at $c = 1000$ by query method: doc2query-T5 (5 queries).

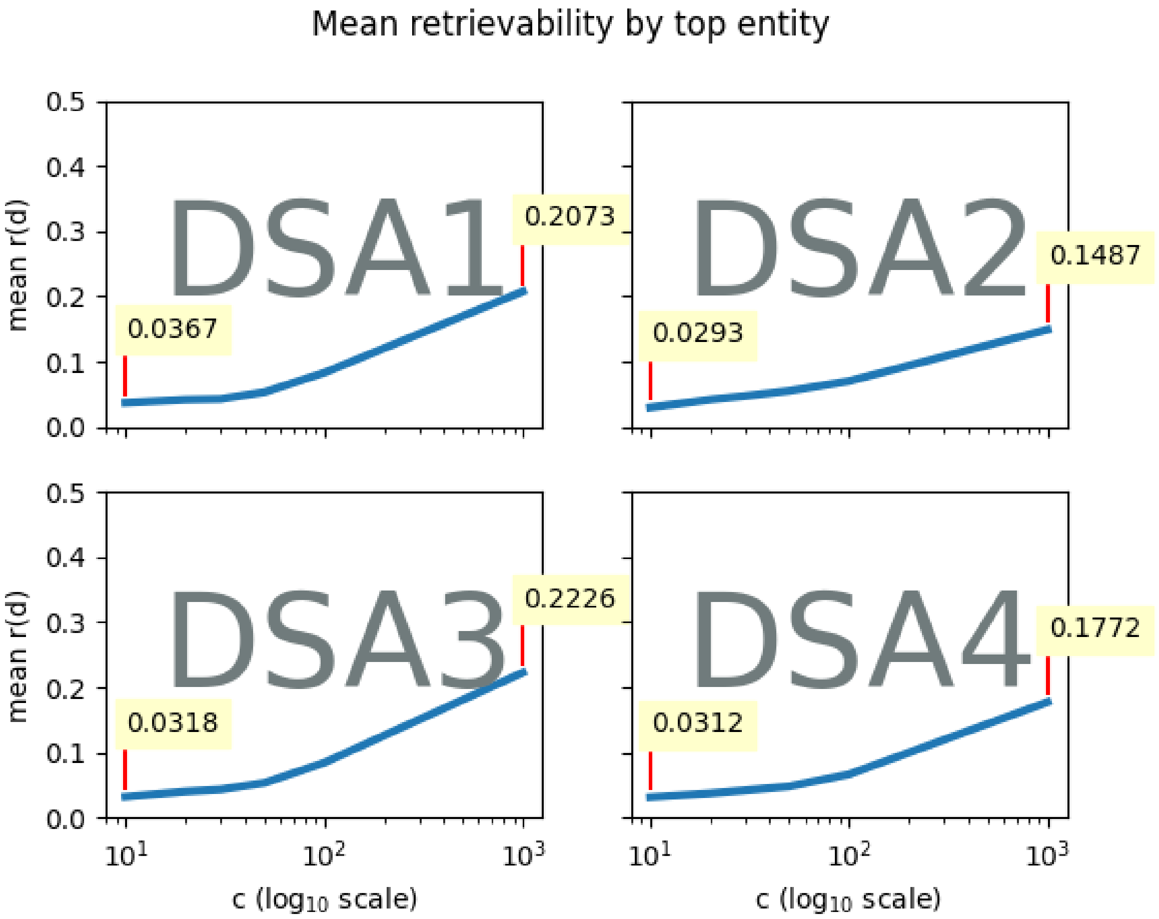


Fig. 98. Mean retrievability per cutoff value by query method: top entity.

TABLE 20

p-values at $c = 10$ ($p < 0.05$ in bold) by query method: top entity.

	DSA1	DSA2	DSA3	DSA4
DSA1	1.00000	0.17943	0.43782	0.36442
DSA2	0.17943	1.00000	0.60295	0.69013
DSA3	0.43782	0.60295	1.00000	0.90070
DSA4	0.36442	0.69013	0.90070	1.00000

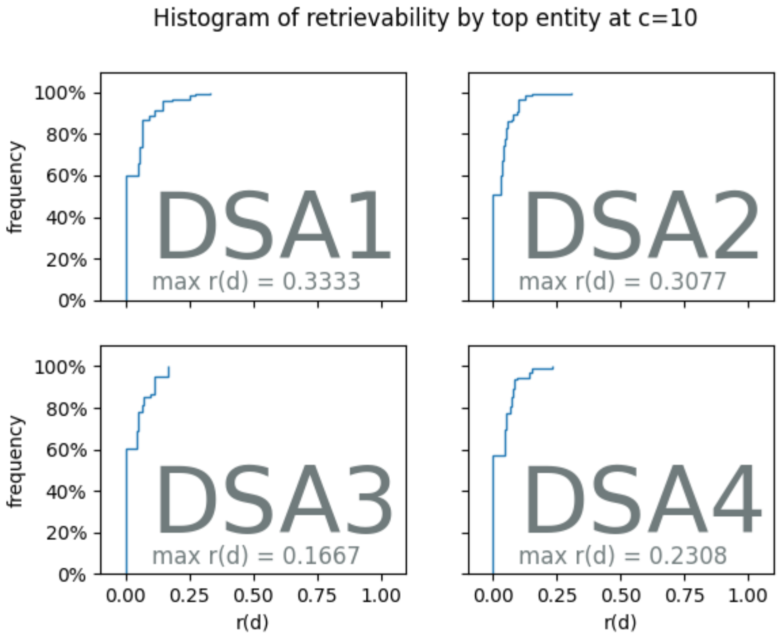


Fig. 99. Histogram at $c = 10$ by query method: top entity.

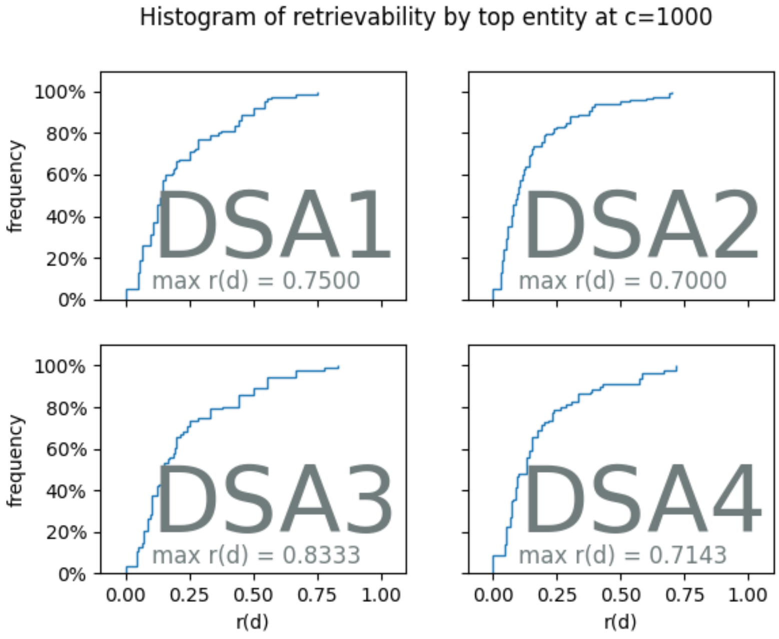


Fig. 100. Histogram at $c = 1000$ by query method: top entity.

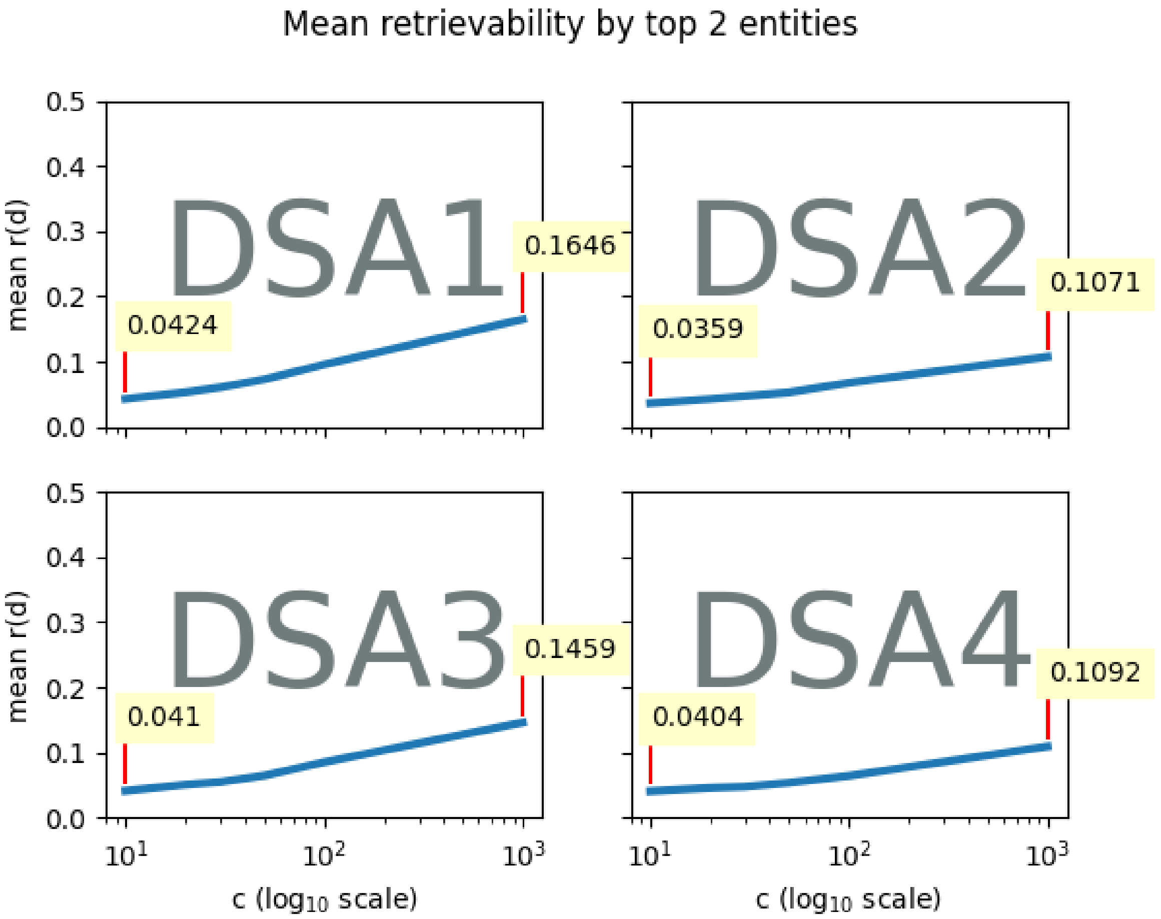


Fig. 101. Mean retrievability per cutoff value by query method: top 2 entities.

TABLE 21

p-values at $c = 10$ ($p < 0.05$ in bold) by query method: top 2 entities.

	DSA1	DSA2	DSA3	DSA4
DSA1	1.00000	0.16798	0.78643	0.69450
DSA2	0.16798	1.00000	0.26856	0.32861
DSA3	0.78643	0.26856	1.00000	0.90140
DSA4	0.69450	0.32861	0.90140	1.00000

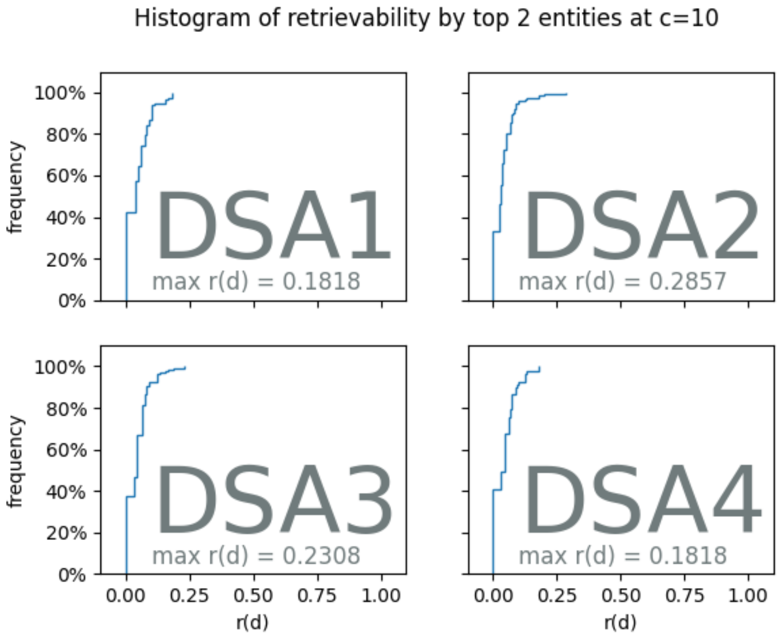


Fig. 102. Histogram at $c = 10$ by query method: top 2 entities.

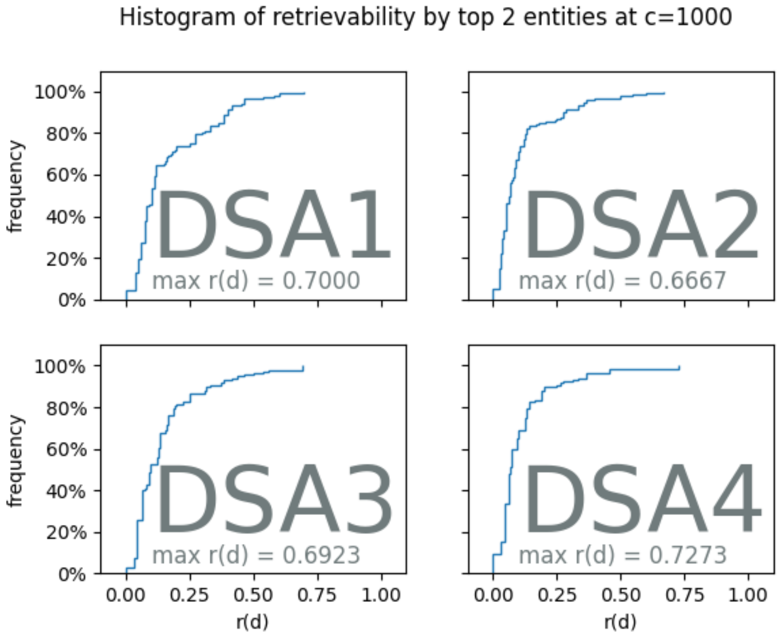


Fig. 103. Histogram at $c = 1000$ by query method: top 2 entities.

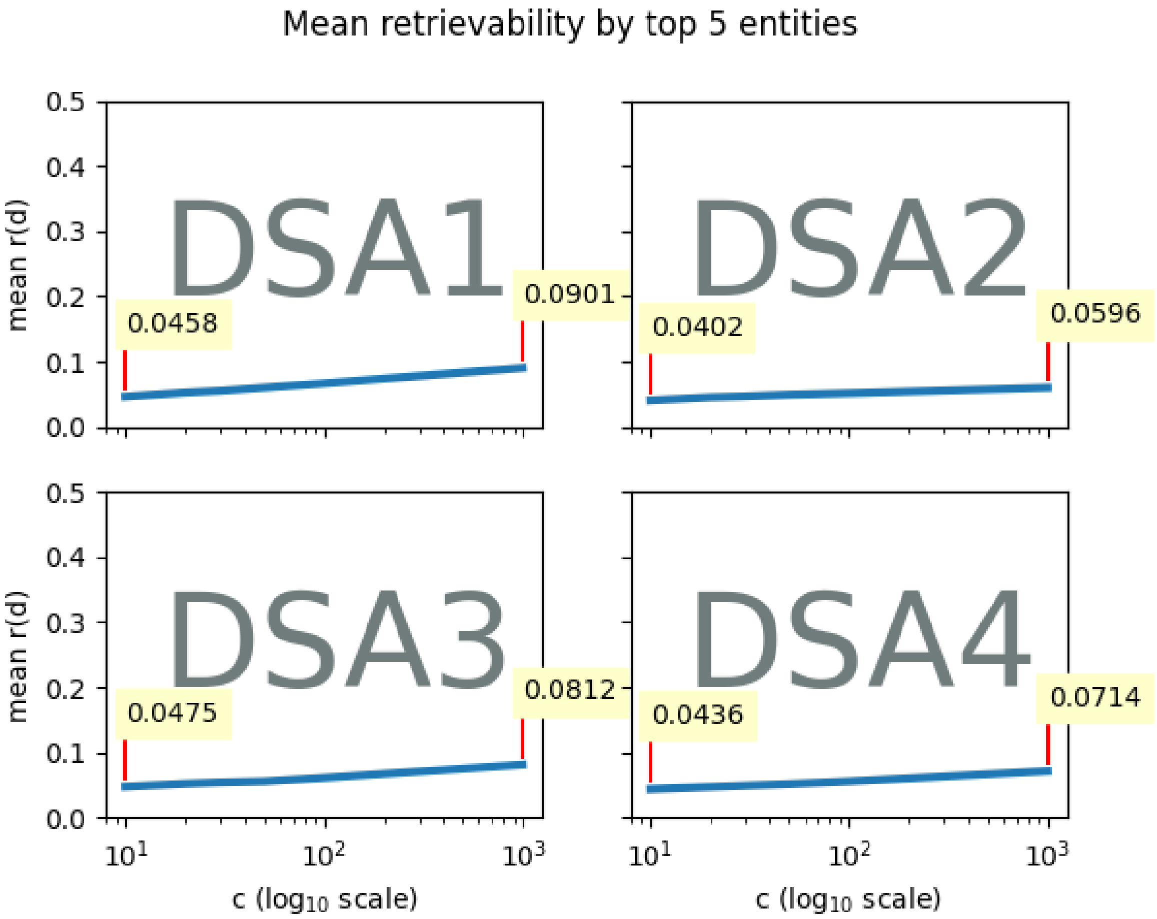


Fig. 104. Mean retrievability per cutoff value by query method: top 5 entities.

TABLE 22

p-values at $c = 10$ ($p < 0.05$ in bold) by query method: top 5 entities.

	DSA1	DSA2	DSA3	DSA4
DSA1	1.00000	0.14967	0.68819	0.63276
DSA2	0.14967	1.00000	**0.03946**	0.34073
DSA3	0.68819	**0.03946**	1.00000	0.34284
DSA4	0.63276	0.34073	0.34284	1.00000

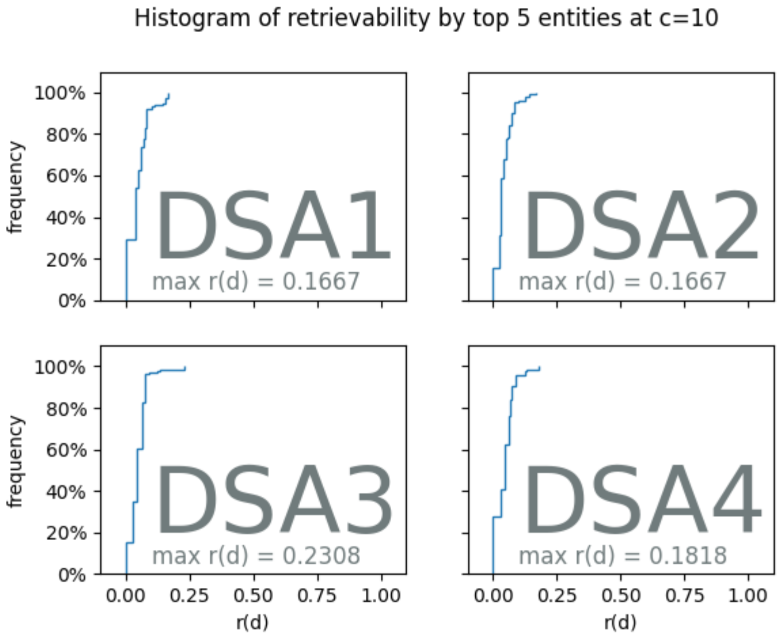


Fig. 105. Histogram at $c = 10$ by query method: top 5 entities.

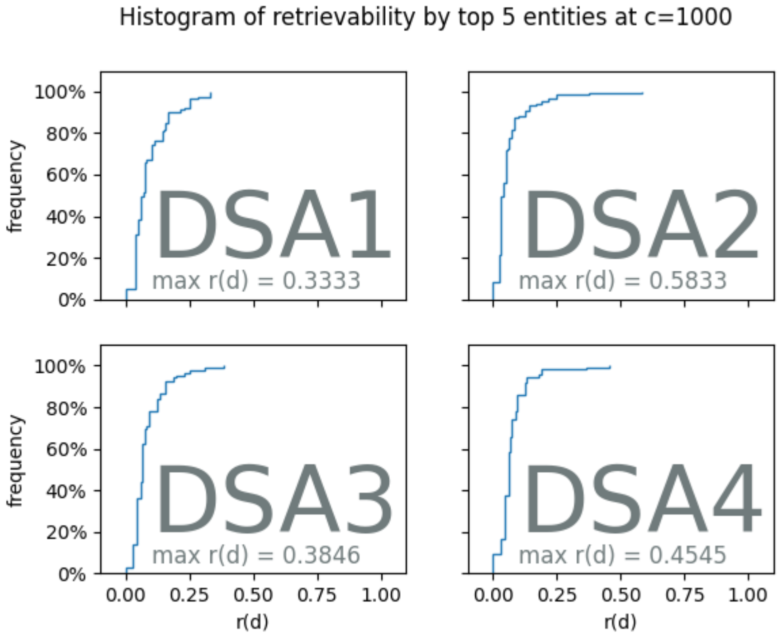


Fig. 106. Histogram at $c = 1000$ by query method: top 5 entities.

statistically significant between DSA algorithms (Figure 20). At $c = 10$ the maximum retrievability is 0.3333 for DSA1 (Figure 99). The minimum is DSA3 with 0.1667. Figure 100 shows that DSA3 reaches a maximum retrievability of 0.8333 $c = 1000$ with DSA1 in second place at 0.7500 when $c = 1000$.

Recall that this is the top entity from each memento in a collection. As noted in Chapter 4.2, collections are typically built upon a central theme. Using this one entity as a query likely unlocks most of the collection, and one can rebuild much – but not all – of all DSA stories if one performs an intensive search of the collection. These results for DSA3 at $c = 1000$ are statistically significant when compared with DSA2 ($p = 0.00010$) and DSA4 ($p = 0.03326$). Recall that Figure 20 does not show these values because it only lists p-values for $c = 10$. These results provide evidence that DSA3 may be meeting its goal of exposing mementos that best match the collection's topic, at least when compared with DSA2's and DSA4's performance. In spite of this, we have not reached a case where a single memento has $r(d) = 1.0$.

As noted by Traub et al. [300], users most often issue 2-term queries, as reflected in the results in Figures 101 through 103. The mean retrievability scores (Figure 101) are lower than those achieved when we applied a single entity as a search term. The comparisons between DSA algorithms are not statistically significant at $c = 10$ (Figure 21). The distributions for DSA2 at $c = 10$ in Figure 102 reaches the maximum retrievability of 0.2857 which is close to its maximum at $c = 10$ with only the top entity. DSA3 comes in second place and DSA1 and DSA4 tie with a maximum retrievability of 0.1818. By $c = 1000$, however, seen in Figure 103, DSA4 has the highest maxium retrievability at 0.7273 compared to DSA3's 0.6923. These results are statistically significant at $c = 1000$ for DSA3 and DSA4 and detract from the idea that DSA3 or DSA4 are fulfilling their intentions successfully.

Results with queries of five entities are represented by Figures 104 through 106. These perform poorer, as seen in the mean retrievability visualized by Figure 104. Comparisons between DSA2 and DSA3 are statistically significant at $c = 10$ (Figure 22). Thus, on average, DSA3's mementos are slighlty more retrievable than those from DSA2 at $c = 10$. This statistical significance holds at $c = 1000$ for DSA2–DSA3 ($p = 0.00351$). Also, at $c = 1000$ the results for DSA1–DSA4 are significant ($p = 0.01381$) with DSA1 having a higher mean retrievability of 0.0901. DSA3's maximum retrievability at $c = 10$ only reaches 0.2308 (Figure 105), but that is better than the others. A memento chosen by DSA2 has the highest retrievability of 0.5833 at $c = 1000$ (Figure 106).

Several reasons exist for why the 5-entity queries did not perform as well as the top entity queries. At first, we anticipated that a 5-entity query would perform better than the others because this method produces queries that provide a specific set of terms to identify the document. However, recall that the means shown in Figures 98, 101, and 104 are means across all entity queries issued for a document. If we issue our $k+1$ 1-entity queries, the top entity is likely common among most queries generated from the mementos in that story. This commonality raises the mean retrievability of all story mementos. By increasing the number of entities, we make each query more specific to a memento. If only one query in our $k+1$ returned the expected story memento, this result would drive down the mean retrievability. The mean number of story mementos is 21.00. Thus $\bar{k} = 21.00$ and $\bar{k}+1 = 22.00$. If, on average, for a set of $k+1$ 5-entity queries only one query returns the correct document, then $\frac{1}{\bar{k}} = \frac{1}{22.00} = 0.04545$. This is close to the mean retrievability values shown in Figure 104 at $c = 10$. It improves by $c = 1000$ but not nearly as much as it does with one or two entity queries.

Lexical signatures allow more types of terms than entities. In Figure 107 we see slightly better mean retrievability scores for the top term from a lexical signature compared with those from a top entity (Figure 98). The top term from the lexical signature is statistically significant (Figure 20) for comparisons between DSA1–DSA2 and DSA2–DSA4; thus, at $c = 10$, DSA1 produces mementos with slightly better retrievability than DSA2 and DSA4 produces mementos with slightly better retrievability than DSA2. Our histogram for the top term from a lexical signature at $c = 10$ (Figures 99) shows that a memento from DSA2 reaches the highest maximum retrievability at 0.2143 compared to DSA1's 0.1818. At 0.7500, DSA1 reaches the highest retrievability for a single memento at $c = 1000$ (Figure 100) but this is only statistically significant when compared with DSA4 ($p = 0.00006$). DSA4, however, has the lowest maximum retrievability for a single memento at 0.4000 and this is statistically significant when compared to DSA1 ($p = 0.00006$), DSA2 ($p = 0.00337$), and DSA3 ($p = 0.00000$). This supports that DSA4 is performing as intended.

Applying the top two terms of a lexical signature (Figure 110) also produces a higher mean for DSA1 at $c = 10$. Per Figure 24, this result is statistically significant only when compared with DSA2. The maximum retrievability of 0.3077 reached by DSA1 for $c = 10$ (Figure 111) is higher than the other algorithms. By $c = 1000$, a memento selected by DSA1 reaches the highest retrievability of all algorithms with this query type, but this is only statistically significant compared with DSA2 ($p = 0.02194$ and DSA4 ($p = 0.00000$). DSA4's low performance, however, is statistically significant when compared to all other algorithms ($p = 0.00000$ for DSA1, $p = 0.01236$ for DSA2, $p = 0.00000$ for DSA3), giving us more evidence that DSA4 may be fulfilling its goal.

Five term lexical signatures are those recommended by Philips and Wilensky [265] and Park et

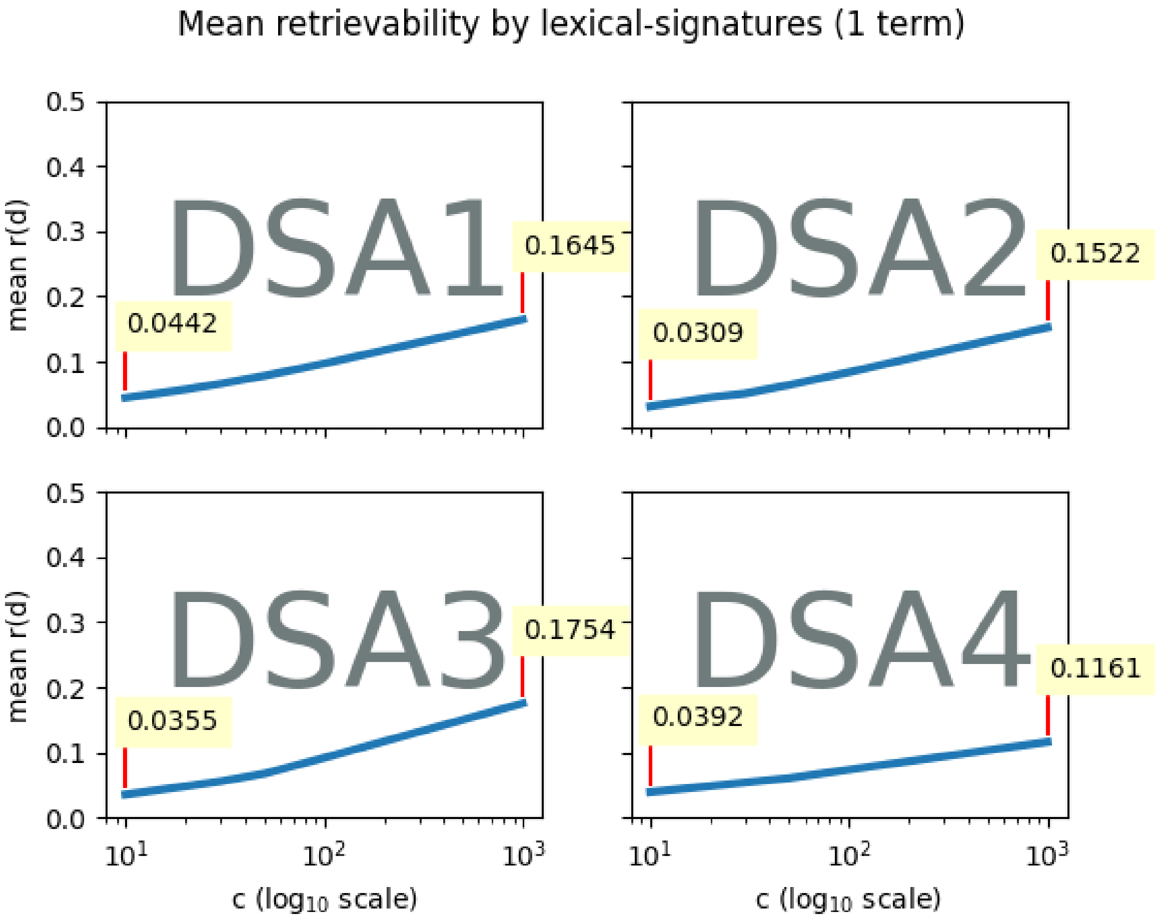


Fig. 107. Mean retrievability per cutoff value by query method: lexical-signatures (1 term).

TABLE 23

p-values at $c = 10$ ($p < 0.05$ in bold) by query method: lexical-signatures (1 term).

	DSA1	DSA2	DSA3	DSA4
DSA1	1.00000	**0.00286**	0.09466	0.30027
DSA2	**0.00286**	1.00000	0.29204	**0.03550**
DSA3	0.09466	0.29204	1.00000	0.42379
DSA4	0.30027	**0.03550**	0.42379	1.00000

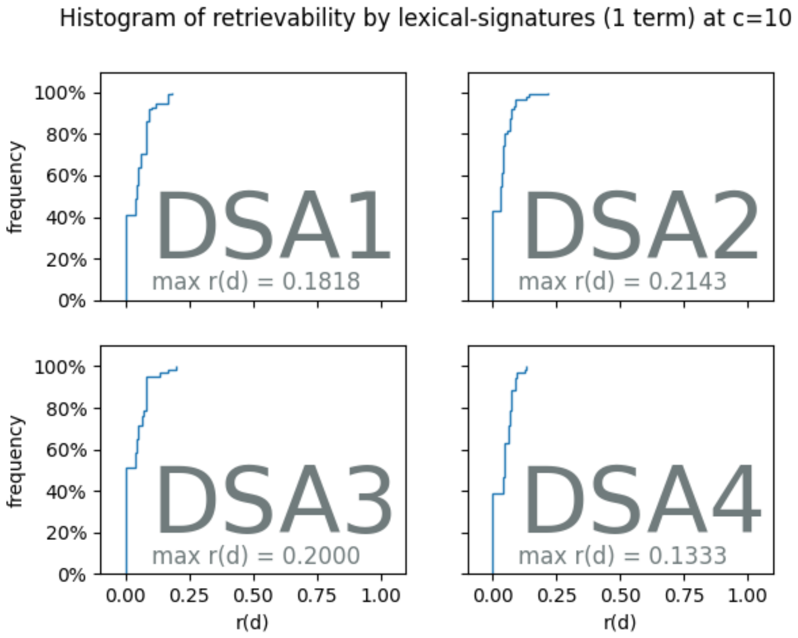


Fig. 108. Histogram at $c = 10$ by query method: lexical-signatures (1 term).

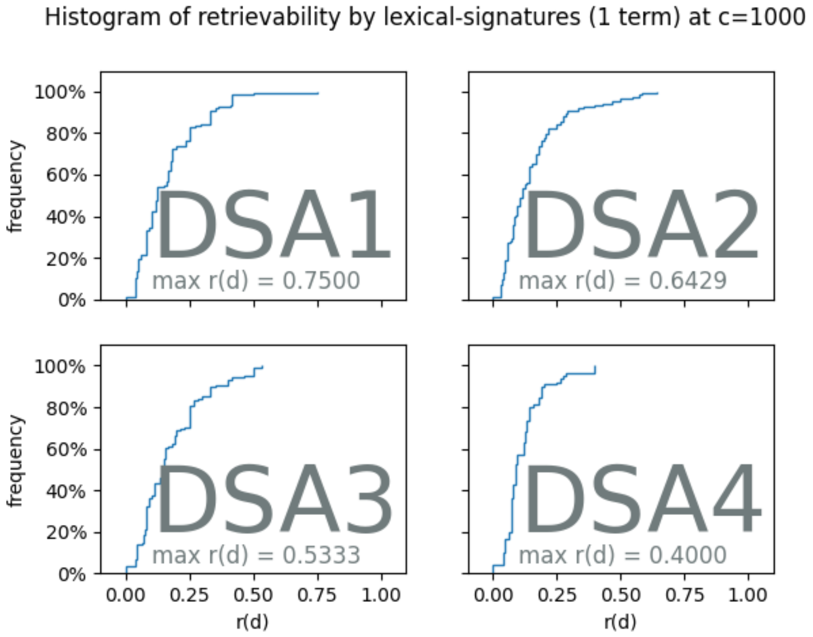


Fig. 109. Histogram at $c = 1000$ by query method: lexical-signatures (1 term).

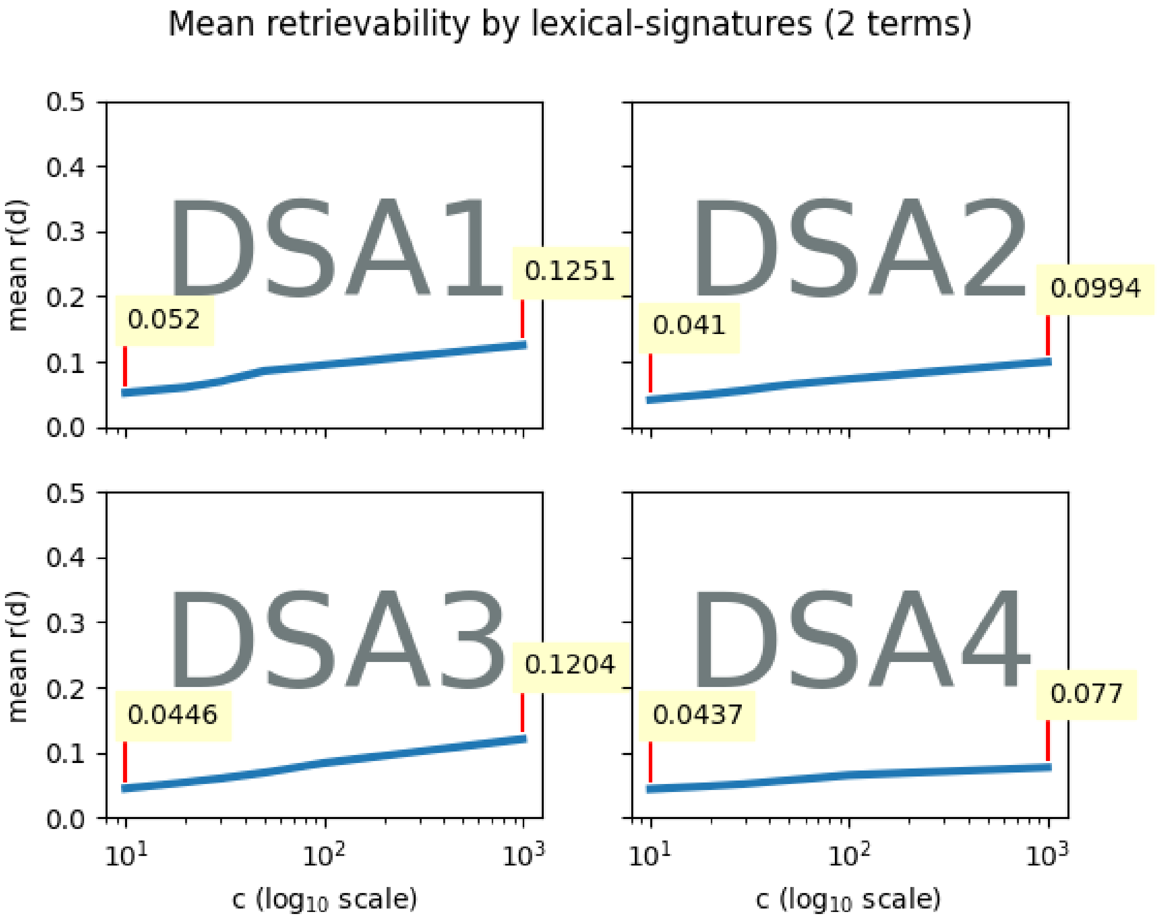


Fig. 110. Mean retrievability per cutoff value by query method: lexical-signatures (2 terms).

TABLE 24

p-values at $c = 10$ ($p < 0.05$ in bold) by query method: lexical-signatures (2 terms).

	DSA1	DSA2	DSA3	DSA4
DSA1	1.00000	**0.02998**	0.19632	0.12272
DSA2	**0.02998**	1.00000	0.38415	0.48701
DSA3	0.19632	0.38415	1.00000	0.82062
DSA4	0.12272	0.48701	0.82062	1.00000

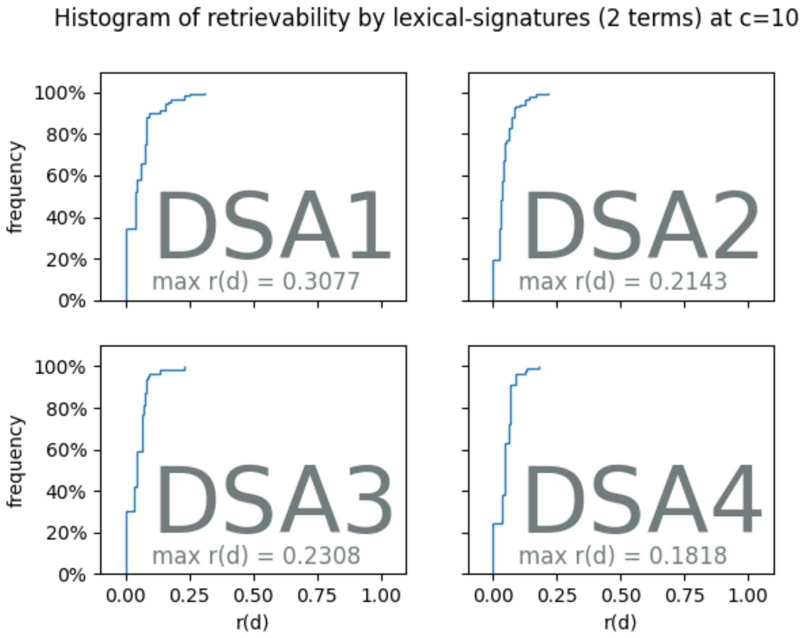


Fig. 111. Histogram at $c = 10$ by query method: lexical-signatures (2 terms).

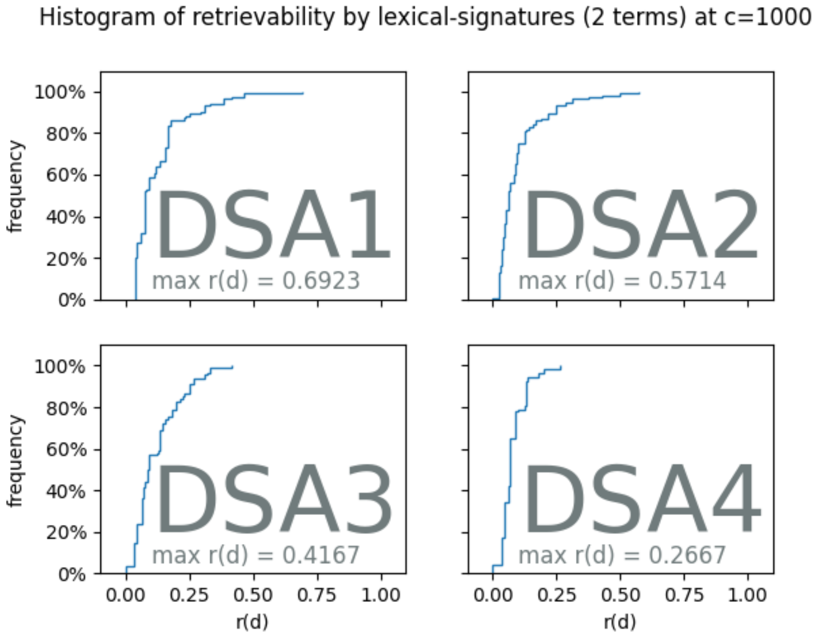


Fig. 112. Histogram at $c = 1000$ by query method: lexical-signatures (2 terms).

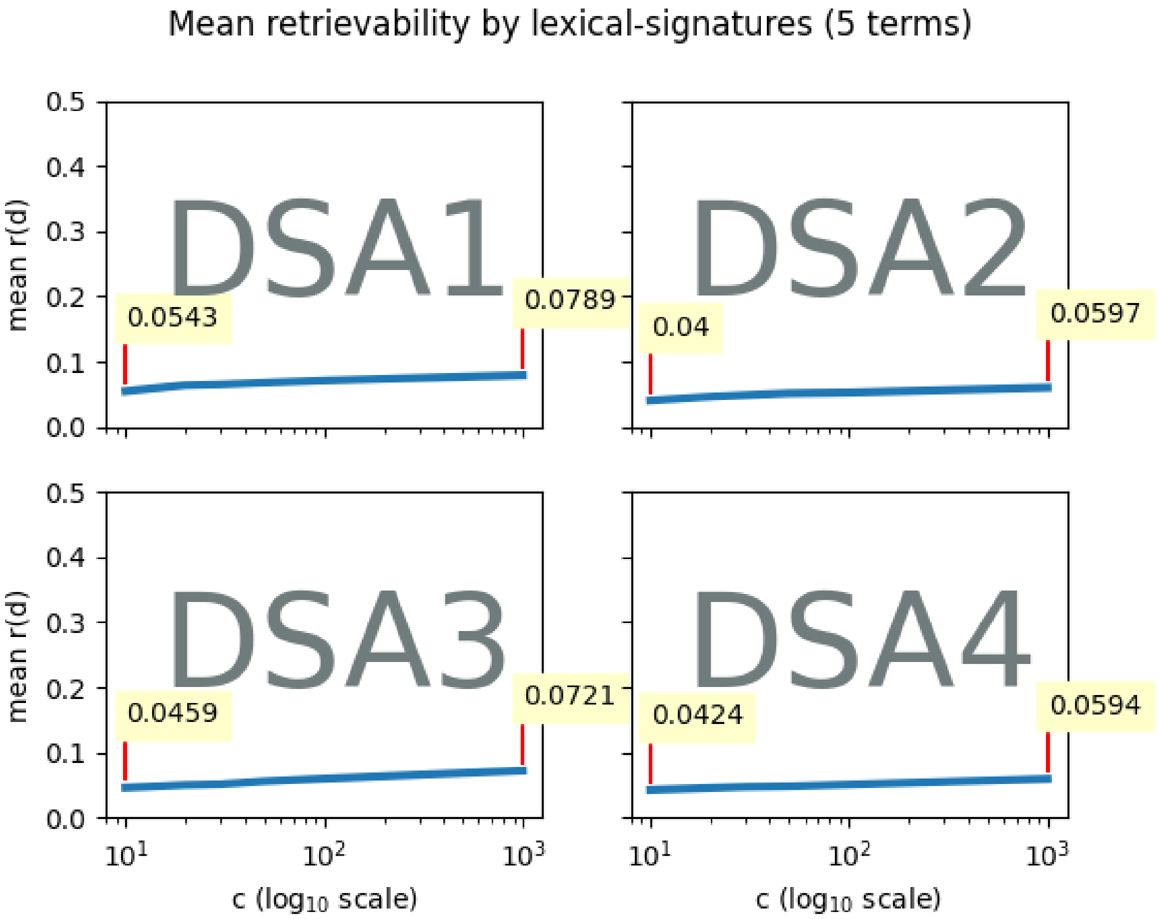


Fig. 113. Mean retrievability per cutoff value by query method: lexical-signatures (5 terms).

TABLE 25

p-values at $c = 10$ ($p < 0.05$ in bold) by query method: lexical-signatures (5 terms).

	DSA1	DSA2	DSA3	DSA4
DSA1	1.00000	**0.00361**	0.13736	**0.02836**
DSA2	**0.00361**	1.00000	0.14136	0.52322
DSA3	0.13736	0.14136	1.00000	0.42165
DSA4	**0.02836**	0.52322	0.42165	1.00000

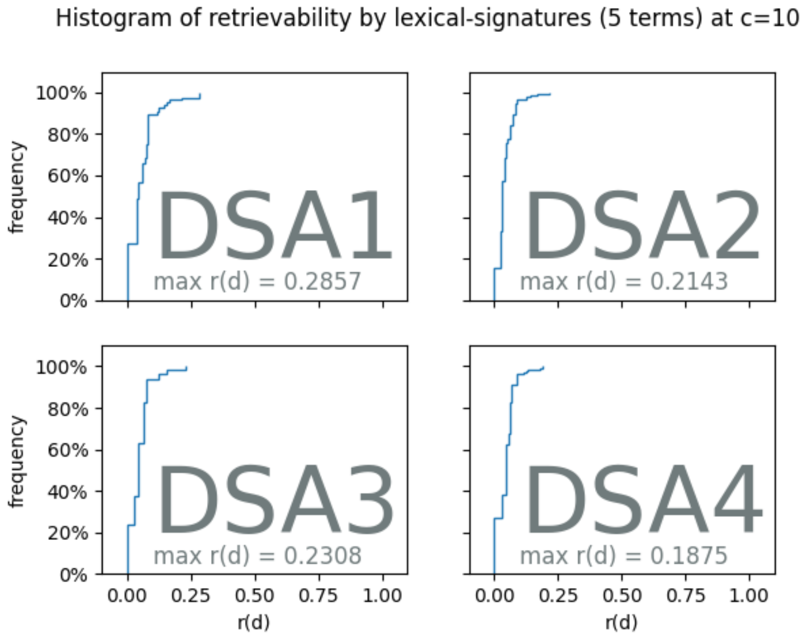


Fig. 114. Histogram at $c = 10$ by query method: lexical-signatures (5 terms).

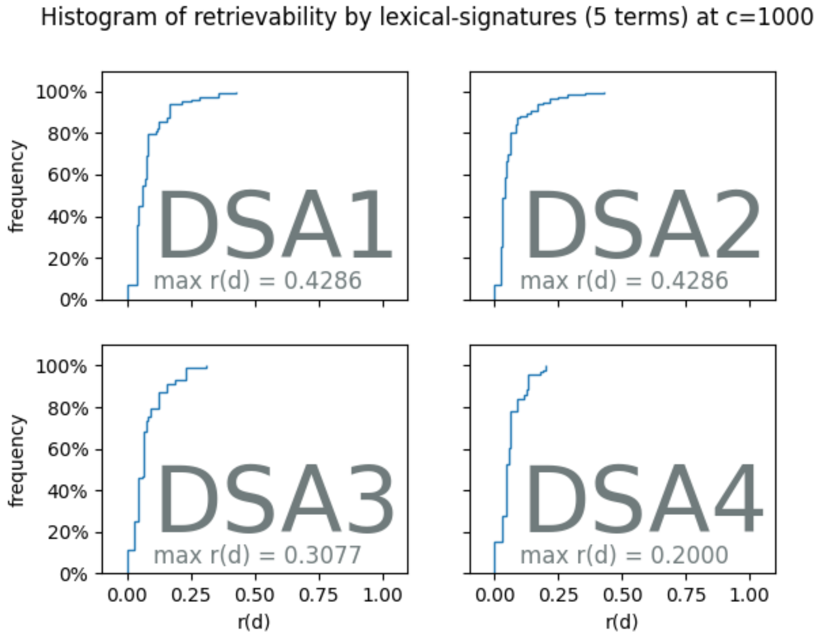


Fig. 115. Histogram at $c = 1000$ by query method: lexical-signatures (5 terms).

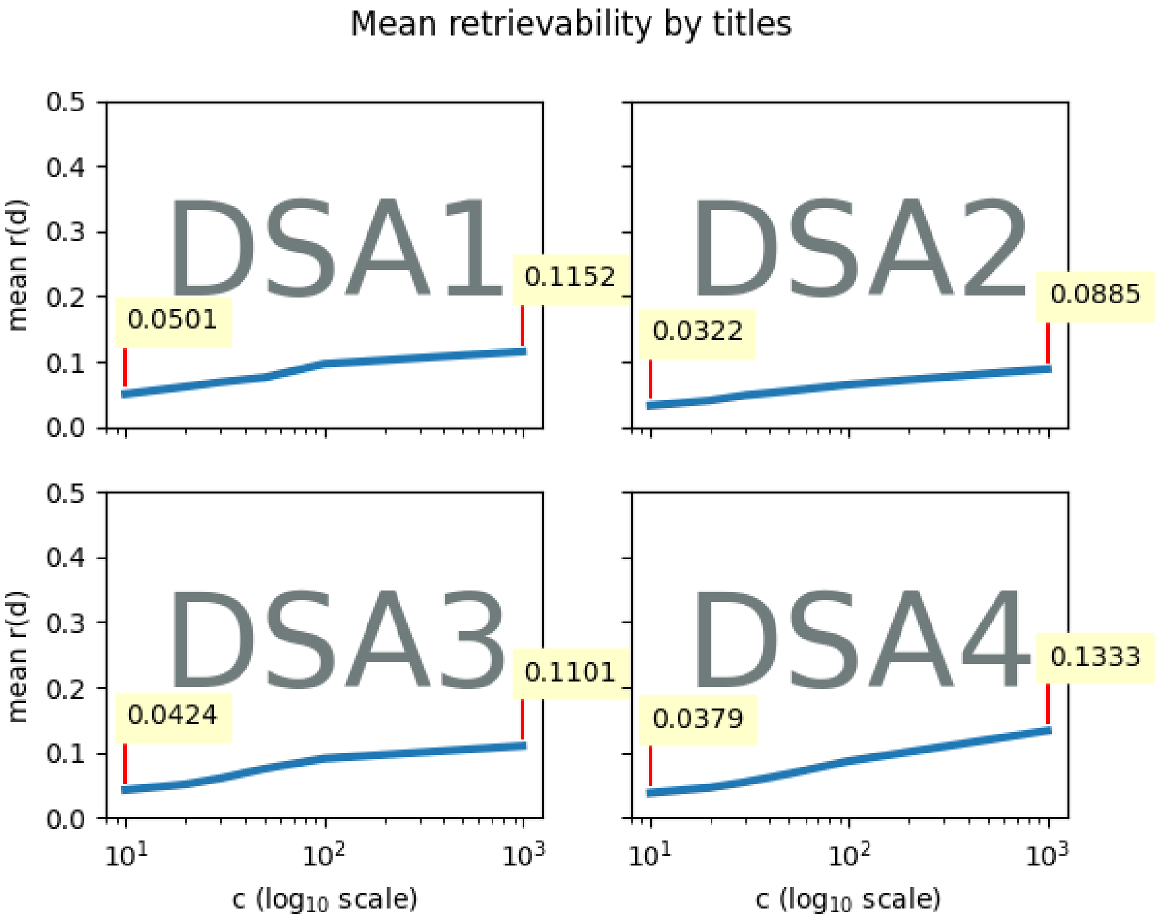


Fig. 116. Mean retrievability per cutoff value by query method: titles.

TABLE 26

p-values at $c = 10$ ($p < 0.05$ in bold) by query method: titles.

	DSA1	DSA2	DSA3	DSA4
DSA1	1.00000	**0.00016**	0.16123	**0.02523**
DSA2	**0.00016**	1.00000	**0.00973**	0.14587
DSA3	0.16123	**0.00973**	1.00000	0.32318
DSA4	**0.02523**	0.14587	0.32318	1.00000

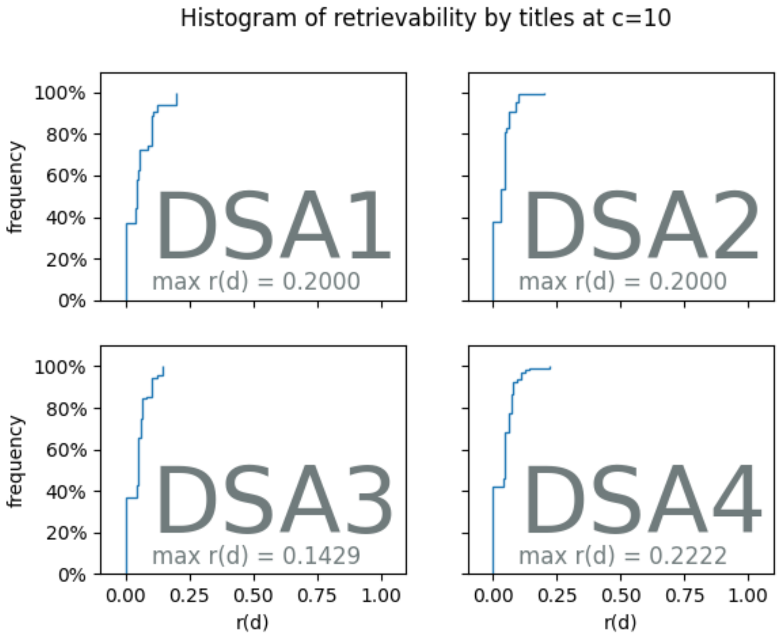


Fig. 117. Histogram at $c = 10$ by query method: titles.

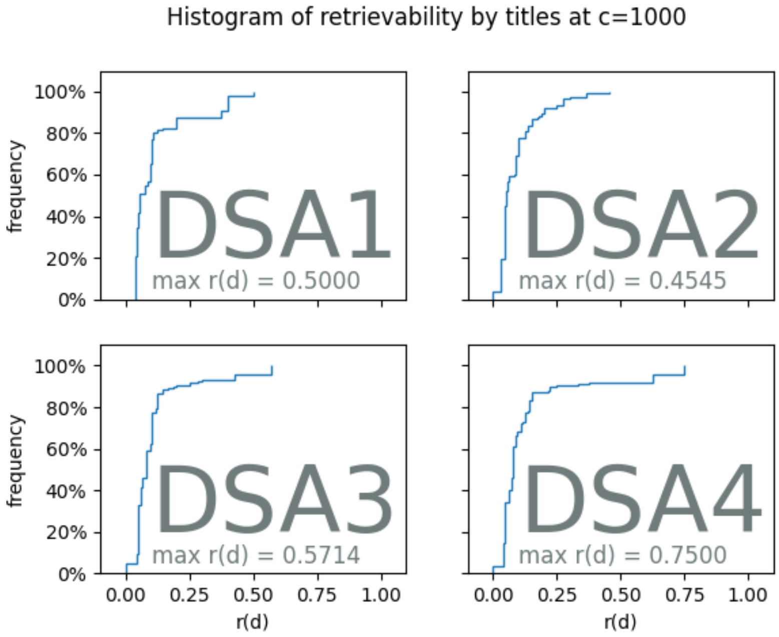


Fig. 118. Histogram at $c = 1000$ by query method: titles.

al. [258] for directly retrieving a document. These queries simulate a user searching for a specific document that they strongly remember. DSA1 has the highest mean retreivability of 0.0789 at $c = 10$ (Figure 113) and this is statistically significant when compared with other DSA2 and DSA4 (Figure 25). Looking at the histograms in Figure 114 we see the lowest maximum retrievability score with DSA4. This trend does hold at $c = 1000$, with DSA4's maximum retrievability only reaching 0.2000. The comparisons for DSA4 are statistically signficant at $c = 1000$ for DSA1 ($p = 0.00524$) and DSA3 ($p = 0.04301$), again supporting that DSA4 is meeting its goal.

Like with the entities, we see the same pattern of the 1-term lexical signature outperforming the 5-term query. The same scenario that occurs with the entities is at play here as well, with all values in Figure 110 starting at roughly $\frac{1}{\bar{k}} = \frac{1}{22.00} = 0.04545$ and slighly improving as $c \to 1000$.

Finally, our title queries simulate a user searching for a specific document. In Figure 116, titles perform comparable to lexical signaturess for locating the same memento within a collection. We again reach values roughly approximating $\frac{1}{\bar{k}} = \frac{1}{22.00} = 0.04545$ at $c = 10$. DSA1 has the highest mean retrievability score at 0.0501, but is only statistically significant (Figure 26) when compared with DSA2 and DSA4. One of DSA4's mementos reaches a retrievability of 0.2222 (Figure 117), giving that algorithm the highest retrievability at $c = 10$. At $c = 1000$, DSA4 has the highest retrievability for a single memento at 0.7500 (Figure 118) and these results are only statistically significant at $c = 1000$ when compared with DSA2.

We see several trends in these results. For doc2query-T5 and lexical signatures of one, two, and five terms, DSA4 consistently has the statistically significant lowest retrievability when compared to other algorithms' performance, supporting the hypothesis that it is meeting its goal of surfacing novel mementos from the collection, at least for some types of queries. DSA3 only has the highest retrievability when using the query method of top entities and only once a user has reviewed 1,000 search results ($c = 1000$). From this result, we have little conclusive evidence that DSA3 is working as intended, in spite of its extra steps to filter out content that does not support the general topic of the collection.

In no case, however, do we see any of the algorithms produce a single memento with 100% retrievability, even at $c = 1000$. This means that the DSA algorithms surface mementos that are difficult to retrieve with a search engine. Retrievability has another facet besides its mean, however. How many mementos had zero retrievability?

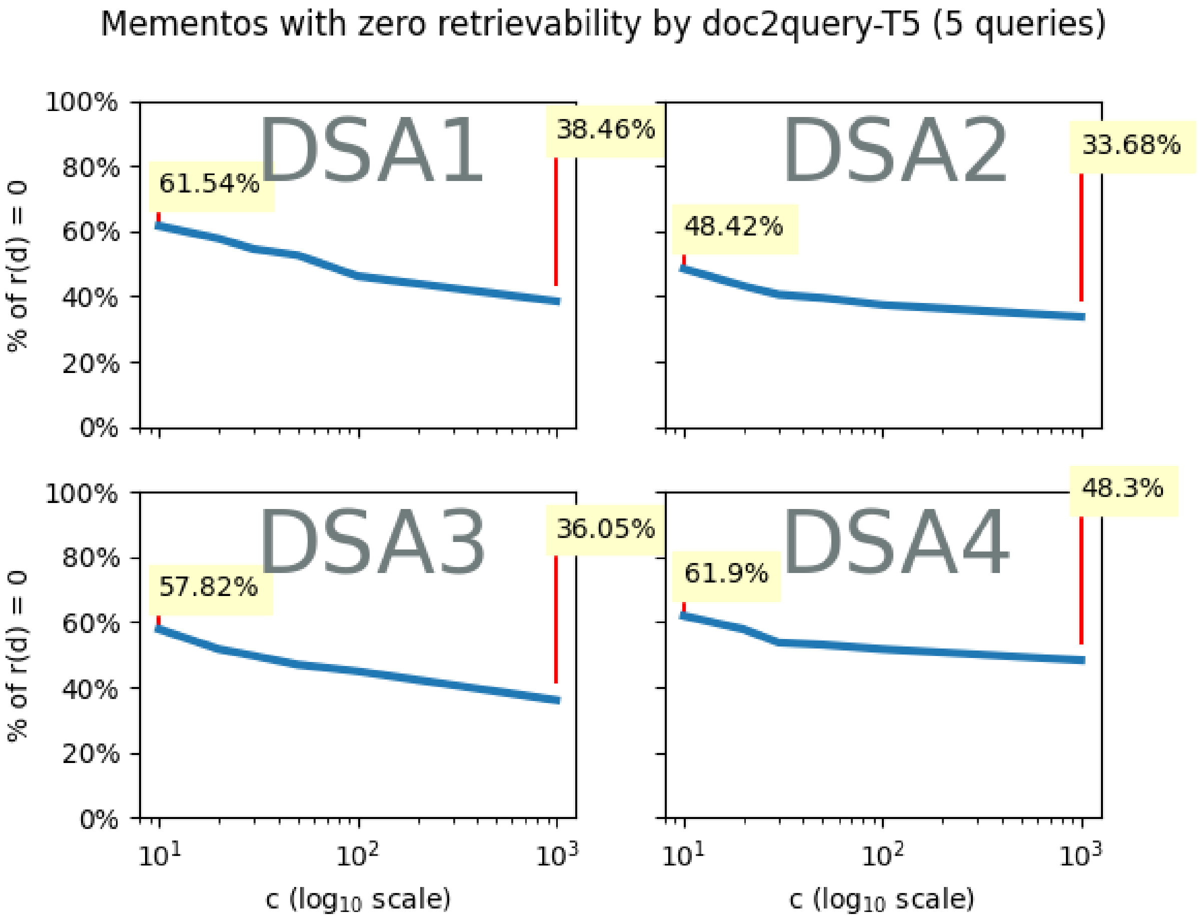


Fig. 119. Percentage of story mementos with zero retrievability at different cutoff values of *c* by query method: doc2query-T5 (5 queries).

4.4.3 ANALYZING ZERO RETRIEVABILITY OF DSA ALGORITHMS

The last section focused on mean and max retrievability for our exemplars. For an exemplar to score a retrievability of 0, none of its $\mathfrak{Q}$ queries would have returned it. In this section, we calculate the percentage of exemplars that have a retrievability of 0. This percentage tells us the number of exemplars that are not found at all. With a percentage of 100%, none of the DSA algorithm's mementos are retrievable by the queries in the query method. With a percentage of 0%, all of the DSA algorithm's mementos are retrieveable by at lease some query in the query method. Figures 119 through 126 show the percentage of mementos chosen by each DSA algorithm that produce zero retrievability when searched by applying the given query method. The x-axis of each visualization shows the cutoff, and the y-axis the percentage. The higher the percentage,

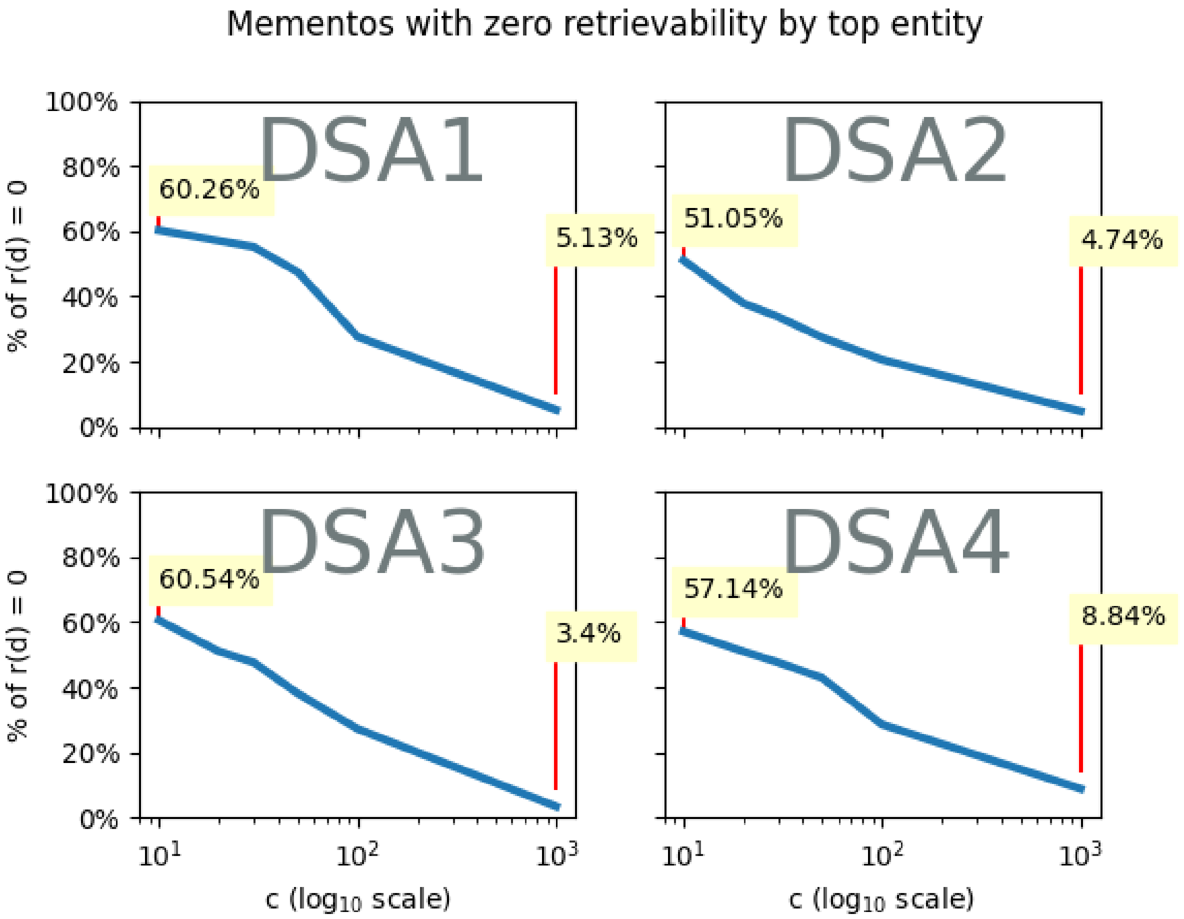


Fig. 120. Percentage of story mementos with zero retrievability at different cutoff values of *c* by query method: top entity.

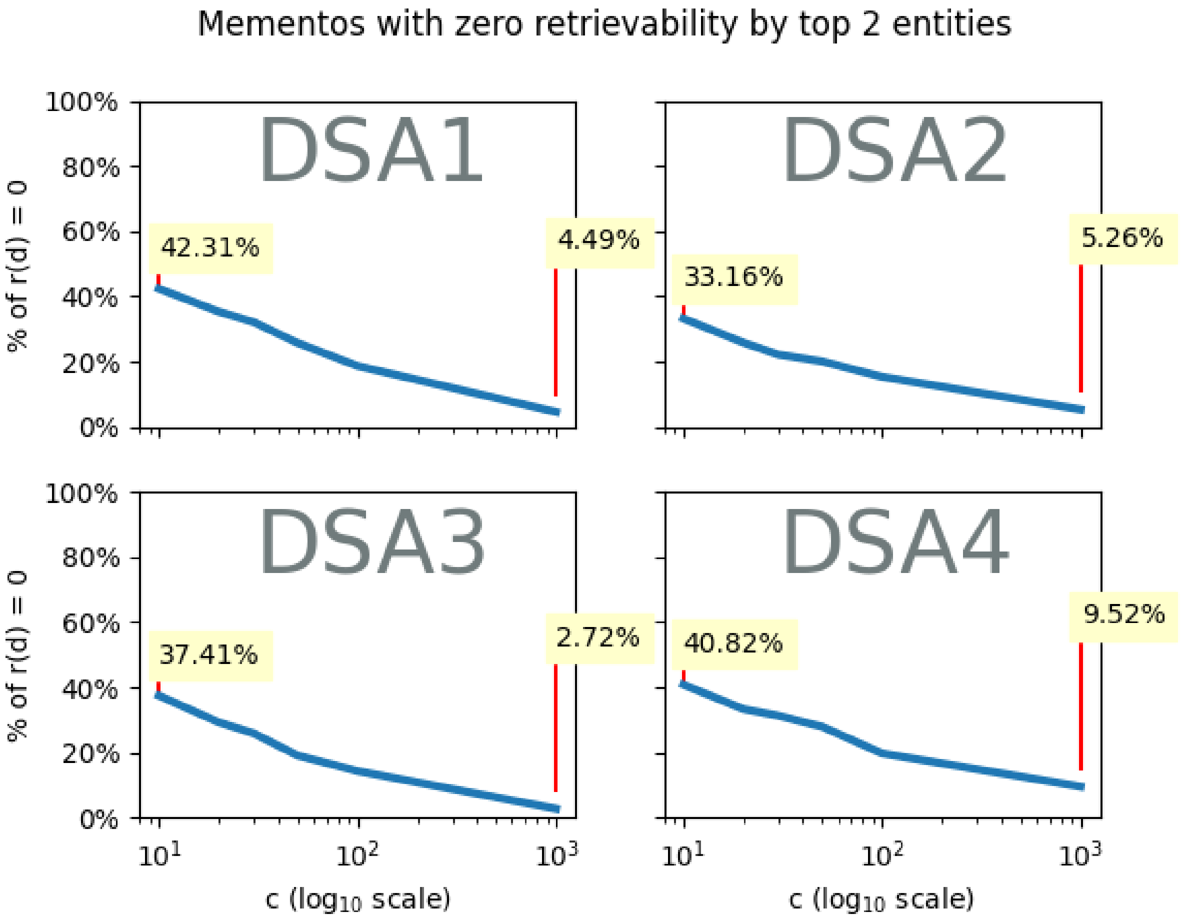


Fig. 121. Percentage of story mementos with zero retrievability at different cutoff values of *c* by query method: top 2 entities.

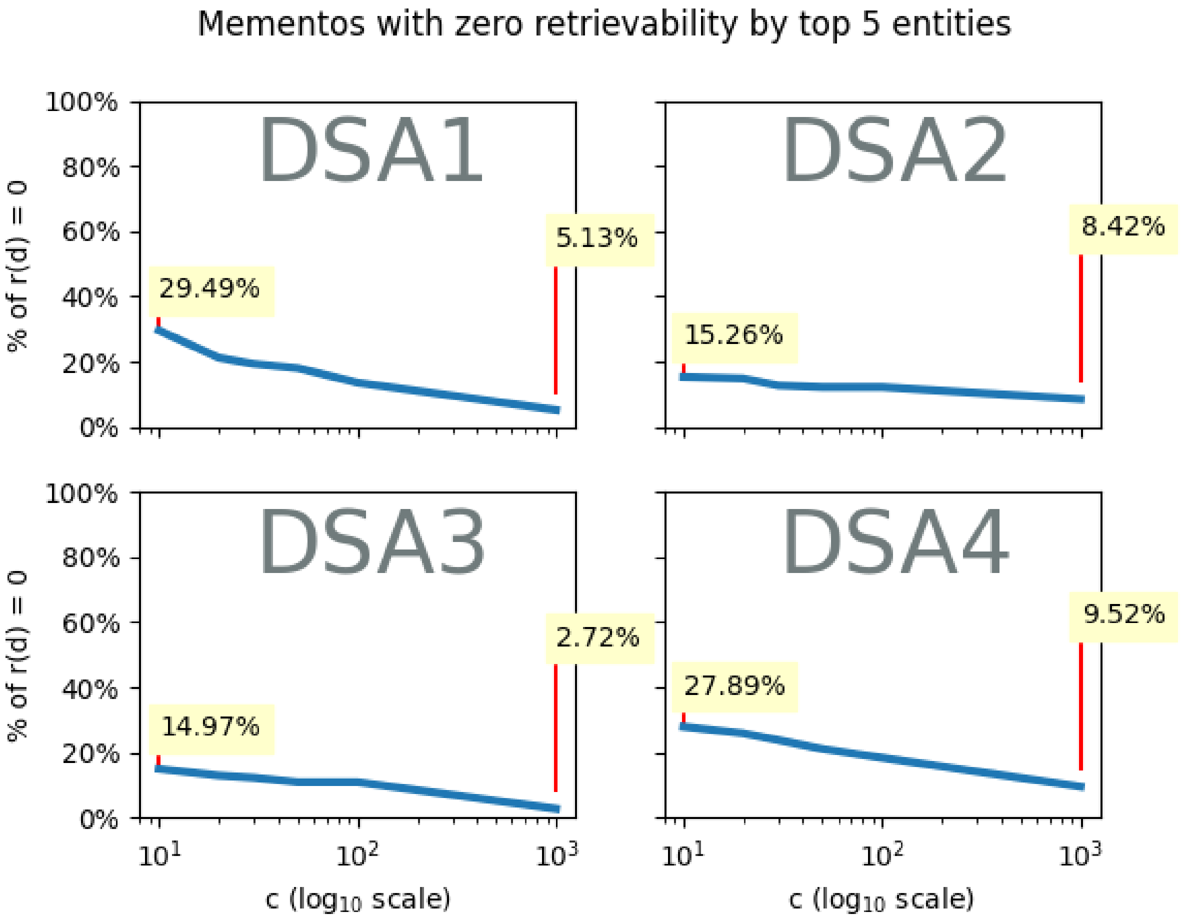


Fig. 122. Percentage of story mementos with zero retrievability at different cutoff values of *c* by query method: top 5 entities.

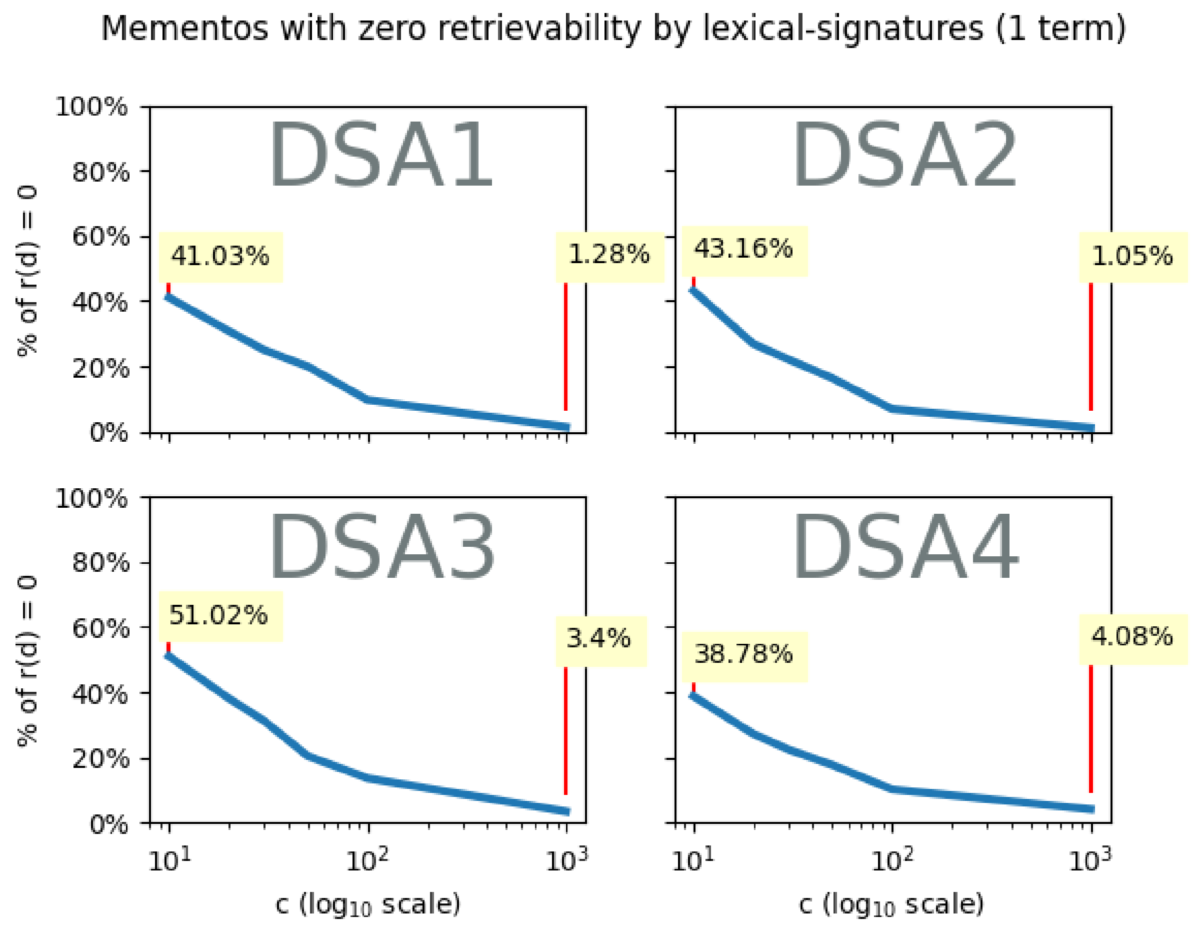


Fig. 123. Percentage of story mementos with zero retrievability at different cutoff values of *c* by query method: lexical signatures (1 term).

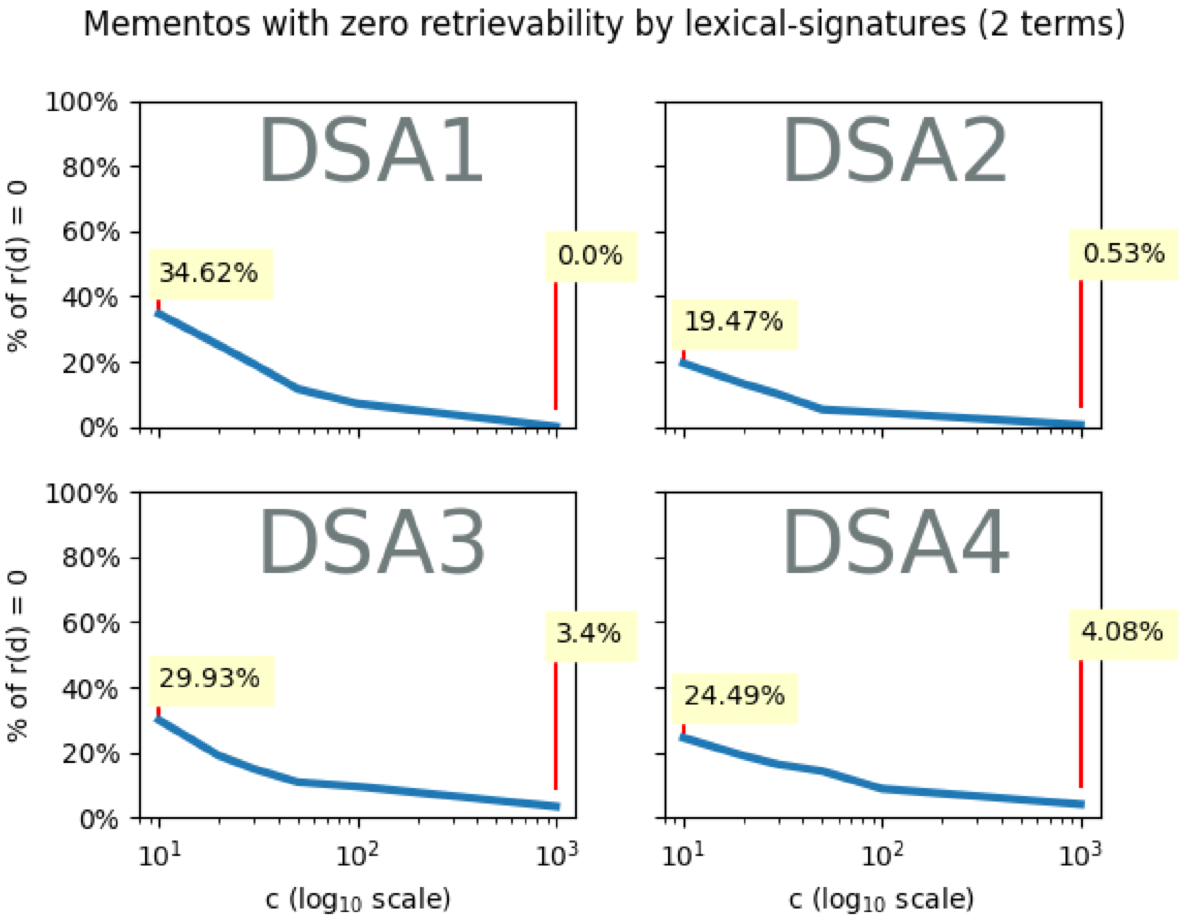


Fig. 124. Percentage of story mementos with zero retrievability at different cutoff values of *c* by query method: lexical signatures (2 terms).

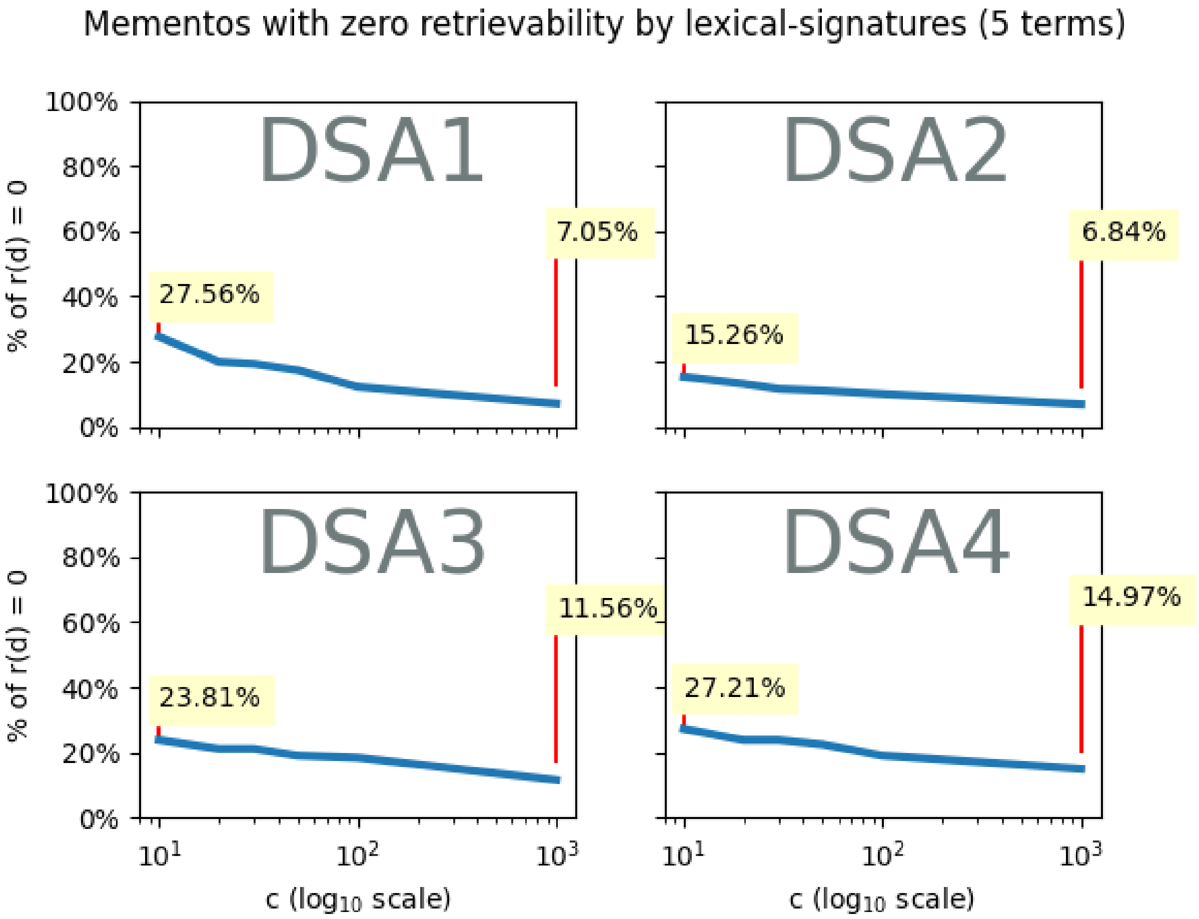


Fig. 125. Percentage of story mementos with zero retrievability at different cutoff values of *c* by query method: lexical signatures (5 terms).

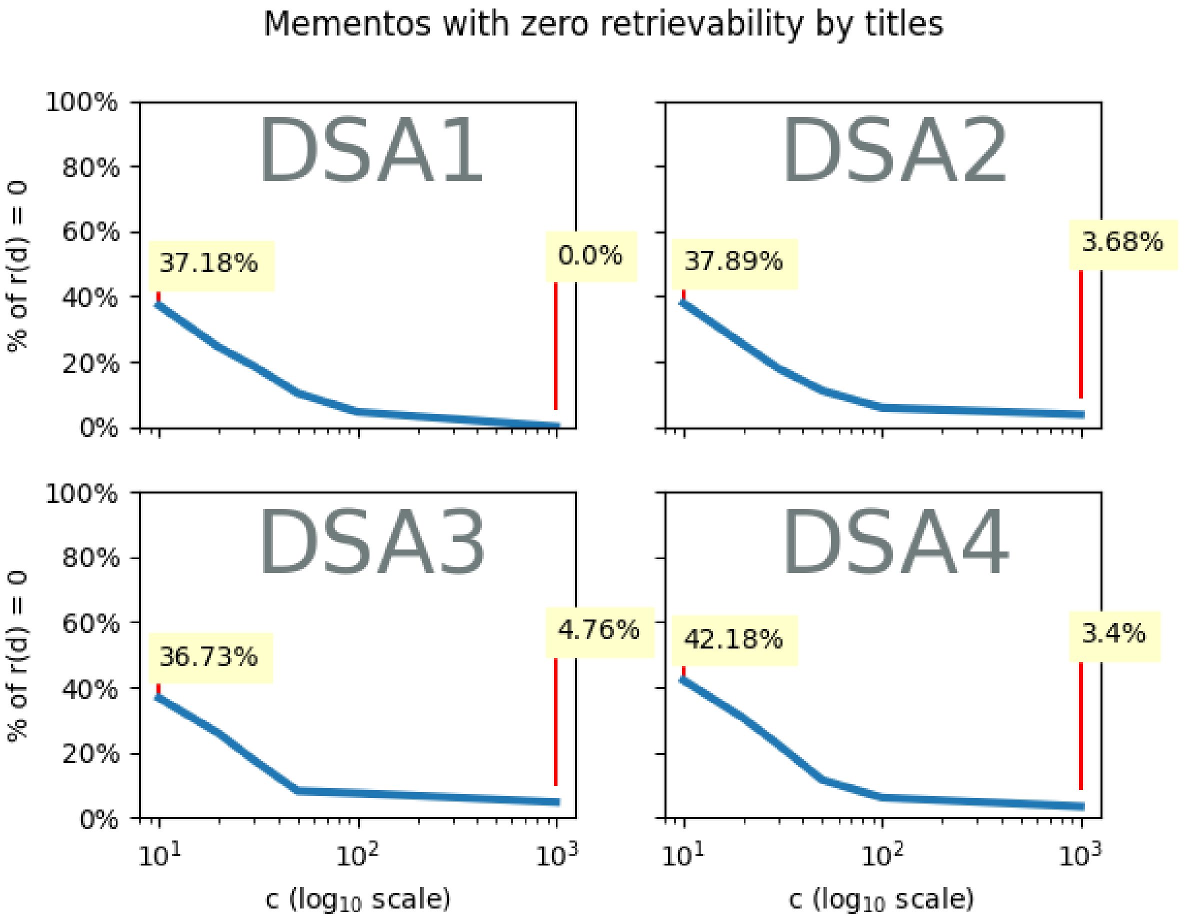


Fig. 126. Percentage of story mementos with zero retrievability at different cutoff values of *c* by query method: titles.

the more mementos are missed by our queries but surfaced by our DSA algorithms. The lower the percentage, the more successful a user will be at applying a search engine to produce these mementos.

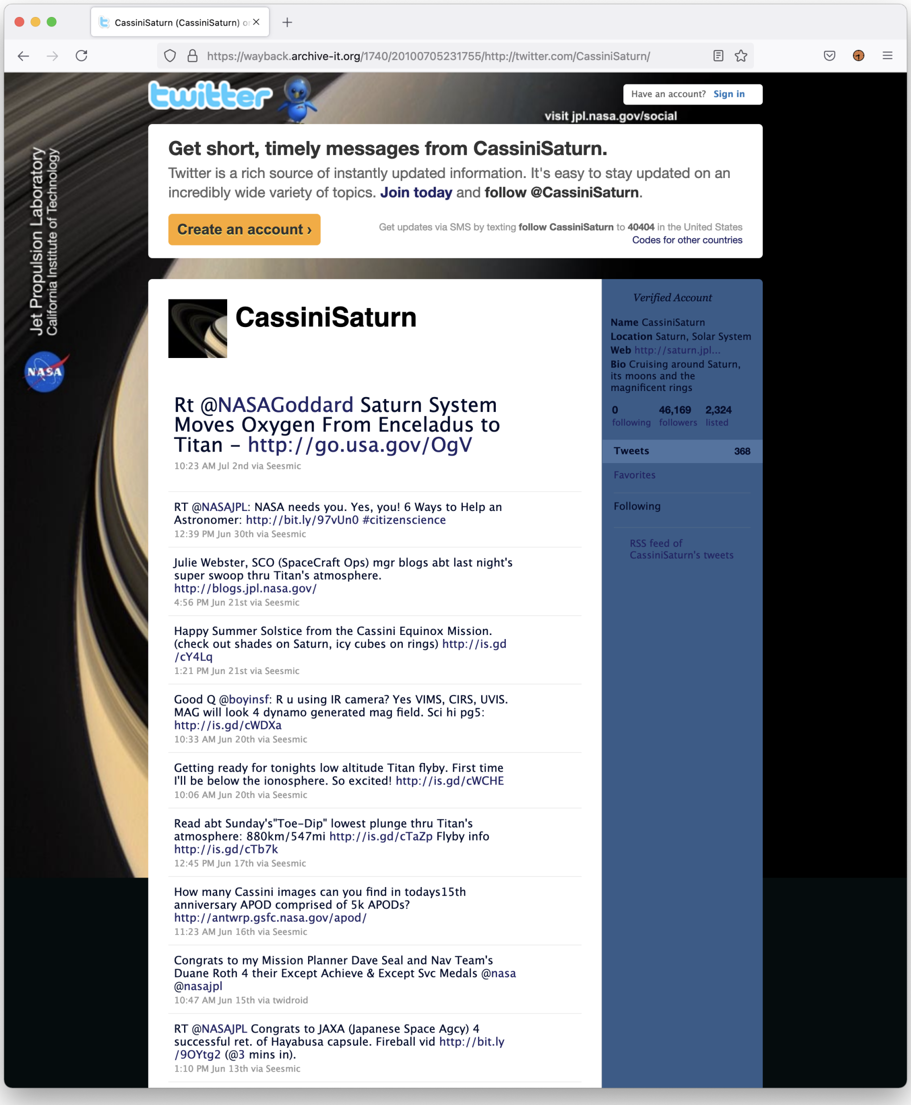
Applying doc2query-T5 produces the highest percentage of mementos with zero retrievability. DSA4 leads with 61.9% of its mementos having zero retrievability at $c = 10$ and 48.3% of its mementos being unretrievable at $c = 1000$. Other algorithms' scores are comparable but not as high. These results are consistent with the low retrievability scores we viewed for doc2query-T5 in the last section. This provides further evidence for DSA4 meeting its goals. DSA2 has the highest percentage of mementos with zero retrievability, not DSA3 as expected.

In Figure 120, DSA3 has the highest percentage of mementos with zero retrievability when queried by the top named entity found in each memento. Other algorithms are comparable. We see a steep drop by $c = 1000$, with DSA4 leading the others with 8.84% of its mementos still not retrievable. Much like we saw in the last section, this behavior provides further evidence that a single named entity surfaced by measuring and ranking entity frequencies can help users retrieve much of the collection. The DSA algorithms demonstrate, though, that between 3% and 9% of the collection remains out of reach by searching.

Figures 121 and 122 demonstrate similar behavior for named entities, but starting with lower percentages of mementos with zero retrievability than we saw for the top entity. DSA1 still has the highest value for queries containing the top 2 and top 5 entities at $c = 10$. With 5-entity queries, we have fewer mementos with zero retrievability, starting with 29.49% via DSA1. If we compare this with the last section, we see that the mean retrievability scores for 5-entity queries were lower than those for 1- and 2-entity queries. By $c = 1000$, however, we see the same steep drop for both query methods. With 9.52% of DSA4's mementos not retrievable with 2-entity queries or 5-entity queries, DSA4 is still surfacing more novel mementos than the search engine.

Single term lexical signatures perform better. Figure 123 shows how DSA3 has the most mementos with zero retrievability at 44.85% at $c = 10$. Of DSA4's mementos, 4.08% are still unreachable with this query method at $c = 1000$. With two-term lexical signatures, we further reduce the number of mementos with zero retrievability. DSA1 has the highest number of mementos with zero retrievability at 34.62% at $c = 10$ (Figure 124). By $c = 1000$, all DSA1 mementos are retrievable with 0% having zero retrievability. For DSA4 24.49% of its mementos have zero retrievability at $c = 10$, but 4.08% of DSA4's mementos are still out of reach by $c = 1000$.

At $c = 10$, five-term lexical signatures produce the lowest percentages of mementos with zero retrievability. Figure 125 shows how DSA1 has the highest percentage of mementos with zero retrievability at 27.45%. This result is the lowest score for DSA1 across all query methods shown


CassiniSaturn (CassiniSaturn) o
https://wayback.archive-it.org/1740/20100705231755/http://twitter.com/CassiniSaturn/
twitter
Have an account? Sign in
visit jpl.nasa.gov/social
Jet Propulsion Laboratory
California Institute of Technology
NASA
Create an account ›
Tweets 368
Favorites
Following
RSS feed of CassiniSaturn's tweets

Get short, timely messages from CassiniSaturn.

Twitter is a rich source of instantly updated information. It's easy to stay updated on an incredibly wide variety of topics. Join today and follow @CassiniSaturn.

Get updates via SMS by texting follow CassiniSaturn to 40404 in the United States
Codes for other countries

CassiniSaturn

Verified Account

Name CassiniSaturn
Location Saturn, Solar System
Web http://saturn.jpl...
Bio Cruising around Saturn, its moons and the magnificent rings

0	46,169	2,324
following	followers	listed

Rt @NASAGoddard Saturn System Moves Oxygen From Enceladus to Titan – http://go.usa.gov/OgV
10:23 AM Jul 2nd via Seesmic

RT @NASAJPL: NASA needs you. Yes, you! 6 Ways to Help an Astronomer: http://bit.ly/97vUn0 #citizenscience
12:39 PM Jun 30th via Seesmic

Julie Webster, SCO (SpaceCraft Ops) mgr blogs abt last night's super swoop thru Titan's atmosphere. http://blogs.jpl.nasa.gov/
4:56 PM Jun 21st via Seesmic

Happy Summer Solstice from the Cassini Equinox Mission. (check out shades on Saturn, icy cubes on rings) http://is.gd/cY4Lq
1:21 PM Jun 21st via Seesmic

Good Q @boyinsf: R u using IR camera? Yes VIMS, CIRS, UVIS. MAG will look 4 dynamo generated mag field. Sci hi pg5: http://is.gd/cWDXa
10:33 AM Jun 20th via Seesmic

Getting ready for tonights low altitude Titan flyby. First time I'll be below the ionosphere. So excited! http://is.gd/cWCHE
10:06 AM Jun 20th via Seesmic

Read abt Sunday's"Toe-Dip" lowest plunge thru Titan's atmosphere: 880km/547mi http://is.gd/cTaZp Flyby info http://is.gd/cTb7k
12:45 PM Jun 17th via Seesmic

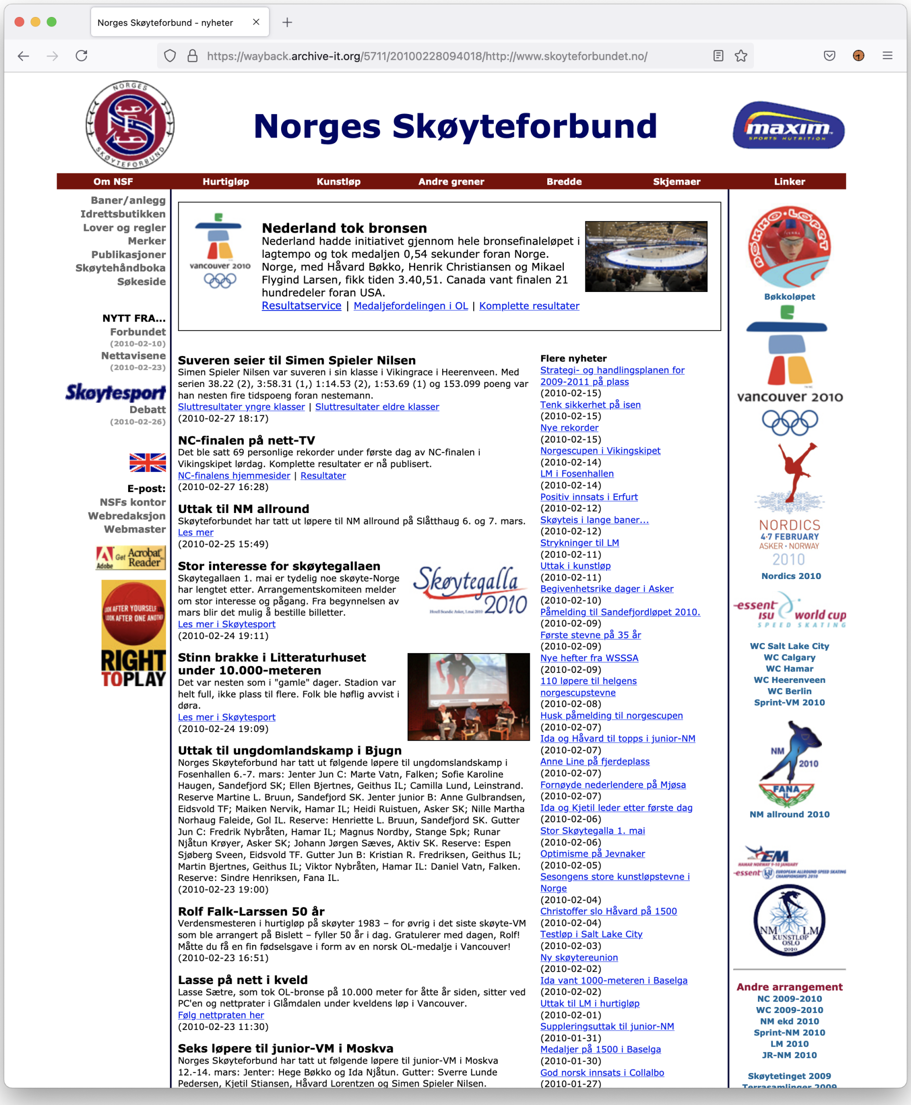
How many Cassini images can you find in todays15th anniversary APOD comprised of 5k APODs? http://antwrp.gsfc.nasa.gov/apod/
11:23 AM Jun 16th via Seesmic

Congrats to my Mission Planner Dave Seal and Nav Team's Duane Roth 4 their Except Achieve & Except Svc Medals @nasa @nasajpl
10:47 AM Jun 15th via twidroid

RT @NASAJPL Congrats to JAXA (Japanese Space Agcy) 4 successful ret. of Hayabusa capsule. Fireball vid http://bit.ly/9OYtg2 (@3 mins in).
1:10 PM Jun 13th via Seesmic

(a) Surfaced by DSA1 from collection 1740
URI-M:
`https://wayback.archive-it.org/1740/20100705231755/http://twitter.com/CassiniSaturn/`

Fig. 127. Examples of mementos with zero retrievability in SolrWayback, but that were surfaced by DSA algorithms.



Norges Skøyteforbund

Nederland tok bronsen

Nederland hadde initiativet gjennom hele bronsefinaleløpet i lagtempo og tok medaljen 0,54 sekunder foran Norge. Norge, med Håvard Bøkko, Henrik Christiansen og Mikael Flygind Larsen, fikk tiden 3.40,51. Canada vant finalen 21 hundredeler foran USA.

Resultatservice | Medaljefordelingen i OL | Komplette resultater

Suveren seier til Simen Spieler Nilsen

Simen Spieler Nilsen var suveren i sin klasse i Vikingrace i Heerenveen. Med serien 38.22 (2), 3:58.31 (1,) 1:14.53 (2), 1:53.69 (1) og 153.099 poeng var han nesten fire tidspoeng foran nestemann.

Sluttresultater yngre klasser | Sluttresultater eldre klasser

(2010-02-27 18:17)

NC-finalen på nett-TV

Det ble satt 69 personlige rekorder under første dag av NC-finalen i Vikingskipet lørdag. Komplette resultater er nå publisert.

NC-finalens hjemmesider | Resultater

(2010-02-27 16:28)

Uttak til NM allround

Skøyteforbundet har tatt ut løpere til NM allround på Slåtthaug 6. og 7. mars.

Les mer

(2010-02-25 15:49)

Stor interesse for skøytegallaen

Skøytegallaen 1. mai er tydelig noe skøyte-Norge har lengtet etter. Arrangementskomiteen melder om stor interesse og pågang. Fra begynnelsen av mars blir det mulig å bestille billetter.

Les mer i Skøytesport

(2010-02-24 19:11)

Stinn brakke i Litteraturhuset under 10.000-meteren

Det var nesten som i "gamle" dager. Stadion var helt full, ikke plass til flere. Folk ble høflig avvist i døra.

Les mer i Skøytesport

(2010-02-24 19:09)

Uttak til ungdomlandskamp i Bjugn

Norges Skøyteforbund har tatt ut følgende løpere til ungdomslandskamp i Fosenhallen 6.-7. mars: Jenter Jun C: Marte Vatn, Falken; Sofie Karoline Haugen, Sandefjord SK; Ellen Bjertnes, Geithus IL; Camilla Lund, Leinstrand. Reserve Martine L. Bruun, Sandefjord SK. Jenter junior B: Anne Gulbrandsen, Eidsvold TF; Maiken Nervik, Hamar IL; Heidi Ruistuen, Asker SK; Nille Martha Norhaug Faleide, Gol IL. Reserve: Henriette L. Bruun, Sandefjord SK. Gutter Jun C: Fredrik Nybråten, Hamar IL; Magnus Nordby, Stange Spk; Runar Njåtun Krøyer, Asker SK; Johann Jørgen Sæves, Aktiv SK. Reserve: Espen Sjøberg Sveen, Eidsvold TF. Gutter Jun B: Kristian R. Fredriksen, Geithus IL; Martin Bjertnes, Geithus IL; Viktor Nybråten, Hamar IL: Daniel Vatn, Falken. Reserve: Sindre Henriksen, Fana IL.

(2010-02-23 19:00)

Rolf Falk-Larssen 50 år

Verdensmesteren i hurtigløp på skøyter 1983 – for øvrig i det siste skøyte-VM som ble arrangert på Bislett – fyller 50 år i dag. Gratulerer med dagen, Rolf! Måtte du få en fin fødselsgave i form av en norsk OL-medalje i Vancouver!

(2010-02-23 16:51)

Lasse på nett i kveld

Lasse Sætre, som tok OL-bronse på 10.000 meter for åtte år siden, sitter ved PC'en og nettprater i Glåmdalen under kveldens løp i Vancouver.

Følg nettpraten her

(2010-02-23 11:30)

Seks løpere til junior-VM i Moskva

Norges Skøyteforbund har tatt ut følgende løpere til junior-VM i Moskva 12.-14. mars: Jenter: Hege Bøkko og Ida Njåtun. Gutter: Sverre Lunde Pedersen, Kjetil Stiansen, Håvard Lorentzen og Simen Spieler Nilsen.

Flere nyheter

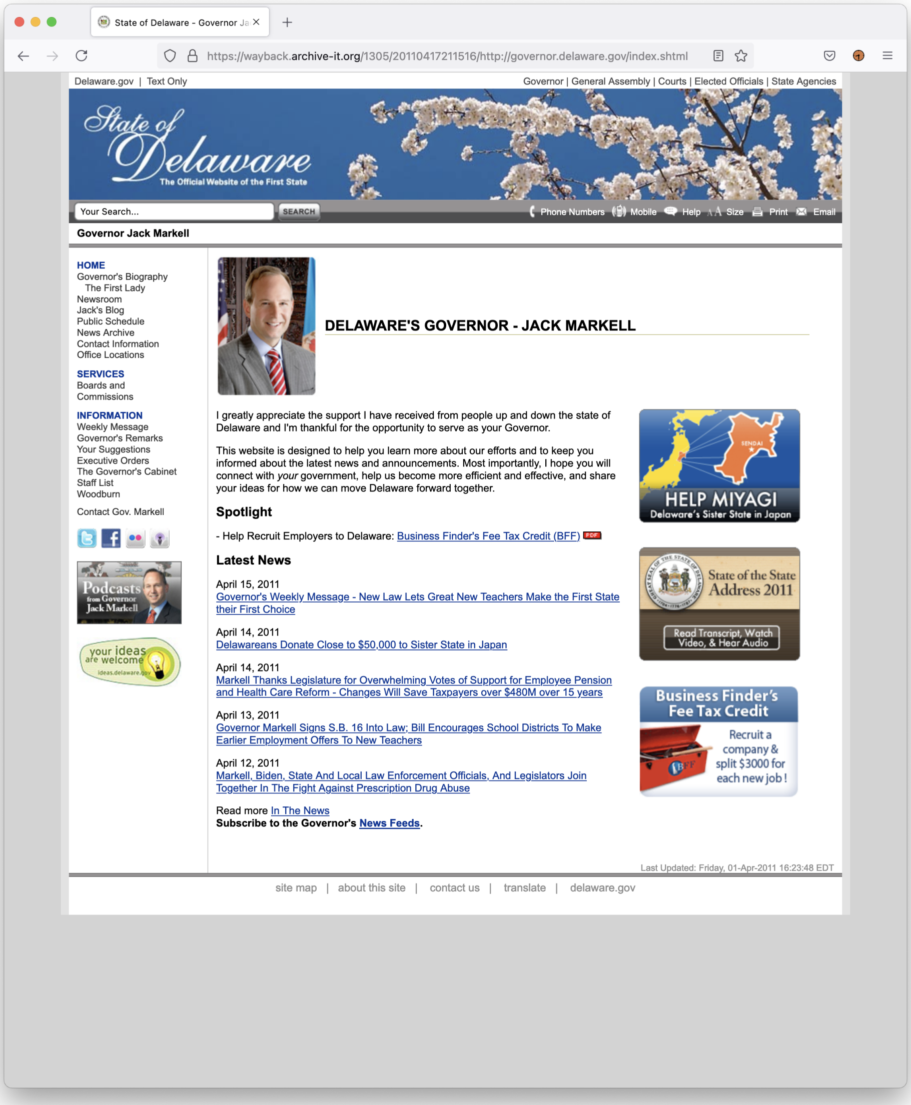
Strategi- og handlingsplanen for 2009-2011 på plass
(2010-02-15)
Tenk sikkerhet på isen
(2010-02-15)
Nye rekorder
(2010-02-15)
Norgescupen i Vikingskipet
(2010-02-14)
LM i Fosenhallen
(2010-02-14)
Positiv innsats i Erfurt
(2010-02-12)
Skøyteis i lange baner...
(2010-02-12)
Strykninger til LM
(2010-02-11)
Uttak i kunstløp
(2010-02-11)
Begivenhetsrike dager i Asker
(2010-02-10)
Påmelding til Sandefjordløpet 2010.
(2010-02-09)
Første stevne på 35 år
(2010-02-09)
Nye hefter fra WSSSA
(2010-02-09)
110 løpere til helgens norgescupstevne
(2010-02-08)
Husk påmelding til norgescupen
(2010-02-07)
Ida og Håvard til topps i junior-NM
(2010-02-07)
Anne Line på fjerdeplass
(2010-02-07)
Fornøyde nederlendere på Mjøsa
(2010-02-07)
Ida og Kjetil leder etter første dag
(2010-02-06)
Stor Skøytegalla 1. mai
(2010-02-06)
Optimisme på Jevnaker
(2010-02-05)
Sesongens store kunstløpstevne i Norge
(2010-02-04)
Christoffer slo Håvard på 1500
(2010-02-04)
Testløp i Salt Lake City
(2010-02-03)
Ny skøytereunion
(2010-02-02)
Ida vant 1000-meteren i Baselga
(2010-02-02)
Uttak til LM i hurtigløp
(2010-02-01)
Suppleringsuttak til junior-NM
(2010-01-31)
Medaljer på 1500 i Baselga
(2010-01-30)
God norsk innsats i Collalbo
(2010-01-27)

(b) Surfaced by DSA2 from collection 5711

URI-M:

`https://wayback.archive-it.org/5711/20100228094018/http://www.skoyteforbundet.no/`

Fig. 127. (Continued) Examples of mementos with zero retrievability in SolrWayback, but that were surfaced by DSA algorithms.


State of Delaware - Governor Ja
https://wayback.archive-it.org/1305/20110417211516/http://governor.delaware.gov/index.shtml
Delaware.gov | Text Only
Governor | General Assembly | Courts | Elected Officials | State Agencies
State of Delaware
The Official Website of the First State
Your Search...
SEARCH
Phone Numbers
Mobile
Help
Size
Print
Email
HOME
Governor's Biography
The First Lady
Newsroom
Jack's Blog
Public Schedule
News Archive
Contact Information
Office Locations
SERVICES
Boards and Commissions
INFORMATION
Weekly Message
Governor's Remarks
Your Suggestions
Executive Orders
The Governor's Cabinet
Staff List
Woodburn
Contact Gov. Markell
Podcasts from Governor Jack Markell
your ideas are welcome
HELP MIYAGI
Delaware's Sister State in Japan
State of the State Address 2011
Read Transcript, Watch Video, & Hear Audio
Business Finder's Fee Tax Credit
Recruit a company & split $3000 for each new job!
site map | about this site | contact us | translate | delaware.gov

Governor Jack Markell

DELAWARE'S GOVERNOR - JACK MARKELL

I greatly appreciate the support I have received from people up and down the state of Delaware and I'm thankful for the opportunity to serve as your Governor.

This website is designed to help you learn more about our efforts and to keep you informed about the latest news and announcements. Most importantly, I hope you will connect with *your* government, help us become more efficient and effective, and share your ideas for how we can move Delaware forward together.

Spotlight

- Help Recruit Employers to Delaware: Business Finder's Fee Tax Credit (BFF) PDF

Latest News

April 15, 2011
Governor's Weekly Message - New Law Lets Great New Teachers Make the First State their First Choice

April 14, 2011
Delawareans Donate Close to $50,000 to Sister State in Japan

April 14, 2011
Markell Thanks Legislature for Overwhelming Votes of Support for Employee Pension and Health Care Reform - Changes Will Save Taxpayers over $480M over 15 years

April 13, 2011
Governor Markell Signs S.B. 16 Into Law; Bill Encourages School Districts To Make Earlier Employment Offers To New Teachers

April 12, 2011
Markell, Biden, State And Local Law Enforcement Officials, And Legislators Join Together In The Fight Against Prescription Drug Abuse

Read more In The News
Subscribe to the Governor's News Feeds.

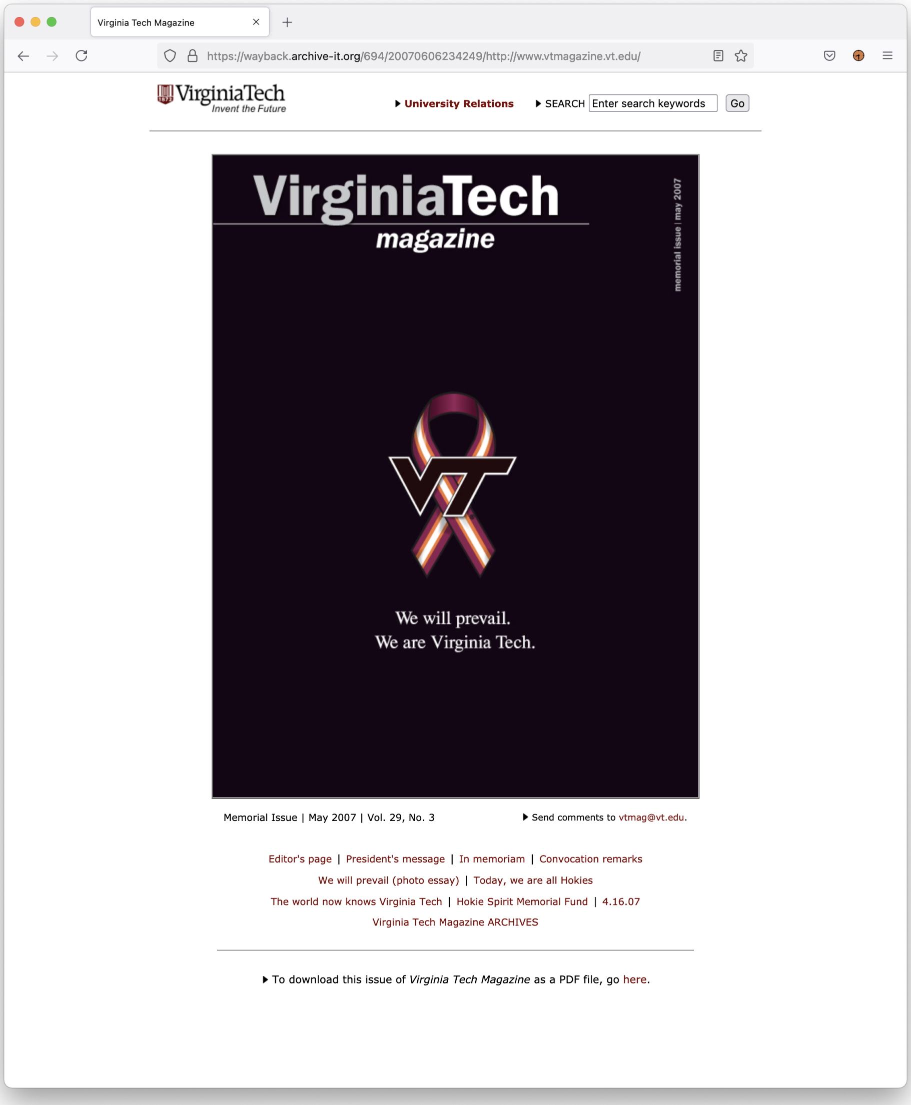
Last Updated: Friday, 01-Apr-2011 16:23:48 EDT

(c) Surfaced by DSA3 from collection 1305

URI-M:

`https://wayback.archive-it.org/1305/20110417211516/http://governor.delaware.gov/index.shtml`

Fig. 127. (Continued) Examples of mementos with zero retrievability in SolrWayback, but that were surfaced by DSA algorithms.



We will prevail.
We are Virginia Tech.

Memorial Issue | May 2007 | Vol. 29, No. 3

Send comments to vtmag@vt.edu.

Editor's page | President's message | In memoriam | Convocation remarks
We will prevail (photo essay) | Today, we are all Hokies
The world now knows Virginia Tech | Hokie Spirit Memorial Fund | 4.16.07
Virginia Tech Magazine ARCHIVES

To download this issue of *Virginia Tech Magazine* as a PDF file, go here.

(d) (Continued) Surfaced by DSA4 from collection 694
URI-M:
`https://wayback.archive-it.org/694/20070606234249/http://www.vtmagazine.vt.edu/`

Fig. 127. (Continued) Examples of mementos with zero retrievability in SolrWayback, but that were surfaced by DSA algorithms.

in Figures 119 through 126. DSA4 has the highest number of mementos with zero retrievability at 14.97% at $c = 1000$ using five-term lexical signatures.

Finally, with titles as queries in Figure 126, we see each algorithm start with 36% to 38% of their mementos undiscovered at $c = 10$. All of DSA1's mementos are discovered by at least one query by $c = 1000$. We discover nearly all of DSA4's mementos with only 3.4% outstanding. Using titles as queries with the search engine, we could not discover 4.76% of the mementos that DSA3 surfaced.

These results tell us that the DSA algorithms surface mementos that remain out of reach of search engine users – some examples of such mementos are shown in Figure 127. At "10 blue links", users will miss between 14% and 62% of mementos that they would otherwise encounter in our stories. If users perform an intensive search, they may find all of the mementos with the correct query but will have to review a thousand search results. DSA3 does not do a good job surfacing mementos that are retrievable with the search engine. Surpisingly, DSA1 did surface mementos that are retrieveable by both titles and two-term lexical signatures. Of these algorithms, DSA4 surfaced the largest percentage of mementos that were out of reach even at $c = 1000$ using the query methods of lexical signatures, entities, and doc2query-Tf5. With titles, DSA3 surpassed. This provides futher supporting evidence that DSA3 has failed to surface mementos with high retrievability and DSA4 surfaces novel mementos that are not easy to retrieved from the collection.

4.4.4 SUMMARY

A user like Natasha will first turn to a search engine to explore a web archive collection. She must start with a query, and she will base that query on the metadata she can find on the collection and any prior knowledge she brings to the endeavor. Our DSA algorithms surface $k \approx 28$ exemplars to describe the collection. This section evaluated how well someone like Natasha could reproduce our exemplars through search engine queries.

We synthesized WARCs from eight Archive-It collections and indexed them with SolrWayback, a web archive search engine. Once these mementos were available in SolrWayback, we generated stories from these collections with the DSA1, DSA2, DSA3, and DSA4 algorithms. To replicate search engine user behavior, we generated search queries from the mementos in these stories. We applied various methods, simulating casual searchers with doc2query-T5, more experienced users with named entities, users who recall the existence of specific documents with lexical signatures, and those who apply document titles to perform known-item searches.

We discovered that our DSA algorithms surface mementos that are not reachable through these search engine queries, even for known-item searches – i.e., with titles and five-term lexical signatures. That is to say that the DSA algorithms choose mementos with low retrievability scores. In DSA3, we tried to create an algorithm with high retrievability, but it did not succeed in its task, still surfacing novel mementos. In DSA4, we tried to build an algorithm with low retrievability and its low retrievabilty persists, even across differnt query types.

This study shows the efficacy of our method for helping surface mementos that are not otherwise discoverable. In addition to saving Natasha time by only sharing $k \approx 28$ exemplars, the DSA algorithms also provide insight not available by the de facto method available to her for collection understanding. In the next section, we introduce Hypercane, an application that helps users select exemplars from web archive collections.

4.5 HYPERCANE: IMPLEMENTING INTELLIGENT SAMPLING

Hypercane [144, 145, 146, 149, 157, 166] is a tool for summarizing web archive collections through intelligent sampling. It is a command-line application (`hc`) that takes a collection of mementos as input and provides various *actions* that can be applied to a web archive collection to produce a subset of URI-Ms as output. The primitives mentioned in Chapter 4.3 map to actions in Hypercane. For example, `hc cluster` performs clustering. Hypercane is an implementation of the model detailed in Chapter 4.3.

Because most Hypercane commands accept a list of URI-Ms as input, outputs from one Hypercane command can be fed into another, forming complex chains similar to Unix shell command pipes. More information on Hypercane's capabilities is available in its documentation [150] but we will provide a summary here. We divide Hypercane's actions into two categories: core and advanced. We intend core actions to be the ones most applied by end-users. Advanced actions can be employed to construct custom algorithms for a collection. Figure 128 shows how the different Hypercane actions map to the storytelling processes outlined in Chapter 1.

4.5.1 CORE ACTIONS

Hypercane's principal focus is to **select exemplars** from a web archive collection, thus its first core action is named `sample` after the same primitive from Chapter 4.3. The `sample` action accepts a list of URI-Ms and performs a specified sampling operation on the input. The user can select from a variety of **probabilistic sampling** algorithms, such as `true-random` which randomly selects k mementos from the collection, and `systematic`, which selects every j^{th} memento from the collection. Hypercane also provides **intelligent sampling** algorithms, such as DSA1, DSA2, DSA3, and DSA4, specified as `dsa1`, `dsa2`, `dsa3`, and `dsa4`. We also combine the two concepts in `filtered-random`, which eliminates off-topic and near-duplicate

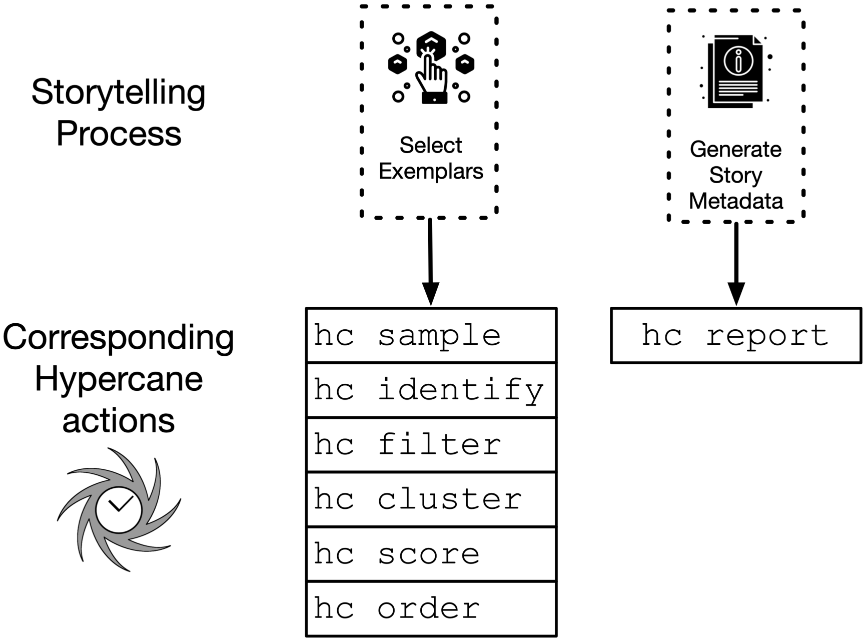


Fig. 128. The mapping of Hypercane actions mapped to the overall storytelling processes outlined in Chapter 1.

mementos before randomly selecting *k* from the remainder.

To fulfill an additional storytelling process, Hypercane can **generate story metadata** from a web archive collection or a list of exemplars through its `report` action. Through `report`, a user can generate a list of `terms` or `entities` from the list of exemplars for use in the final story. They can also apply `image-data` to the list of exemplars to ask Hypercane to choose the best representative image from all documents in the input. Likewise, `hc report metadata` can be run on the whole collection to extract metadata from the collection pages of the web archive collection itself. Users can apply this report to provide the collection title, creators, and other metadata for downstream storytelling.

The `report` action is not just limited to generating story metadata for storytelling. For the experiment run in Chapter 4.4, we generated queries through Hypercane's `hc report generate-queries` command. Hypercane also provides reports on `http-status` indicating the availability of mementos and Memento Protocol data from a web archive collection. A user can analyze individual HTML metadata fields with `hc report html-metadata`. A user can also create metadata based on the structural features from Chapter 4.1 through `hc report seed-statistics` and `hc report growth`. These capabilities are helpful for understanding and analyzing a collection but not necessarily helpful for storytelling.

Unconnected to our storytelling process is Hypercane's `synthesize` action, which allows users to convert their list of mementos into other formats, such as a directory of WARCs. The `synthesize` action is similar to the export capability of many tools and, except for `raintale-story` output, is more of a convenience feature than related to the storytelling process outlined in Chapter 1. We applied the `synthesize` action to generate WARCs from Archive-It collections in the study from Chapter 4.4.

4.5.2 ADVANCED ACTIONS

Hypercane's advanced actions allow users to create their own algorithms. Again, the output from most Hypercane commands is a list of URI-Ms, and any command can accept a list of URI-Ms, allowing users to chain Hypercane commands together. Internally, Hypercane implements intelligent sampling algorithms, such as `dsa1`, by executing scripts that contain these chains.

The `identify` action allows a user to convert one list of Memento Protocol objects into another. They can convert their list of mementos into a list of TimeMaps, or vice-versa. They can generate a list of URI-Rs from TimeMaps. It also includes convenience features for extracting all URI-Ms, URI-Ts, or URI-Rs from a given Trove, Archive-It, or PANDORA collection. We built the `identify` action into each of these commands so we can run it as part of their input handling. One can also run it alone to save their conversion for use outside of storytelling. Because Hypercane needs to download content to find mementos using `identify`, it is often best to save the output of `identify` before moving to the next step.

From Chapter 4.3, we have primitives for `filter`, `cluster`, `score`, and `order`. Each of these have their own corresponding Hypercane command combining the primitive with additional arguments (e.g., `hc order memento-datetime`).

Hypercane leverages many existing libraries to meet its goals. Recall our experiment with detecting off-topic mementos in Chapter 4.2. We implemented that algorithm in the Off-Topic Memento Toolkit (OTMT) [138, 141, 164]. Hypercane applies the OTMT to filter off-topic mementos via the `hc filter exclude off-topic` command. It leverages many other libraries, such as Gensim [342] for topic modeling (e.g., `hc cluster lda`), spaCy [121] for named entity recognition (e.g., `hc report entities`), and scikit-learn [261] for clustering (e.g., `hc cluster dbscan`).

4.5.3 CREATING ALGORITHMS WITH HYPERCANE

Figure 129 shows the Hypercane commands that execute AlNoamany's Algorithm. As noted in Chapter 3, AlNoamany's algorithm reduces the mementos under consideration by filtering out

```
hc identify timemaps -i archiveit -a 13529 -o timemaps.tsv
hc filter
   include-only on-topic -i timemaps -a timemaps.tsv -o ontopic.tsv
hc filter exclude
   near-duplicates -i mementos -a ontopic.tsv -o non-duplicates.tsv
hc filter include-only languages --lang en -i mementos -a non-
   duplicates.tsv -o english-only.tsv
hc cluster time-slice -i mementos -a english-only.tsv -o sliced.tsv
hc cluster dbscan -i mementos -a sliced.tsv -o sliced-and-clustered
   .tsv
hc score dsa1-scoring -i mementos -a sliced-and-clustered.tsv -o
   scored.tsv
hc filter include-only highest-score-per-cluster -i mementos -a
   scored.tsv -o highest-scored.tsv
hc order pubdate-else-memento-datetime -i mementos -a highest-
   scored.tsv -o ordered.tsv
```

Fig. 129. DSA1 run via Hypercane's advanced actions.

```
hc identify timemaps -i archiveit -a 13529 -o timemaps.tsv
hc filter
   include-only on-topic -i timemaps -a timemaps.tsv -o ontopic.tsv
hc filter exclude near-duplicates -i mementos -a ontopic.tsv -o non
   -duplicates.tsv
hc sample true-random -i mementos -k 28 -a non-duplicates.tsv -o
   random-sample.tsv
```

Fig. 130. Hypercane's `filtered-random` sampling algorithm.

```
hc identify timemaps -i archiveit -a 13529 -o timemaps.tsv
hc filter
   include-only on-topic -i timemaps -a timemaps.tsv -o ontopic.tsv
hc filter exclude near-duplicates -i mementos -a ontopic.tsv -o non-
   duplicates.tsv
hc cluster kmeans -i mementos -k 28 -a non-duplicates.tsv -o time-
   clustered.tsv --feature memento-datetime
hc cluster lda -i mementos -a time-clustered.tsv -o topic-clustered.tsv
hc score dsa2-scoring -i mementos -a topic-clustered.tsv -o scored.tsv
hc filter include-only highest-score-per-cluster -i mementos -a scored.
   tsv -o highest-scored.tsv
hc order pubdate-else-memento-datetime -i mementos -a highest-scored.
   tsv -o ordered.tsv
```

Fig. 131. The DSA2 Algorithm run via Hypercane's advanced actions.

those that are off-topic (`filter include-only on-topic`), near-duplicates (`exclude near-duplicates`), and non-English (`include-only languages --lang en`). Then it clusters by AlNomany's time slices (`cluster time-slice`) so that each slice contains roughly the same number of mementos. From there, it clusters again via DBSCAN (`cluster dbscan`) using the Simhash distance of each memento in the cluster from each other. Finally, it scores the mementos in each cluster based on a scoring function (`score dsa1-scoring`) that weights memento-damage, path depth, and page category. Hypercane then filters out all but the highest scoring memento (`filter include-only highest-score-per-cluster`) from each cluster and orders them by publication date (`order pubdate-else-memento-datetime`). A user executes these same steps when they run the command `hc sample dsa1`.

In Figure 130 we demonstrate how to construct the `filtered-random` algorithm from Chapter 4.3 that filters out off-topic and near-duplicate mementos before randomly choosing 28. In this example, we show how one can feed the output of other commands through an additional `sample` command to reduce the set of mementos further.

Figures 131, 132, and 133 show the algorithms DSA2, DSA3, and DSA4, respectively, as detailed in Chapter 4.3 and used in the experiment from Chapter 4.4. A user executes them with `hc sample dsa2`, `hc sample dsa3`, `hc sample dsa4`.

```
hc identify timemaps -i archiveit -a 13529 -o timemaps.tsv
hc filter
   include-only on-topic -i timemaps -a timemaps.tsv -o ontopic.tsv
hc filter exclude near-duplicates -i mementos -a ontopic.tsv -o non-
   duplicates.tsv
hc cluster dbscan -i mementos -a non-duplicates.tsv -o dbscan-clustered
   .tsv --feature tfidf --eps 1.3
hc filter exclude with-cluster-id -i mementos -a dbscan-clustered.tsv -
   o no-outliers.tsv --cluster-id -1
hc synthesize
   cluster-free -i mementos -a no-outliers.tsv -o clusters-stripped.tsv
hc cluster kmeans -i mementos -a clusters-stripped.tsv -o time-
   clustered.tsv --feature memento-datetime
hc cluster kmeans -i mementos -a time-clustered.tsv -o term-clustered.
   tsv --feature tfidf
hc score top-entities-and-bm25 -i mementos -a term-clustered.tsv -o
   scored.tsv
hc filter include-only highest-score-per-cluster -i mementos -a scored.
   tsv -o highest-scored.tsv
hc order pubdate-else-memento-datetime -i mementos -a highest-scored.
   tsv -o ordered.tsv
```

Fig. 132. The DSA3 Algorithm run via Hypercane’s advanced actions.

```
hc identify timemaps -i archiveit -a 13529 -o timemaps.tsv
hc filter
   include-only on-topic -i timemaps -a timemaps.tsv -o ontopic.tsv
hc filter exclude near-duplicates -i mementos -a ontopic.tsv -o non
   -duplicates.tsv
hc cluster kmeans -i mementos -a non-duplicates.tsv -o time-
   clustered.tsv --feature memento-datetime
hc cluster kmeans -i mementos -a time-clustered.tsv -o term-
   clustered.tsv --feature tfidf
hc score distance-from-centroid -i mementos -a term-clustered.tsv -
   o scored.tsv
hc filter include-only highest-score-per-cluster -i mementos -a
   scored.tsv -o highest-scored.tsv
hc order pubdate-else-memento-datetime -i mementos -a highest-
   scored.tsv -o ordered.tsv
```

Fig. 133. The DSA4 Algorithm run via Hypercane's advanced actions.

```
hc filter include-only containing-url-pattern \
    --url-pattern "http://bostinno.streetwise.co/2013/04/19/boston-
       marathon-suspect-dead-another-at-large-after-standoff/" \
    -i archiveit -a 3649 -o story-mementos.tsv
```

Fig. 134. The filter for Rustam's FPST story.

```
hc filter include-only near-datetime \
    --start-datetime "2018-07-11T00:00:00" \
    --end-datetime "2018-07-30T11:00:00"
    -i mementos -a news-mementos.tsv -o story-mementos.tsv
```

Fig. 135. A filter for Olayinka's SPFT story.

Elbert wants to create stories to promote his collection for people like Natasha. He can select exemplars by executing one of these commands. Then he can generate story metadata with `hc report`. Ling can do the same, so she has a set of exemplars from her collection. Ling can also apply the other `report` commands to explore her collection through structural features and additional metadata.

Hypercane does not accept WARCs as input. Its focus is to provide anyone access to a public web archive collection's mementos through the Memento Protocol. Not only can archivists apply Hypercane to promote or explore their own collections, but users can select their own exemplars to meet their needs. Recall that Rustam wants to tell a *fixed page, sliding time* (FPST) story where he can view a specific page on the Boston Marathon Bombing. He can execute Hypercane's `filter` action on a specific URI-R as shown in Figure 134.

Hypercane is **not** limited to Archive-It. It supports lists of URI-Ms from any Memento-compliant archive. Olayinka wanted to tell a *sliding page, fixed time* (SPFT) story centered on a specific time period. She can execute a `filter` action for a specific time range on that list as shown in Figure 135.

4.5.4 SUMMARY

Hypercane (`hc`) is a command-line application that implements the primitives detailed in Chapter 4.3. It has helped us address RQ1 and RQ2 throughout this chapter. First, Hypercane can sample mementos from a web archive collection. With Hypercane, a user can apply probabilistic sampling techniques, like true random, or intelligent sampling techniques, like DSA1. With intelligent sampling techniques, Hypercane can satisfy the storytelling process of selecting exemplars. Second, Hypercane's report action provides information about the collection. It can generate reports that contain data on a collection's structural features, as mentioned in Chapter 4.1, helping us better understand a collection overall. The `report` action also provides metadata, both generated and captured, on the collection or a list of exemplars, satisfying the storytelling process for generating story metadata.

Hypercane's potential is not just limited to existing sampling algorithms. We applied Hypercane to the experiments in this chapter. It allows the user to build their own algorithms out of advanced actions. Because Hypercane applies the Memento Protocol, it can sample from any public web archive collection — not just Archive-It — allowing users to select their own exemplars, generate their own story metadata, and explore web archive collections in ways not supported with existing search engines.

4.6 CHAPTER SUMMARY

In this chapter, we covered the storytelling processes of selecting exemplars and generating story metadata. These two processes are our first steps toward storytelling. Generating story metadata provides us with the information necessary to augment our story. It may also be used to help us select exemplars.

RQ1: What types of web archive collections exist and what structural features do they have?

We introduced Archive-It collections' structural features and semantic categories, providing different methods of characterizing a collection. We discovered that four semantic categories exist for Archive-It collections. The most common is *Self-Archiving*, where organizations are archiving their own content. The least common is *Time Bounded - Spontaneous*, where an archivist tries to archive the coverage of a natural disaster or other unexpected event.

We detailed seed features, like domain diversity, that allow a user to compare how diverse the sources are for a collection. We discussed growth curves, that demonstrate the temporal skew of a collection's contents. Clustering by time is an essential part of selecting exemplars from web archive collections. This temporal skew has implications for choosing which clustering algorithm produces mementos that best represent the temporal skew of the collection.

RQ2: Which approaches work best for selecting exemplars from web archive collections?

An essential ability to select exemplars is removing off-topic mementos from consideration. We evaluated different measures for determining if a memento is off-topic compared to the others in its TimeMap and concluded that word count performs best on our gold standard dataset with an F_1 score of 0.788.

We introduced a model for selecting exemplars through primitives. Most algorithms that select exemplars covered in Chapter 3.1 apply one or more primitives of *filter*, *cluster*, *score*, and *order*. We demonstrated how we combine these primitives to create AlNoamany's Algorithm (DSA1) from Chapter 3.5. We then discussed how our knowledge of structural features could be applied to improve AlNoamany's Algorithm with a new algorithm named DSA2. We further introduced DSA3 for selecting exemplars that best match the overall topic of the collection and DSA4 for selecting more novel exemplars.

We addressed how difficult it is for a user to apply a search engine and arrive at the exemplars selected by the DSA algorithms. We synthesized WARCs from Archive-It collections and indexed them with SolrWayback. Once in SolrWayback, we ran the DSA algorithms to select exemplars. We automatically generated queries that simulated how different types of users might search the collection from these exemplars. We then recorded the search engine results for each of these queries and discovered that the search queries did not surface all of the mementos selected by the DSA algorithms. These results show that the DSA algorithms are a valuable method for selecting exemplars thare are not easily retrieved with a search engine.

Finally, we introduced Hypercane, a tool for sampling from web archive collections. Hypercane provides several actions that allow the user to tailor algorithms to intelligently sample from the collection, producing the exemplars that best tell the story they desire. Hypercane's `sample` action satisfies the select exemplars storytelling process through pre-defined probabilistic and intelligent sampling algorithms. Hypercane's `report` action satisfies the generate story metadata storytelling process, providing collection-wide metadata and structural feature analysis. Hypercane's `synthesize` action allows one to convert mementos into WARCs. Hypercane's advanced actions permit the user to identify, filter, cluster, score, and order mementos as part of developing their own algorithms.

With our exemplars selected, in our next chapter, we summarize each document to generate document metadata. After that, we will combine the story metadata and the document metadata to tell stories.

CHAPTER 5

GENERATING DOCUMENT METADATA

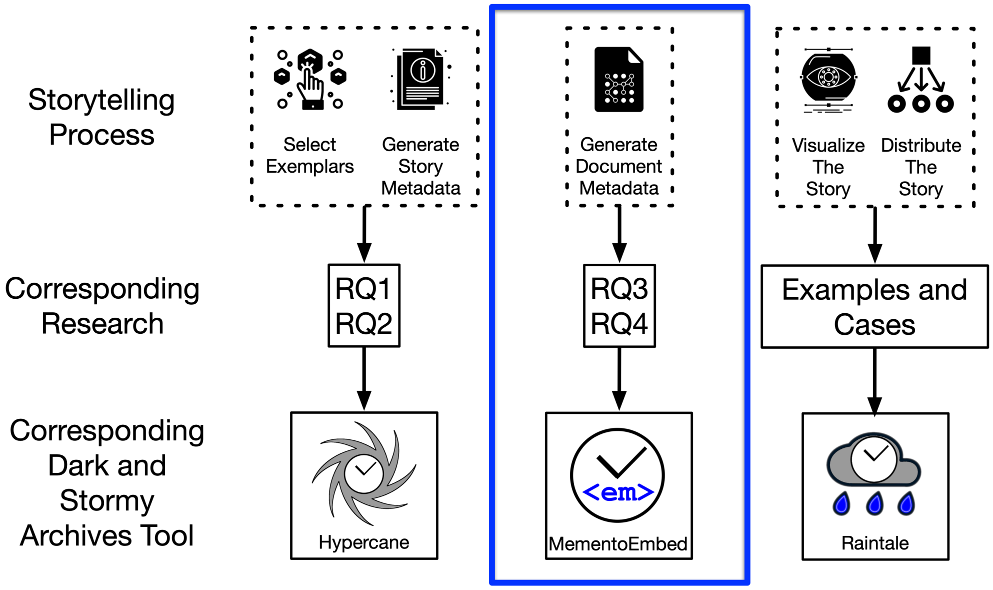


Fig. 136. The processes for storytelling that are covered in this chapter are indicated by the blue box.

Chapter 4 discussed issues with selecting exemplars and generating story metadata. Natasha wants to compare collections. Rustam wants to compare different versions of the same page over time. Olayinka wants to compare the content of different pages captured around the same time. Ling wants to understand her own collection. Elbert wants to promote a collection for others. All of these personas need someone to **generate document metadata** so they can understand the exemplars in their stories, thus this chapter covers that process (Figure 136). A surrogate is a visualization of an individual document's metadata. It is not easy for end-users to evaluate metadata without also evaluating a surrogate visualizing it.

In Chapter 5.1, we will summarize an evaluation of 55 platforms to find one that completes this storytelling process for mementos. We will discover that these platforms do a poor job for mementos, leading us to conduct a user study in Chapter 5.2 that helps us determine that social cards probably work best for understanding collections of mementos (RQ3). The results of this user study give us a surrogate to focus on, so in Chapter 5.3, we analyze how common social

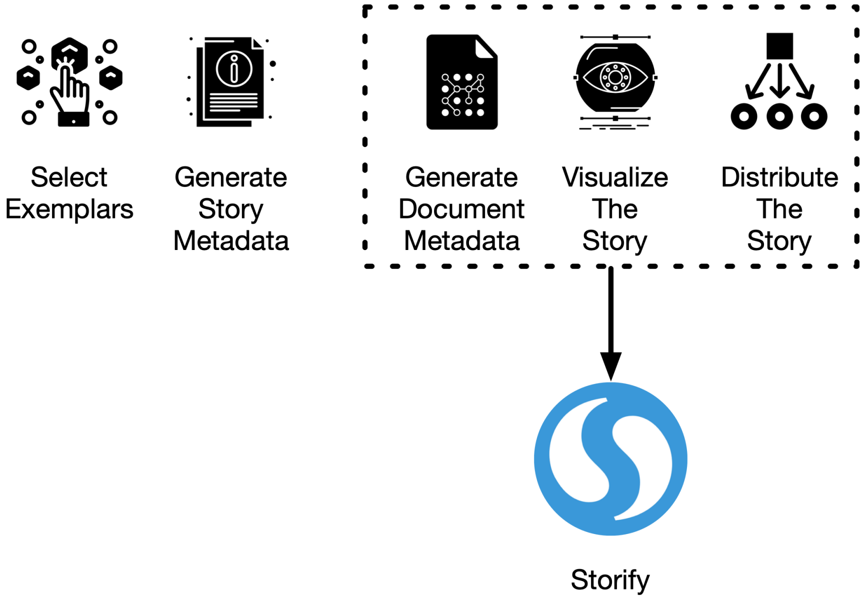


Fig. 137. Storify combined the storytelling processes of generating document metadata, visualizing the story, and distributing it. Its API allowed AlNoamany to augment its surrogates with additional story and document metadata.

card metadata is for mementos of news articles. We discover that this metadata was quickly adopted when Facebook first released social card standards in 2010. However, mementos captured before 2010 will require some automated solution for generating the document metadata necessary to produce cards. Mementos captured prior to 2010 represent 150 billion pages in the Internet Archive [125] and content from 1,486 Archive-It collections. We analyze the authors' behavior with respect to card metadata fields in order to describe how to emulate humans' behavior for generating descriptions and selecting striking images. Because we are choosing the "best" image from those available in a document, automating the selection of striking images is yet another case of selecting exemplars. We close by introducing MementoEmbed (Chapter 5.4), an archive-aware document metadata and surrogate service that incorporates these lessons.

5.1 HOW EXISTING PLATFORMS DO NOT SATISFY OUR STORYTELLING PROCESS FOR MEMENTOS

Storify was the initial tool applied by AlNoamany's work to gather document metadata, visualize, and distribute web archive stories. Storify's definition of **story** influenced the visual summaries that we developed. Its application of surrogates as a summarization technique alongside additional metadata made it a popular tool [328] for journalists [78, 294, 295, 296] and educators

[20, 96, 102, 201]. Livefyre acquired Storify in 2013 [112] and Adobe acquired Livefyre in 2016 [283]. Adobe created Storify 2 as a content creation platform so authors could create and publish stories to their own web sites, but removed the public distribution capabilities provided by the original Storify [116]. Adobe then shut the original Storify down in 2018 [43, 282, 289].

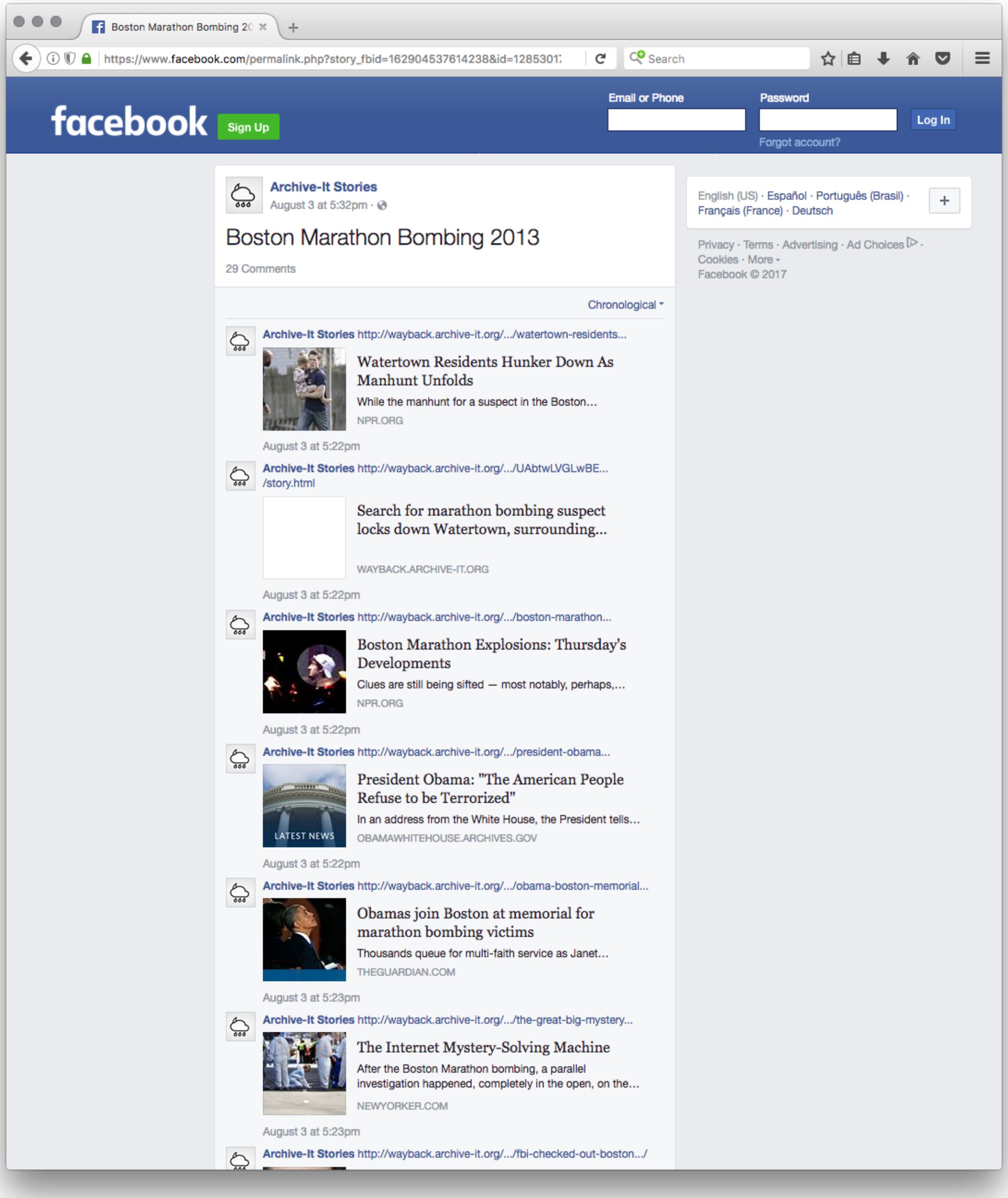
The loss of Storify was problematic for the use cases embodied by our characters. As demonstrated in Figure 137, Storify combined the storytelling processes of generating document metadata, visualizing the story, and distributing it. Without it, Natasha cannot view the stories created by others, and Elbert cannot promote his collections. We needed a replacement so we could still tell stories, and Storify 2, as part of Adobe's Livefyre, had evolved into a product requiring a local client and manual interaction; thus, it no longer met our needs.

We evaluated 55 platforms in 2017 with respect to their ability to select exemplars, generate document metadata, visualize, and distribute summaries of mementos [140, 156]. A more complete discussion on that study exists in work by Jones et al. [140, 156]. Here we summarize the results. We cannot see the internal representations of document metadata so we are forced to judge the tool's ability to generate this metadata based on the surrogates that the tool creates. Even though this chapter focuses on the generating document metadata process, there are many tools that promise to address other parts of the storytelling process and they fail with mementos. Of these, 30 promise to select exemplars in some way, but none do so for mementos. Many will not accept exemplars from an outside source, meaning that we cannot take advantage of their ability to complete other storytelling processes. We could not evaluate seven because they are no longer in service. Here we highlight some of the few finalists and why they were not suitable. Though this survey is four years old, nothing rivaling Storify has emerged in the commercial market since.

5.1.1 FACEBOOK

With 1.88 billion daily active users [85], Facebook is the most popular social media tool. Facebook supports social cards in posts and comments. Facebook also supports creating albums of photos but not of posts. Posts contain comments, however. As shown in Figure 138, in order to generate a series of social cards as a story, we created a post that contained the general collection metadata (e.g., title, link to collection) and supplied each URI-M in the story as a separate comment. Ultimately, we had to supply our own story metadata.

As seen in Figure 139, Facebook occasionally fails to generate social cards for links. The problem is not that Facebook is blocking web archives. Facebook can create cards for web archives by reading the OGP metadata found in the memento. In Figure 140, both comments contain the



Boston Marathon Bombing 20
https://www.facebook.com/permalink.php?story_fbid=162904537614238&id=1285301
Search
facebook
Sign Up
Email or Phone
Password
Log In
Forgot account?

Archive-It Stories
August 3 at 5:32pm ·

Boston Marathon Bombing 2013

29 Comments

Archive-It Stories http://wayback.archive-it.org/.../watertown-residents...

Watertown Residents Hunker Down As Manhunt Unfolds
While the manhunt for a suspect in the Boston...
NPR.ORG

August 3 at 5:22pm

Archive-It Stories http://wayback.archive-it.org/.../UAbtwLVGLwBE.../story.html

Search for marathon bombing suspect locks down Watertown, surrounding...
WAYBACK.ARCHIVE-IT.ORG

August 3 at 5:22pm

Archive-It Stories http://wayback.archive-it.org/.../boston-marathon...

Boston Marathon Explosions: Thursday's Developments
Clues are still being sifted — most notably, perhaps,...
NPR.ORG

August 3 at 5:22pm

Archive-It Stories http://wayback.archive-it.org/.../president-obama...

President Obama: "The American People Refuse to be Terrorized"
In an address from the White House, the President tells...
OBAMAWHITEHOUSE.ARCHIVES.GOV

August 3 at 5:22pm

Archive-It Stories http://wayback.archive-it.org/.../obama-boston-memorial...

Obamas join Boston at memorial for marathon bombing victims
Thousands queue for multi-faith service as Janet...
THEGUARDIAN.COM

August 3 at 5:23pm

Archive-It Stories http://wayback.archive-it.org/.../the-great-big-mystery...

The Internet Mystery-Solving Machine
After the Boston Marathon bombing, a parallel investigation happened, completely in the open, on the...
NEWYORKER.COM

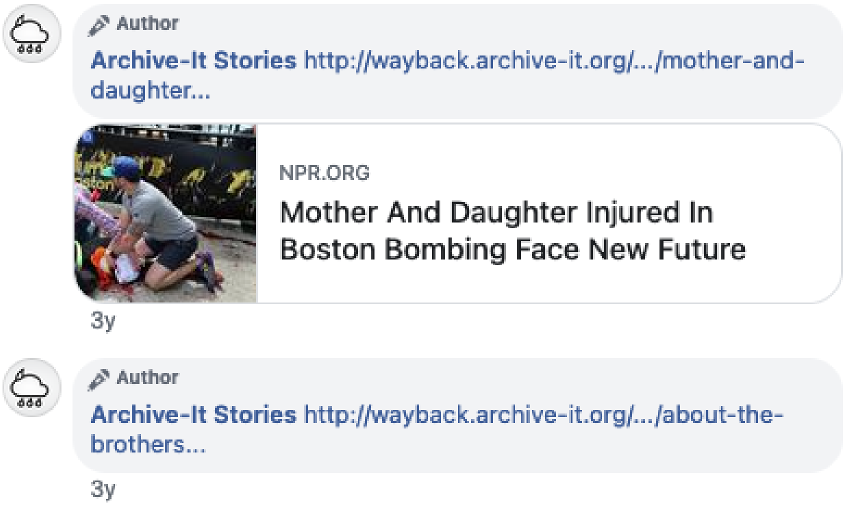
August 3 at 5:23pm

Archive-It Stories http://wayback.archive-it.org/.../fbi-checked-out-boston.../

English (US) · Español · Português (Brasil) · Français (France) · Deutsch

Privacy · Terms · Advertising · Ad Choices · Cookies · More
Facebook © 2017

Fig. 138. We can create Facebook stories through comments. (Screenshot taken in 2017.)



Author

Archive-It Stories http://wayback.archive-it.org/.../mother-and-daughter...

NPR.ORG

Mother And Daughter Injured In Boston Bombing Face New Future

3y

Author

Archive-It Stories http://wayback.archive-it.org/.../about-the-brothers...

3y

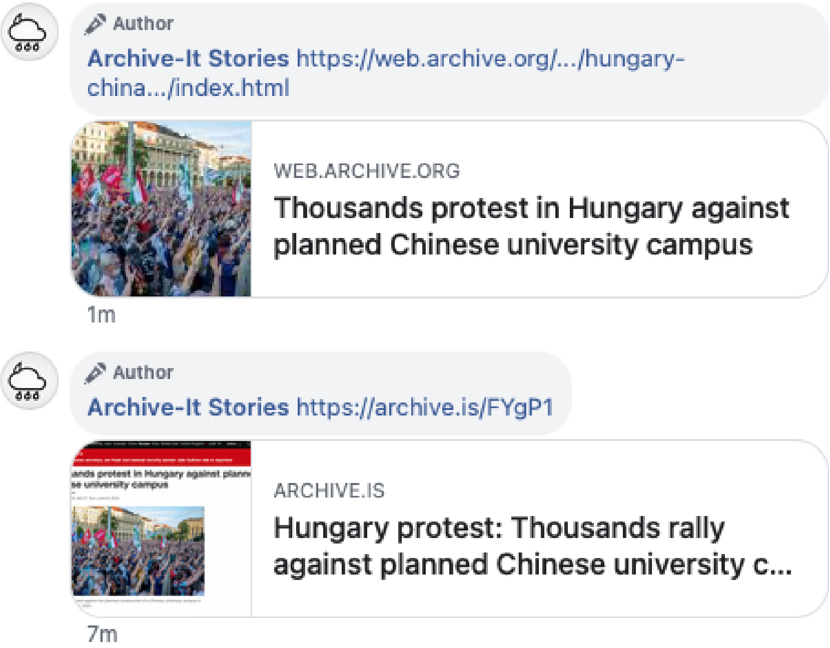
Fig. 139. A screenshot of two Facebook comments. The URI-M above generated a social card, but the URI-M below did not. (Screenshot taken in June 2021.)

same URI-R stored in the Internet Archive (top) and Archive.Today (bottom) captured on 2021-06-07 16:57:36 UTC and 2021-06-07 17:43:21, respectively. The Internet Archive presents the original resource OGP metadata for Facebook. Archive.Today generates its own values for OGP metadata, including changing the striking image to a browser thumbnail representing how the original resource looked when it was captured.

Facebook is unreliable for generating document metadata for mementos and incapable of visualizing our stories. Additionally, Facebook does not keep the comments in order once they are submitted. A story ordered by Hypercane through some feature, like publication date, will be reordered by Facebook, thus losing the meaning we sought to impart. We could employ the Facebook API to update such comments with a photo and a snippet, if necessary. Providing additional images is impossible, as Facebook posts and comments will not generate a social card if the post/-comment already has an image. Also, at that point, Facebook would only be distributing our story, and we still will not have solved the ordering problem.

5.1.2 TWITTER

Twitter has 186 million daily active users worldwide [307]. Twitter also supports social cards in posts. For Twitter, comments are also posts, as every interaction on Twitter is a *Tweet*. Users can group tweets together to create their own stories through Twitter Moments. All tools we reviewed required that the user title the story in some way, and Twitter Moments required that the



Author

Archive-It Stories https://web.archive.org/.../hungary-china.../index.html

WEB.ARCHIVE.ORG

Thousands protest in Hungary against planned Chinese university campus

1m

Author

Archive-It Stories https://archive.is/FYgP1

ARCHIVE.IS

Hungary protest: Thousands rally against planned Chinese university c...

7m

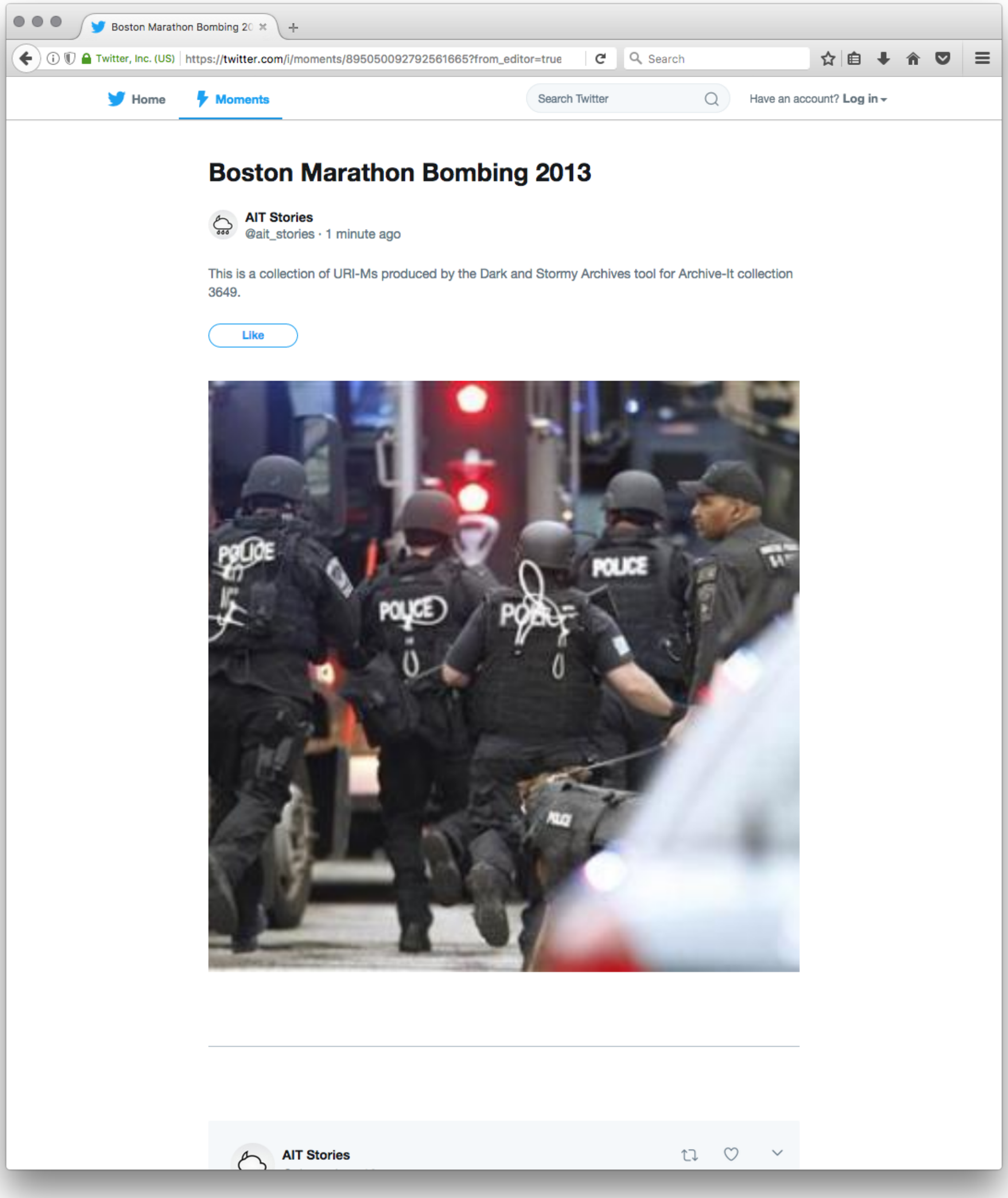
Fig. 140. Facebook can create cards for web archives by reading the OGP metadata found in the memento. Both comments contain the same URI-R stored in the Internet Archive (top) and Archive.Today (bottom). (Screenshot taken in June 2021.)

user upload an image separately in order to create a story, as shown in Figure 141. This image serves as the striking image for the collection. Twitter compels the user to supply a description. Thus, while a user can supply story metadata, the tool will not generate it.

Twitter does not reliably generate social cards for URI-Ms. Figure 142 displays a screenshot of a Twitter Moment containing mementos. Twitter displays the individual URI-Ms tweets with no additional visualization. We could use the Twitter API to add images and additional text (up to 280 characters) to supplement these tweets. At that point, we are building our own surrogates out of tweets and just using Twitter for distribution.

5.1.3 TUMBLR

Tumblr had 319 million visitors in May 2021 [285]. Tumblr is a complex social media tool supporting many different types of posts. A user selects which type of post they desire and then supplies the necessary data or metadata. For example, if a user wanted to generate something like a Facebook post or a Twitter tweet, they would choose *Text*. The post type *Link* produces a social card for the supplied link. In addition to the social card generated by Tumblr, a user can adorn

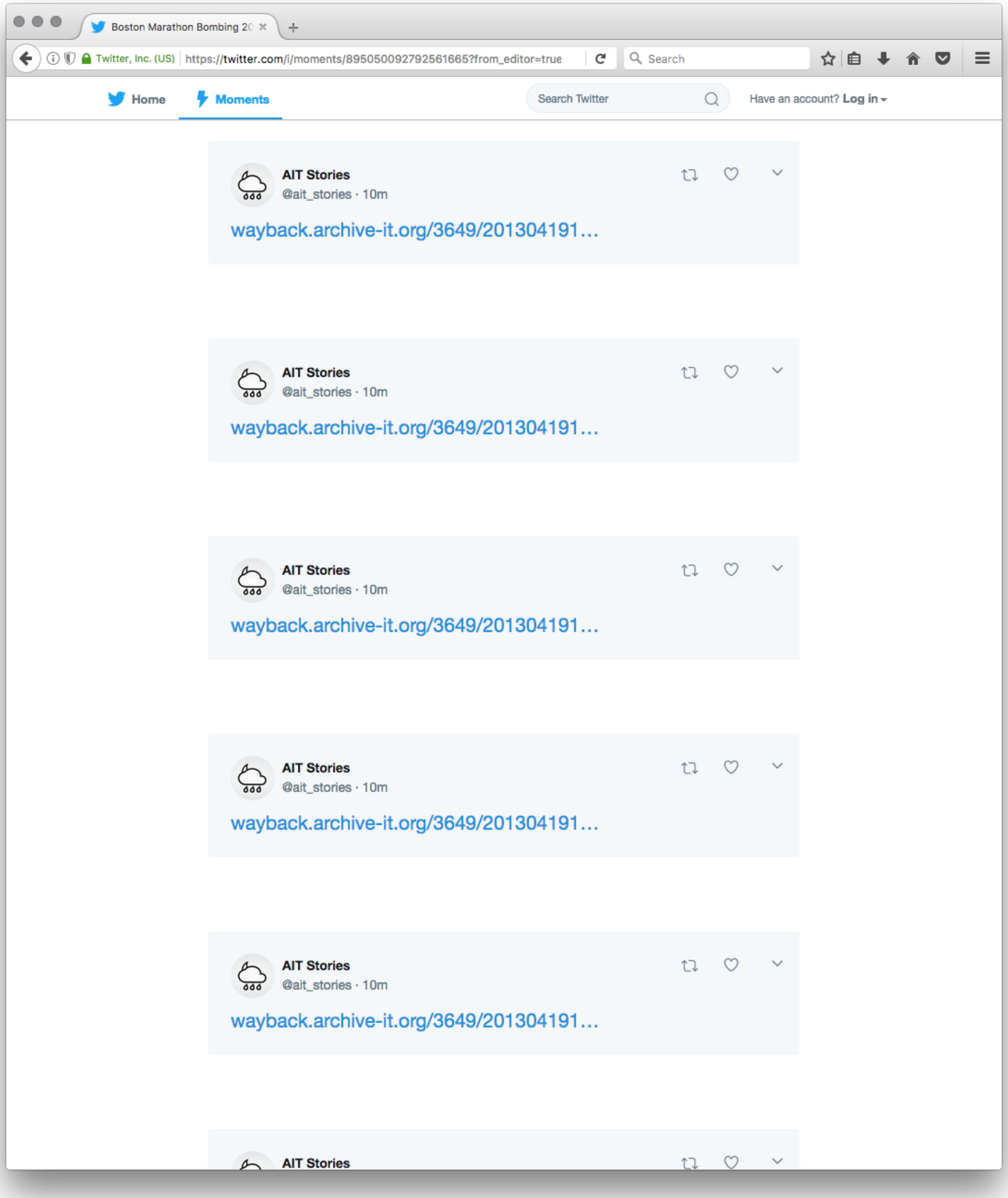

Boston Marathon Bombing 2(
Twitter, Inc. (US)
https://twitter.com/i/moments/895050092792561665?from_editor=true
Search
Home
Moments
Search Twitter
Have an account? Log in
Like
AIT Stories

Boston Marathon Bombing 2013

AIT Stories
@ait_stories · 1 minute ago

This is a collection of URI-Ms produced by the Dark and Stormy Archives tool for Archive-It collection 3649.

Fig. 141. We can create stories on Twitter through Twitter Moments. (Screenshot taken in 2017.)


Boston Marathon Bombing 2(
Twitter, Inc. (US)
https://twitter.com/i/moments/895050092792561665?from_editor=true
Search
Home
Moments
Search Twitter
Have an account? Log in

AIT Stories
@ait_stories · 10m
wayback.archive-it.org/3649/201304191…

AIT Stories
@ait_stories · 10m
wayback.archive-it.org/3649/201304191…

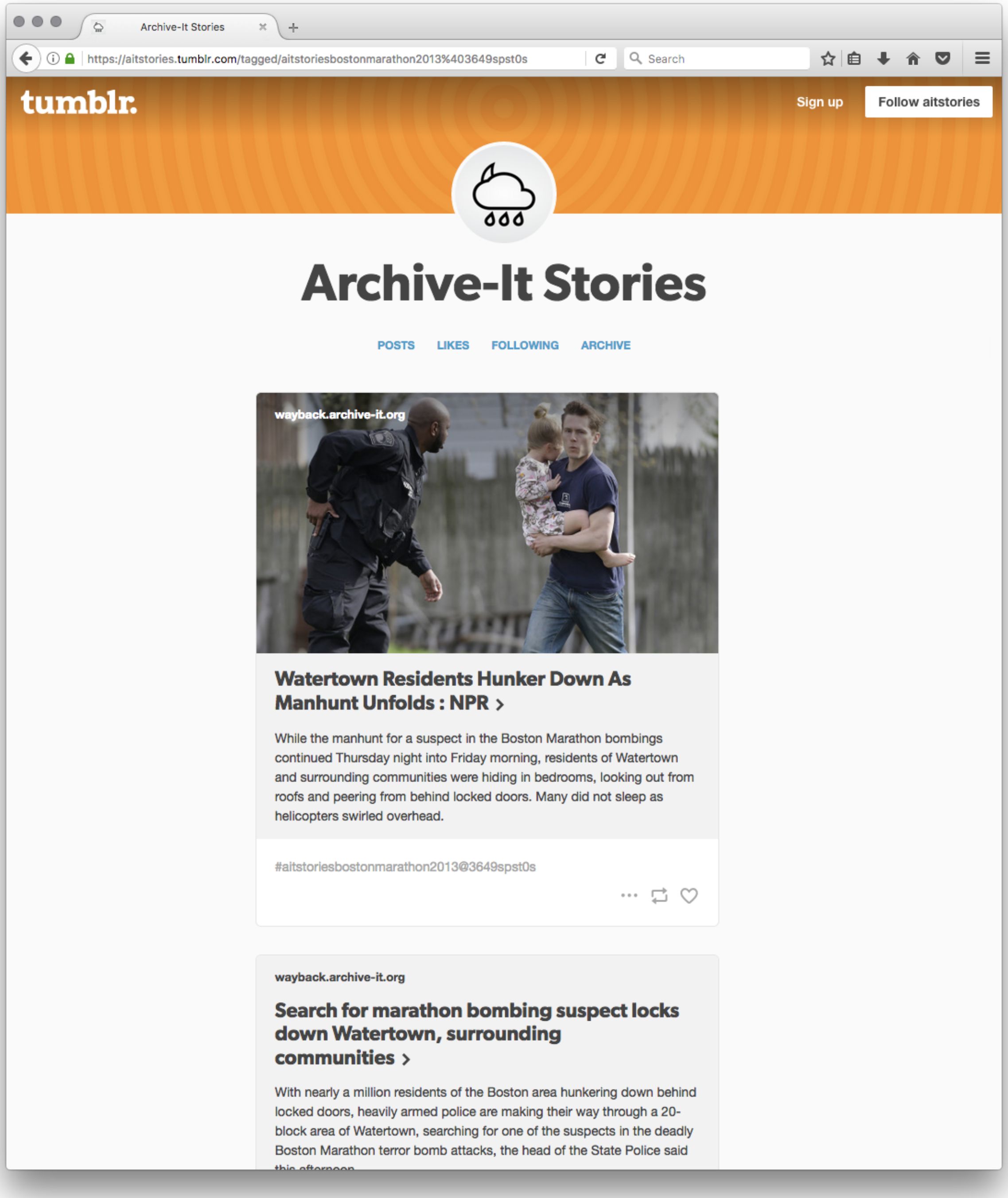
AIT Stories
@ait_stories · 10m
wayback.archive-it.org/3649/201304191…

AIT Stories
@ait_stories · 10m
wayback.archive-it.org/3649/201304191…

AIT Stories
@ait_stories · 10m
wayback.archive-it.org/3649/201304191…

AIT Stories

Fig. 142. Tweets containing Archive-It URI-Ms do not render social cards. (Screenshot taken in 2017.)



Archive-It Stories

Watertown Residents Hunker Down As Manhunt Unfolds : NPR ›

While the manhunt for a suspect in the Boston Marathon bombings continued Thursday night into Friday morning, residents of Watertown and surrounding communities were hiding in bedrooms, looking out from roofs and peering from behind locked doors. Many did not sleep as helicopters swirled overhead.

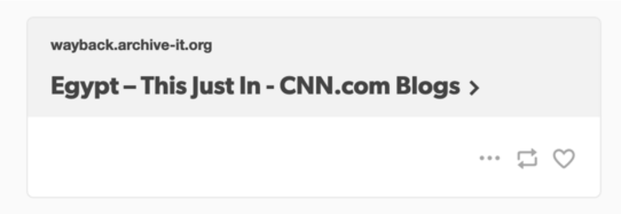
#aitstoriesbostonmarathon2013@3649spst0s

wayback.archive-it.org

Search for marathon bombing suspect locks down Watertown, surrounding communities ›

With nearly a million residents of the Boston area hunkering down behind locked doors, heavily armed police are making their way through a 20-block area of Watertown, searching for one of the suspects in the deadly Boston Marathon terror bomb attacks, the head of the State Police said this afternoon.

Fig. 143. We can create Tumblr stories by bounding posts together with a hashtag. (Screenshot taken in 2017.)



wayback.archive-it.org

Egypt – This Just In - CNN.com Blogs >

Fig. 144. Tumblr does not reliably produce full social cards for mementos, as seen in this example. (Screenshot taken in 2021.)

the *Link* post with an additional photo, video, or text. To create a story on Tumblr (Figure 143), we apply hashtags. Tumblr confines the hashtags to a specific blog controlled by that blog's user; hence posts outside of the blog do not intrude into the collection, as would happen with hashtags on Twitter or Facebook.

Unfortunately, as shown in Figure 144, Tumblr has similar problems to other tools when trying to generate surrogates for mementos. In this example, this *CNN* memento produces a surrogate containing a title, but nothing else. It also misattributes CNN's content to wayback.archive-it.org, potentially confusing readers as to the actual source of this article. Tumblr has an API that we can apply to fix some of these issues. If a surrogate lacks an image, or if we need to supply additional text, the post can be updated appropriately, but, like with Facebook and Twitter, we are again building our own surrogate service at that point.

5.1.4 *STORIES* BUT NOT STORIFY STORYTELLING

Since Storify, some platforms have co-opted the term *story* to mean something impermanent. This impermanence is contrary to the original definition employed by Storify and, more importantly, contrary to our preservation goals. We want the stories to be artifacts that themselves can be archived. Snapchat [288], Instagram Stories [124], Facebook Stories [86], and Twitter Fleets [305] all disappear within 24-hours of posting. Twitter has even announced that the Fleets capability will not longer be available after August 3, 2021 [44]. This impermanence makes those platforms unsuitable for our needs. We want stories that are persistent, reusable, and archivable.

5.1.5 SUMMARY

After reviewing 55 platforms, we could not find a service that would reliably generate document metadata and visualize it as a surrogate for a memento. Some tools are a poor fit for what we are trying to do, but others that seem like a good fit still require that we generate our own document metadata. Without a reliable target visualization, we conducted a user study to determine which document metadata, through surrogates, helped users best understand web archive collections.

5.2 EVALUATING DOCUMENT SURROGATES FOR UNDERSTANDING

Platforms build surrogates from metadata. Metadata can consist of concepts like title, striking image, or browser thumbnail. Even though a browser thumbnail is typically considered a surrogate in its own right, it does convey information about the document. Here we consider it metadata just like a title or description. Even though a striking image is a multimedia object, authors still supply it through a metadata field, as mentioned in Chapter 2.4. If we understand which surrogate works best to understand a collection, we know what document metadata fields we need to generate for storytelling.

This section considers six different types of surrogates and how well they might work to convey an understanding of a collection. We compare the existing Archive-It surrogates, thumbnails of page screenshots, social cards, and three combinations of social cards and thumbnails. We hypothesize that surrogates with more information drawn from the source document produce better results, both in terms of time and understanding. We report the results of a user study that analyzed how users interacted with each of these surrogate types and determined that social cards a probably the best surrogate type for understanding web archive collections.

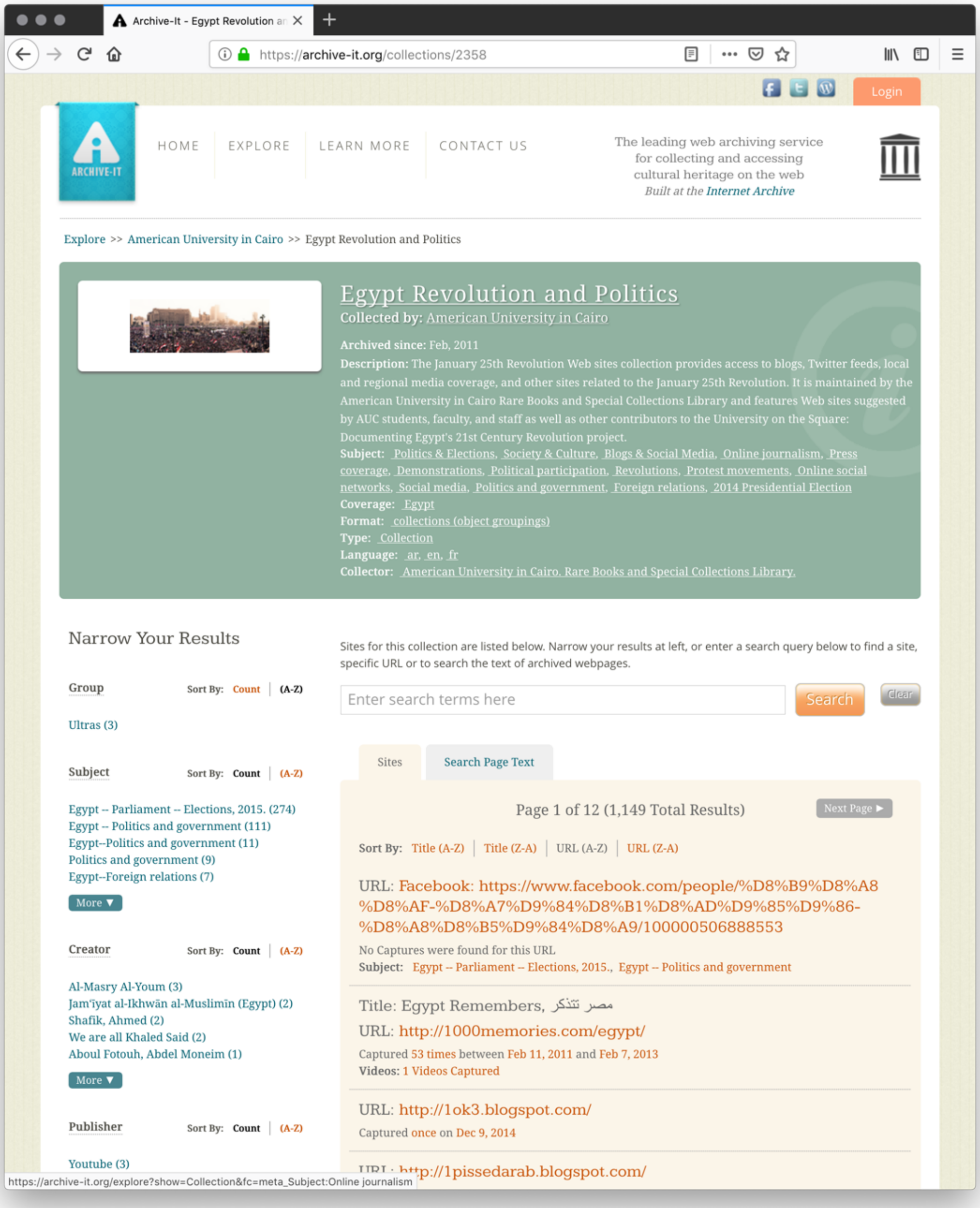
The existing Archive-It surrogate is very sparse, providing very little information for the reader. Before moving on to the user study, we first examine what information a user like Natasha might glean from an Archive-It surrogate.

5.2.1 EVALUATION OF ARCHIVE-IT SURROGATES

Before discussing the results of evaluating different surrogates against each other, we first quantify the information available from Archive-It surrogates. We used our AIU Python package [153] to collect the metadata of 5,857 public Archive-It collections in March 2019. Our goal was to understand the amount of information available with Archive-It surrogates.

Archive-It surrogates contain two sources of metadata: the archiving process and the original resource. A seed's **minimal Archive-It surrogate** contains information from the archiving process: the seed's URI, the dates of the first and last memento, and the number of mementos available. The `title` field is an example of metadata derived from the original resource. An archivist may manually extract the title when they add the seed to the collection or manually add it later.

In addition to sometimes being nonexistent, metadata can also be inconsistently applied among



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

Explore >> American University in Cairo >> Egypt Revolution and Politics

Egypt Revolution and Politics

Collected by: American University in Cairo

Archived since: Feb, 2011

Description: The January 25th Revolution Web sites collection provides access to blogs, Twitter feeds, local and regional media coverage, and other sites related to the January 25th Revolution. It is maintained by the American University in Cairo Rare Books and Special Collections Library and features Web sites suggested by AUC students, faculty, and staff as well as other contributors to the University on the Square: Documenting Egypt's 21st Century Revolution project.

Subject: Politics & Elections, Society & Culture, Blogs & Social Media, Online journalism, Press coverage, Demonstrations, Political participation, Revolutions, Protest movements, Online social networks, Social media, Politics and government, Foreign relations, 2014 Presidential Election

Coverage: Egypt

Format: collections (object groupings)

Type: Collection

Language: ar, en, fr

Collector: American University in Cairo. Rare Books and Special Collections Library.

Narrow Your Results

Group Sort By: Count | (A-Z)

Ultras (3)

Subject Sort By: Count | (A-Z)

Egypt -- Parliament -- Elections, 2015. (274)
Egypt -- Politics and government (111)
Egypt--Politics and government (11)
Politics and government (9)
Egypt--Foreign relations (7)

Creator Sort By: Count | (A-Z)

Al-Masry Al-Youm (3)
Jamʻīyat al-Ikhwān al-Muslimīn (Egypt) (2)
Shafik, Ahmed (2)
We are all Khaled Said (2)
Aboul Fotouh, Abdel Moneim (1)

Publisher Sort By: Count | (A-Z)

Youtube (3)

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

Page 1 of 12 (1,149 Total Results)

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

URL: Facebook: https://www.facebook.com/people/%D8%B9%D8%A8%D8%AF-%D8%A7%D9%84%D8%B1%D8%AD%D9%85%D9%86-%D8%A8%D8%B5%D9%84%D8%A9/100000506888553

No Captures were found for this URL

Subject: Egypt -- Parliament -- Elections, 2015., Egypt -- Politics and government

Title: Egypt Remembers, مصر تتذكر

URL: http://1000memories.com/egypt/

Captured 53 times between Feb 11, 2011 and Feb 7, 2013

Videos: 1 Videos Captured

URL: http://1ok3.blogspot.com/

Captured once on Dec 9, 2014

URL: http://1pissedarab.blogspot.com/

https://archive-it.org/explore?show=Collection&fc=meta_Subject:Online journalism

Fig. 145. This screenshot demonstrates multiple Archive-It surrogates with differing amounts of metadata.

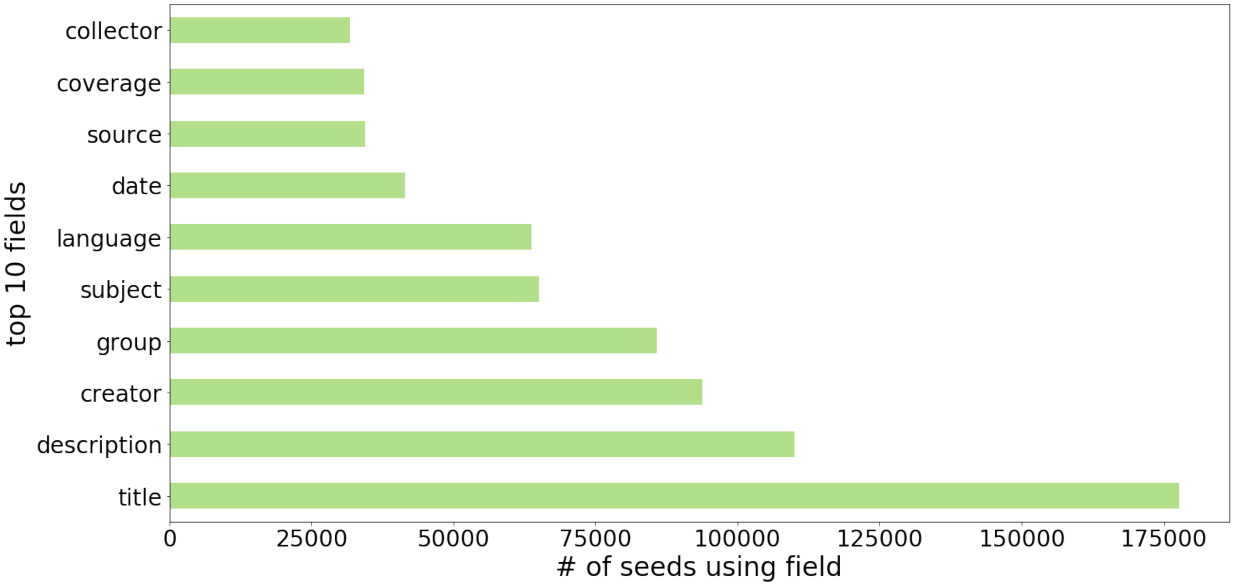


Fig. 146. The top 10 metadata fields used by seeds in Archive-It.

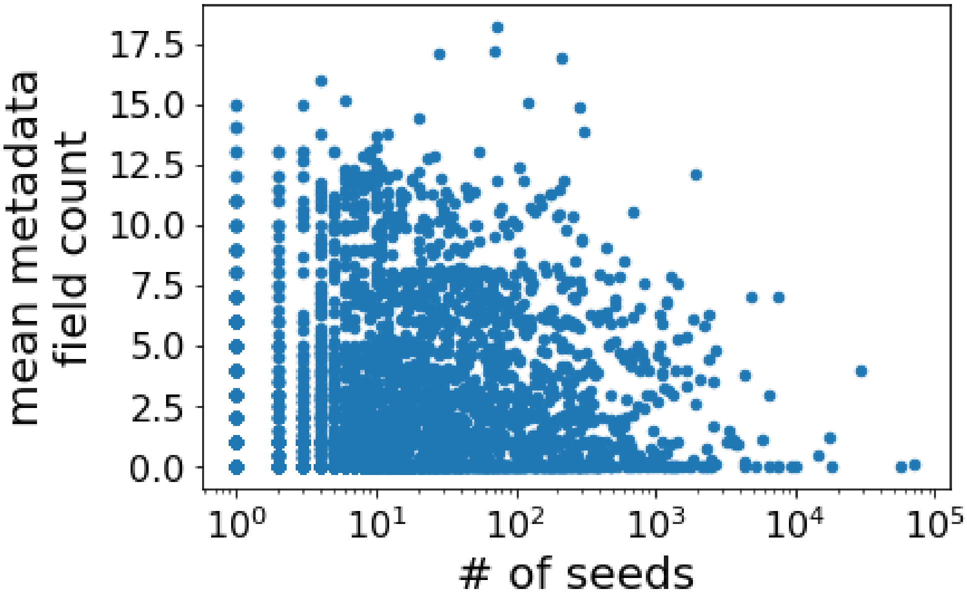


Fig. 147. A plot of mean metadata field count per collection vs. the number of seeds in an Archive-It collection for 5,857 public Archive-It collections.

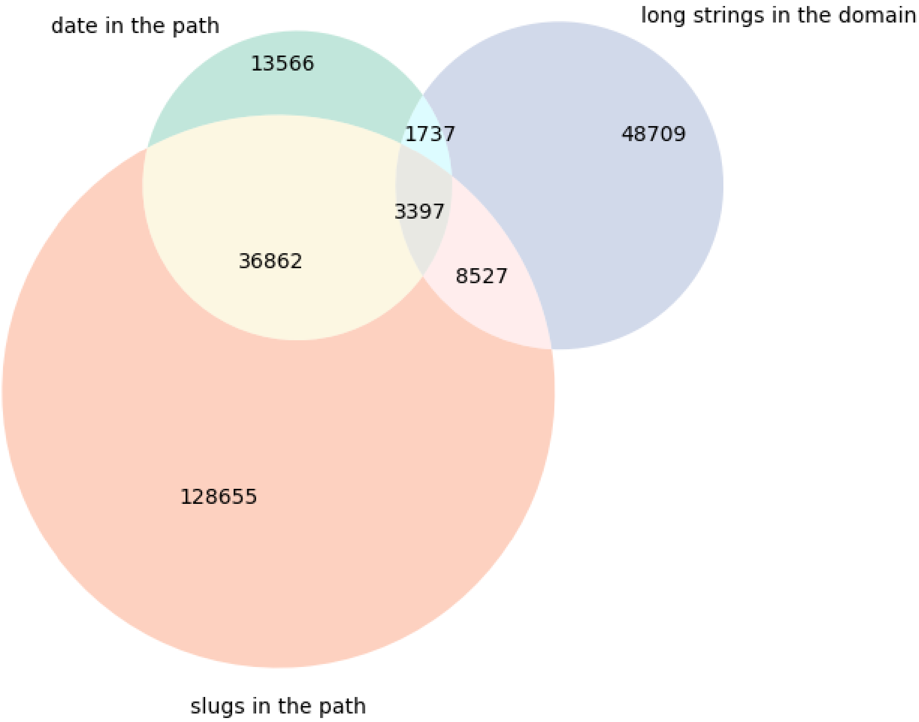


Fig. 148. Overlap of different information classes of Archive-It seed URIs.

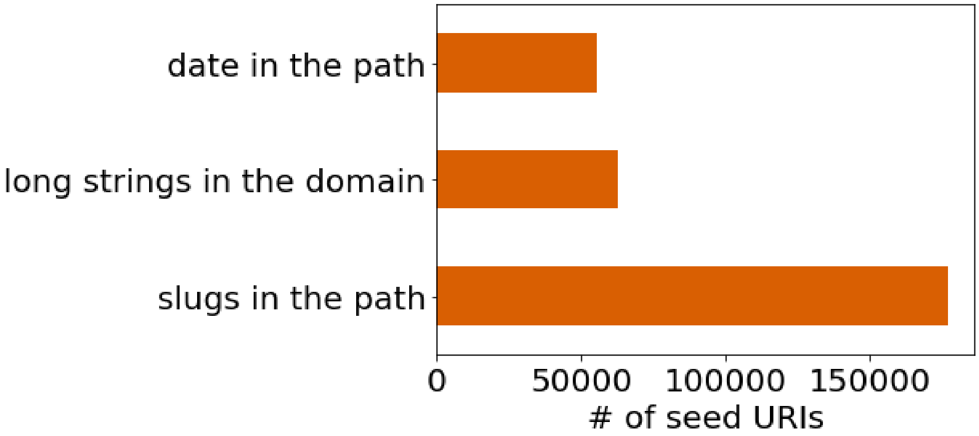


Fig. 149. A bar chart quantifying the potential information detected in Archive-It seed URIs.

surrogates, as seen in the screenshot in Figure 145. From the 602,944 seeds gathered, Archive-It represents 329,178 (54.60%) of them with the minimal Archive-It surrogate. These seeds convey only the URI and information from the archiving process. Suppose the archivist provides metadata fields, such as a `title` or `description`. In that case, the Archive-It surrogate begins to resemble surrogates typically found in search engine results. The two fields together are used on 75,575/602,944 seeds, meaning that 12.53% of Archive-It seeds contain the same metadata fields as a Google surrogate. Figure 146 shows the top ten fields in use by any seed, regardless of the number of fields per seed. In this case `title` is the most widely used metadata field, being present in 177,680/602,944 (29.5%) seeds. The `description` field is in use by 110,065/602,944 (18.3%) seeds.

Some collections, such as *Government of Canada Publications* (ID 3572), have hundreds of thousands of seeds, making the addition of metadata a costly proposition in terms of manual time and effort. Does this cost affect the behavior of the curator? For each collection, we counted how many metadata fields were applied to all seeds in the collection, regardless of size. We then divided the number of fields counted by the number of seeds to produce the mean metadata field count per collection. Figure 147 (also seen as Figure 5 in Chapter 1) shows a point for each collection where the y-axis is the mean metadata field count and the x-axis is the number of seeds in $\log_{10}$ scale. This graph displays a pattern whereby an increase in the number of seeds corresponds to a decrease in the number of metadata fields used to describe those seeds. This result matches our intuition that because each metadata field requires some effort to maintain, the curator supplies fewer metadata fields as the number of seeds increases. The mean metadata field count for 3,096/5,867 (52.86%) collections is 0, again indicating that most collections only contain minimal Archive-It surrogates.

These results appear to support our intuition that many of the Archive-It surrogates contain little information, but a user can still gather information from a seed URI. As noted, there are many collections about the same topic, so there is some overlap in the choice of seed URIs by different curators. There are 14,179 repeated seed URIs across Archive-It collections, meaning that only 588,749 unique seed URIs exist in Archive-It. From those seed URIs, we employed regular expressions from Alkwai et al. [5, 6] to detect different forms of crude information available in the seed URIs from Archive-It. As shown in Figure 148, a seed URI can belong to all three information classes. Figure 149 displays the results of this analysis. Long strings of five or more characters separated by an underscore or other character may indicate the presence of phrases or sentences. We discovered 62,370/588,749 (10.59%) URIs contained long strings in their domain names. Our regular expressions only detected dates in the paths of 55,924/588,749 (9.44%) seed

URIs. Such dates are typically the publication dates of blog posts or news articles. Dates can provide the viewer with a concept of aboutness for the time period of a collection. A slug is a special type of long string indicating a shortened title for an article. Note that a single URI can belong to all three categories. Slugs form the largest category with 177,441/588,749 (30.14%) seed URIs containing slugs in their path. Figure 149 displays the results of this analysis. These results indicate that, despite missing metadata, a user like Natasha can still glean information from the URIs found in Archive-It surrogates.

5.2.2 A USER STUDY COMPARING SURROGATES

Methodology

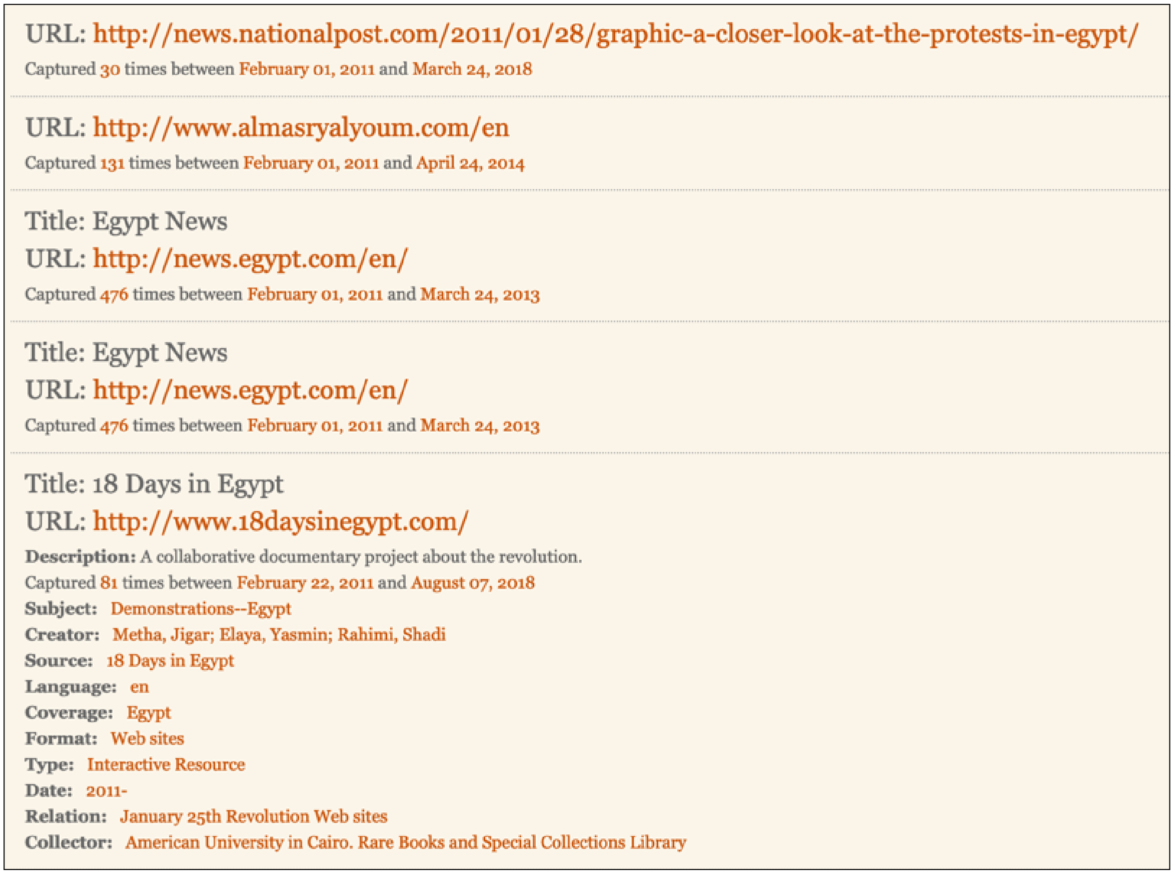
In January 2019, we presented 120 Mechanical Turk (MT) participants with a link to a survey hosted at Old Dominion University. We produced four stories represented by six different surrogates for 24 different combinations of surrogates and stories. This gave us five participants per story-surrogate combination, providing 20 participants per surrogate type. The MT participants were required to have the Master Turker qualification and an acceptance rate of greater than 95%. To control for the effects of learning [177], we employed UniqueTurker[1] to ensure that the same participant did not provide results for multiple surveys. We paid each participant $0.50.

After reading the instructions, each participant had 30 seconds to view a story using a given surrogate. The task asked them a question about what they had just seen. This process is similar to the sentence verification task used in reading comprehension studies [167]. As is common practice for externally hosted surveys on MT, once they submitted their results, they were given a completion code for the MT HIT to map their results to those collected by our survey.

As a source of stories to display to the participants, we selected four stories from AlNoamany's 2016 dataset [13]. Each story consists of collection exemplars selected by a human curator to describe their collection. Table 27 shows the details of the complete dataset. Some collections have mementos that are no longer available, possibly because the archivist removed them. Some collections also have mementos that produce poor-quality thumbnails. If a thumbnail failed to contain at least a heading describing some of the content within the memento, we considered it to be of poor quality. The last column in this table lists the percentage of the story that produced good-quality surrogates.

Our four selections represent a variety of structural and semantic considerations. *Occupy*

[1] `http://uniqueturker.myleott.com`



URL: http://news.nationalpost.com/2011/01/28/graphic-a-closer-look-at-the-protests-in-egypt/
Captured 30 times between February 01, 2011 and March 24, 2018

URL: http://www.almasryalyoum.com/en
Captured 131 times between February 01, 2011 and April 24, 2014

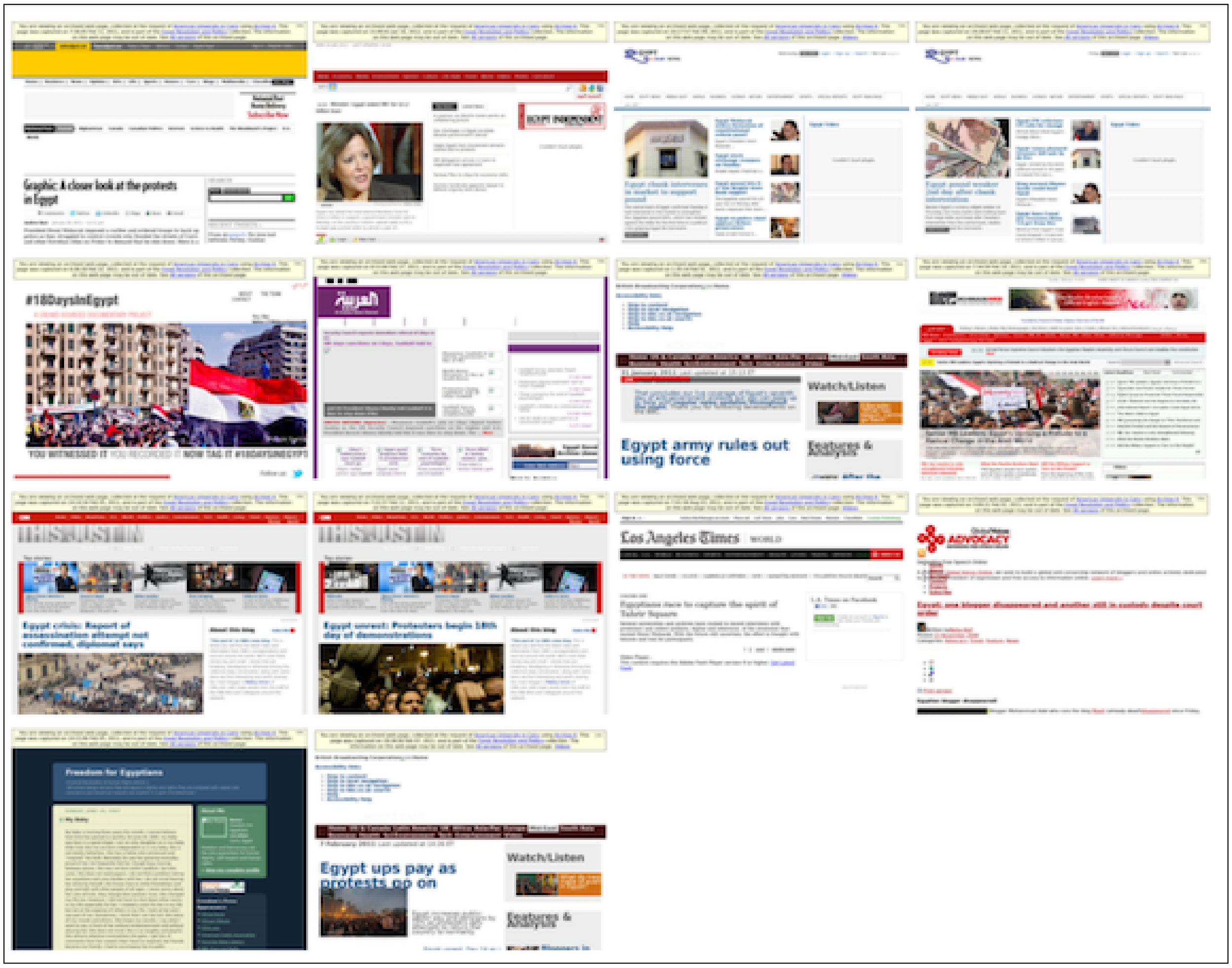
Title: Egypt News
URL: http://news.egypt.com/en/
Captured 476 times between February 01, 2011 and March 24, 2013

Title: Egypt News
URL: http://news.egypt.com/en/
Captured 476 times between February 01, 2011 and March 24, 2013

Title: 18 Days in Egypt
URL: http://www.18daysinegypt.com/
Description: A collaborative documentary project about the revolution.
Captured 81 times between February 22, 2011 and August 07, 2018
Subject: Demonstrations--Egypt
Creator: Metha, Jigar; Elaya, Yasmin; Rahimi, Shadi
Source: 18 Days in Egypt
Language: en
Coverage: Egypt
Format: Web sites
Type: Interactive Resource
Date: 2011-
Relation: January 25th Revolution Web sites
Collector: American University in Cairo. Rare Books and Special Collections Library

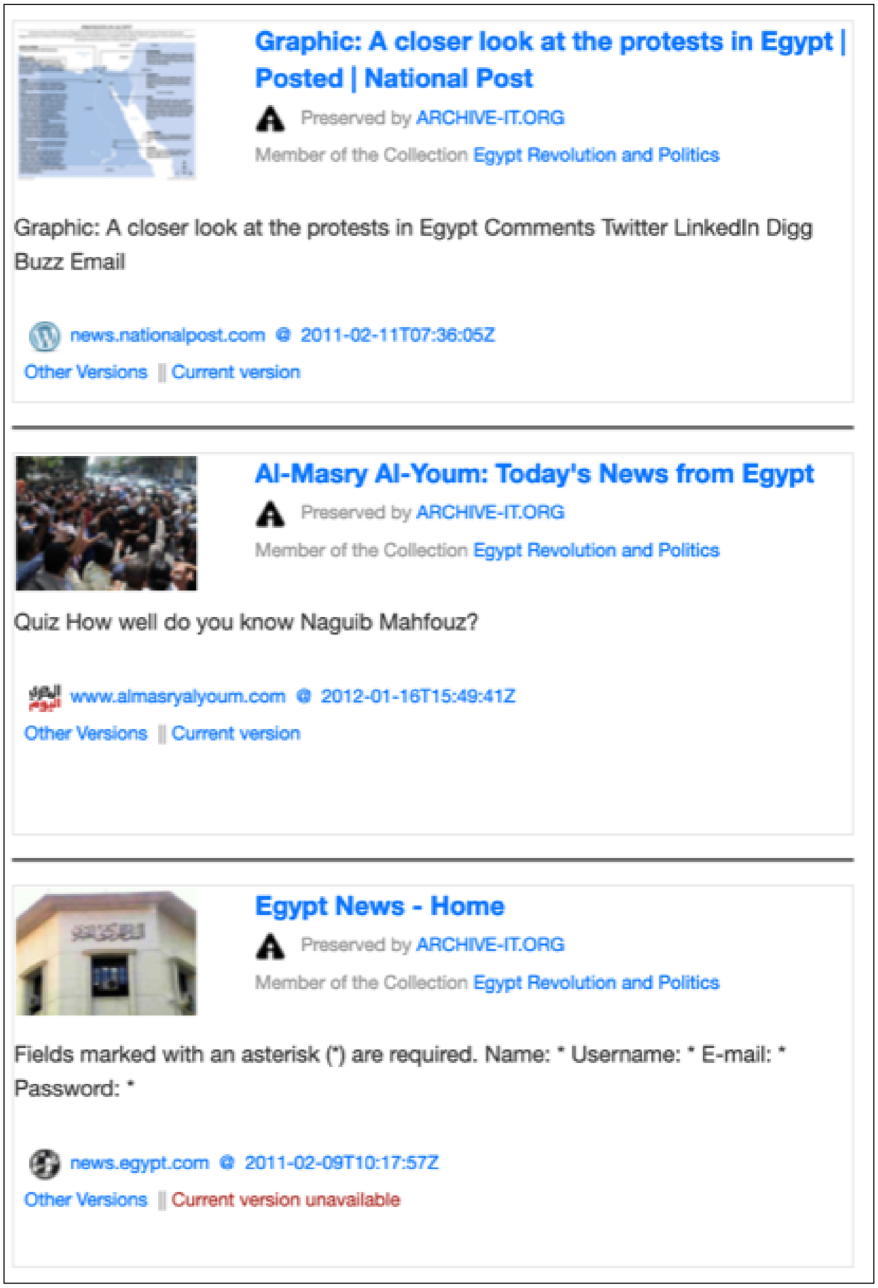
(a) Archive-It Facsimile

Fig. 150. Screenshots of parts of stories constructed using different surrogates from a sample of 16 mementos selected from the 80,484 mementos found in collection *Egypt Revolution and Politics* (ID 2358).


Graphic: A closer look at the protests in Egypt
EGYPT INDEPENDENT
#18DaysInEgypt
YOU WITNESSED IT YOU RECORDED IT NOW TAG IT #18DAYSINEGYPT
Watch/Listen
Egypt army rules out using force
Features & Analysis
THIS JUST IN
Egypt crisis: Report of assassination attempt not confirmed, diplomat says
Egypt unrest: Protesters begin 18th day of demonstrations
Los Angeles Times WORLD
ADVOCACY
Freedom for Egyptians
Egypt ups pay as protests go on

(b) Browser thumbnail

Fig. 150. (Continued) Screenshots of parts of stories constructed using different surrogates from a sample of 16 mementos selected from the 80,484 mementos found in collection *Egypt Revolution and Politics* (ID 2358).



Graphic: A closer look at the protests in Egypt | Posted | National Post

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

Graphic: A closer look at the protests in Egypt Comments Twitter LinkedIn Digg Buzz Email

news.nationalpost.com @ 2011-02-11T07:36:05Z

Other Versions || Current version

Al-Masry Al-Youm: Today's News from Egypt

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

Quiz How well do you know Naguib Mahfouz?

www.almasryalyoum.com @ 2012-01-16T15:49:41Z

Other Versions || Current version

Egypt News - Home

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

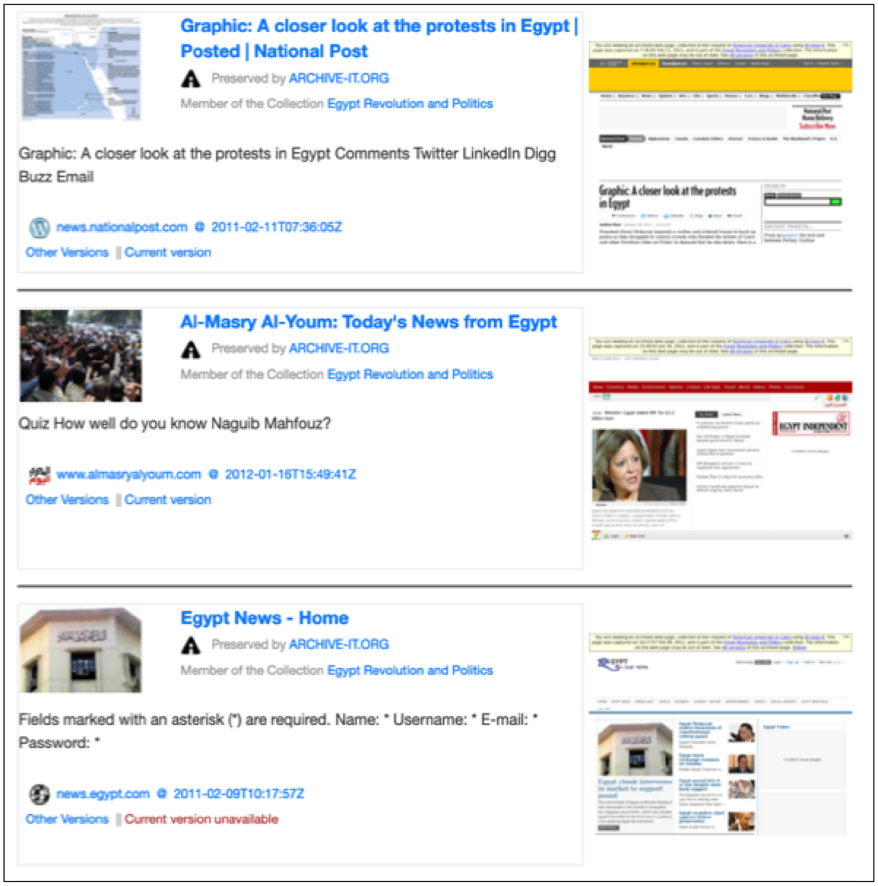
Fields marked with an asterisk (*) are required. Name: * Username: * E-mail: * Password: *

news.egypt.com @ 2011-02-09T10:17:57Z

Other Versions || Current version unavailable

(c) Social Card

Fig. 150. (Continued) Screenshots of parts of stories constructed using different surrogates from a sample of 16 mementos selected from the 80, 484 mementos found in collection *Egypt Revolution and Politics* (ID 2358).



Graphic: A closer look at the protests in Egypt | Posted | National Post

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

Graphic: A closer look at the protests in Egypt Comments Twitter LinkedIn Digg Buzz Email

news.nationalpost.com @ 2011-02-11T07:36:05Z

Other Versions | Current version

Al-Masry Al-Youm: Today's News from Egypt

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

Quiz How well do you know Naguib Mahfouz?

www.almasryalyoum.com @ 2012-01-16T15:49:41Z

Other Versions | Current version

Egypt News - Home

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

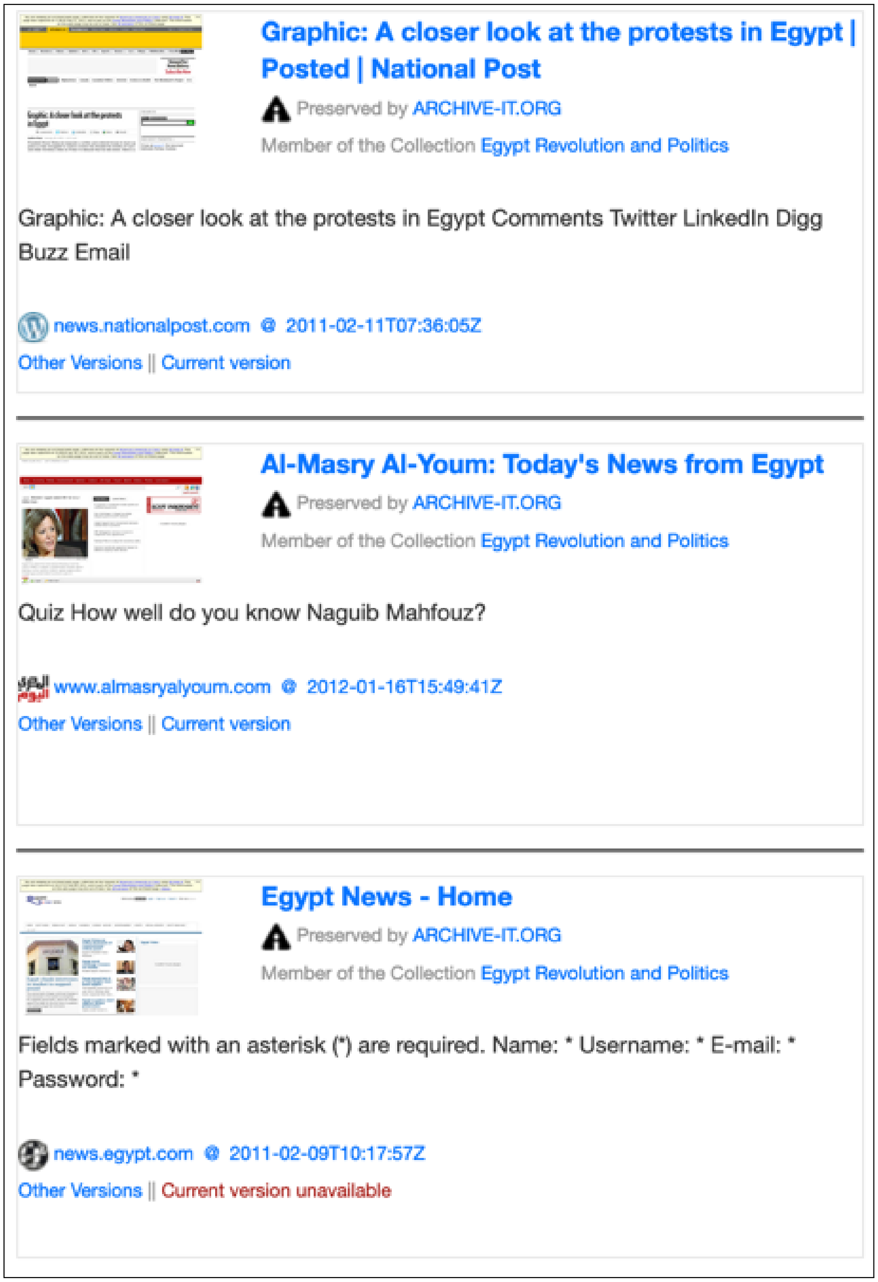
Fields marked with an asterisk (*) are required. Name: * Username: * E-mail: * Password: *

news.egypt.com @ 2011-02-09T10:17:57Z

Other Versions | Current version unavailable

(d) sc+t

Fig. 150. (Continued) Screenshots of parts of stories constructed using different surrogates from a sample of 16 mementos selected from the 80,484 mementos found in collection *Egypt Revolution and Politics* (ID 2358).



Graphic: A closer look at the protests in Egypt | Posted | National Post

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

Graphic: A closer look at the protests in Egypt Comments Twitter LinkedIn Digg Buzz Email

news.nationalpost.com @ 2011-02-11T07:36:05Z

Other Versions || Current version

Al-Masry Al-Youm: Today's News from Egypt

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

Quiz How well do you know Naguib Mahfouz?

www.almasryalyoum.com @ 2012-01-16T15:49:41Z

Other Versions || Current version

Egypt News - Home

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

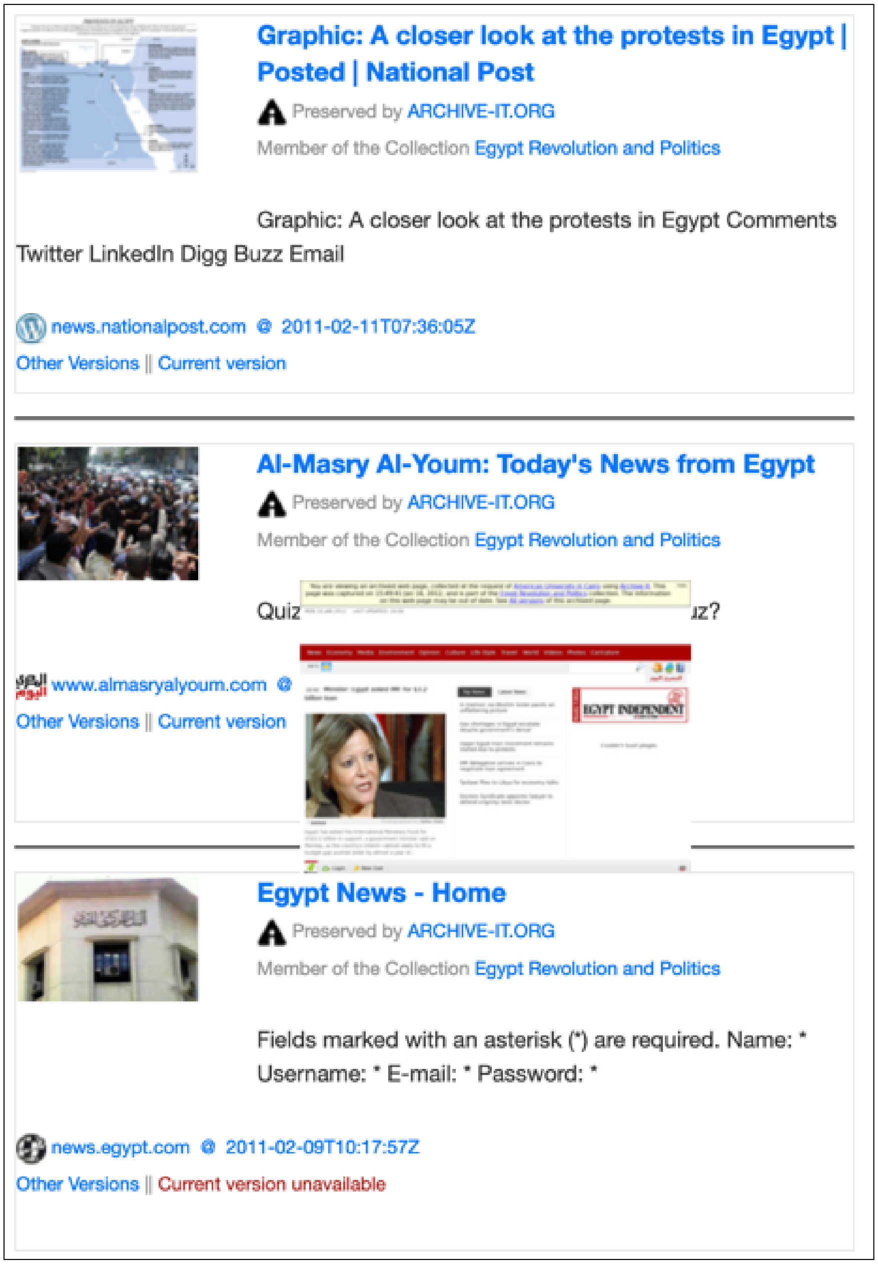
Fields marked with an asterisk (*) are required. Name: * Username: * E-mail: * Password: *

news.egypt.com @ 2011-02-09T10:17:57Z

Other Versions || Current version unavailable

(e) sc/t

Fig. 150. (Continued) Screenshots of parts of stories constructed using different surrogates from a sample of 16 mementos selected from the 80,484 mementos found in collection *Egypt Revolution and Politics* (ID 2358).



Graphic: A closer look at the protests in Egypt | Posted | National Post

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

Graphic: A closer look at the protests in Egypt Comments Twitter LinkedIn Digg Buzz Email

news.nationalpost.com @ 2011-02-11T07:36:05Z

Other Versions || Current version

Al-Masry Al-Youm: Today's News from Egypt

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

Quiz uz?

www.almasryalyoum.com @

Other Versions || Current version

Egypt News - Home

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

Fields marked with an asterisk (*) are required. Name: * Username: * E-mail: * Password: *

news.egypt.com @ 2011-02-09T10:17:57Z

Other Versions || Current version unavailable

(f) sc^t

Fig. 150. (Continued) Screenshots of parts of stories constructed using different surrogates from a sample of 16 mementos selected from the 80, 484 mementos found in collection *Egypt Revolution and Politics* (ID 2358).

Movement 2011/2012 (ID 2950) was selected because it produces the most high-quality thumbnails. *April 16 Archive* (ID 694) has the highest diversity of original resource domain names in its URIs [160]. *Egypt Revolution and Politics* (ID 2358) is a collection that the archivist is still currently maintaining and hence is the longest-lived collection in the set. Collection *Russia Plane Crash Sept 7,2011* (ID 2823) is about an event that is likely not familiar to American MT participants.

To compare against the as-is interface at Archive-It, we generated a facsimile of the Archive-It surrogates using Archive-It's stylesheets as well as metadata gathered using AIU [153]. Figure 150a shows an example story using the Archive-It Facsimile surrogate. We generated a visualization of each story represented as thumbnails (Figure 150b) and again as social cards (Figure 150c). From these, we developed three additional surrogate types combining social cards and thumbnails in order to see if a combination of the two improved results. The surrogates for the story in Figure 150d, noted in this paper as sc+t, display the thumbnail to the right of the existing social card. To produce the surrogates for the story in Figure 150e, noted as sc/t, we replace the social card's striking image with the thumbnail. To conserve space and utilize interactivity, we use JavaScript in the surrogate shown in Figure 150f, noted as sc^t, to allow the user to display the thumbnail if they hover their mouse over the striking image. All visualizations represented the same URI-Ms in the same order.

The first page presented in the survey told the participant that they would view a story for 30 seconds, and then the task would ask them a question. We informed each participant that others may not be viewing the same visualization and that their visualization might respond to mouse hovers, clicks, and other interactions. We did not provide any specific instruction on how to interact with the visualization beyond this. Once the participant had clicked through the instructions, the survey presented them the story. We recorded the initial timestamp of the story page load. The survey system used this timestamp to ensure that the participant had 30 seconds to view the story. We employed JavaScript to ensure that the participant did not view the story for more than 30 seconds.

Once the 30 seconds had expired, then the survey system presented the participant with a new question. This question consisted of checkboxes next to six new surrogates of the same type as the story they had just viewed. We allowed participants as much time as possible to answer the question. The survey system instructed the participant to select the two mementos drawn from the same collection that they had just viewed. We randomly generated the order of these surrogates, but we kept the same order for each collection. Our primary goal was to record how long the users took to answer each question, expecting them to find the two correct answers in all cases.

TABLE 27

A description of the dataset of collections and stories used in this study.

Collection ID	Collection Name	Collected By	Diversity of Original Resource Domain Names	Collection Lifespan	Collection Size (# of Mementos from Seeds)	Story Size	% of Story with Good Surrogates
694	April 16 Archive	VT: Crisis, Tragedy, etc.	0.8391	48 weeks	374	17	88.24%
1784	Earthquake in Haiti	IA Global Events	0.7656	9 weeks	1,080	27	85.19%
2017	Wikileaks 2010	IA Global Events	0.575	3 years	3,333	24	70.83%
2358	Egypt Revolution and Politics	American University in Cairo	0.2585	7 years	80,484	16	87.50%
2535	Brazilian School Shooting	VT: Crisis, Tragedy, etc.	0.2604	5 days	1,540	26	73.08%
2823	Russia Plane Crash Sept 7,2011	VT: Crisis, Tragedy, etc.	0.7843	1 week	603	27	77.78%
2950	Occupy Movement 2011/2012	IA Global Events	0.5585	44 weeks	31,863	15	100.00%
3649	2013 Boston Marathon Bombing	IA Global Events	0.3766	1.9 years	2,421	27	96.30%
3936	US Government Shutdowns	IA Global Events	0.1177	4 years	24,583	16	93.75%
4887	Global Health Events web archive	NLM	0.2723	4 years	9,204	35	100.00%

In addition to instructing users to only select two responses, we also included JavaScript that prevented the user from selecting more or fewer than two. Our question follows Kittur's MT advice to use explicit, verifiable questions as part of the task [180]. The simplicity of our question also avoids user fatigue [177].

To produce the two correct answers, we randomly selected two URI-Ms from the same collection as the story shown to the participants. In choosing these URI-Ms, we discarded ones that used the same original resource domain as any memento in the story, avoiding issues where simple banners or logos might indicate that they are from the same collection.

To produce the four incorrect answers, we selected four other URI-Ms from semantically different collections. To determine which collections were semantically different from our story collection, we extracted entities from each collection in AlNoamany's dataset using Stanford NLP [210]. We then computed the Jaccard distance between these entity sets and selected two collections with the greatest distance from our story collection. We randomly selected two URI-Ms from the most distant and second most distant collections. Table 28 shows the Jaccard scores and which collections were most distant and second-most distant.

In all cases, we discarded URI-Ms that produced poor quality thumbnails to ensure that the quality of the memento did not affect the participant's choice. We also discarded URI-Ms that were off-topic, such as maintenance pages or 404 pages, as described in Section 4.2. If this process discarded a URI-M, we selected again to ensure two selections from the collection they had just viewed, two selections from the most semantically distant collection, and two selections from the second most distant collection. We then randomly sorted the six URI-Ms and generated the surrogates.

Our survey system recorded a timestamp for the load of the question page. It then recorded the timestamp for the load of the completion code. The time the participant took to answer the question is the difference between these two timestamps. We employed JavaScript to record all link clicks and hovers over images and links. This process provides us with several data points for those surrogates: the correctness of their answers, the time the user took to answer the question, and how they interacted with the story. We ran Student's t-test between all pairs of surrogates for completion times and the number of correct answers.

Results

Table 29 displays the mean and median question completion times for each surrogate. At 149.53 seconds, the Archive-It Facsimile surrogates have the highest mean time for answering the question. Browser thumbnails come in second highest at 111.22 seconds. Social cards have the lowest overall mean at 46.12 seconds. The sc+t and sc^t have means slightly greater than 62 seconds. The sc/t surrogate comes in slightly higher at 62.86 seconds. We executed the Student's t-test on the times for all pairs of surrogates. No values are statistically significant at $p < 0.05$. Social cards compared to browser thumbnails produces the lowest p-value at $p = 0.190$. The next lowest p-value is $p = 0.202$ for social cards compared to the Archive-It Facsimile. Despite the mean values, these p-values indicate that our results provide weak evidence that the Archive-It Facsimile takes the most time to evaluate or that social cards take less time to evaluate. The medians demonstrate that some outliers are skewing these means. The Archive-It Facsimile has the lowest median at 33.46 seconds, followed by social cards at 35.89 seconds. The median completion time for browser thumbnails is highest at 53.30 seconds. The combinations of social card and thumbnail all have medians between 38 and 40 seconds. Thus, even though the browser thumbnails still have the highest median, the p-values still demonstrate that we have not established that thumbnails take longer to process.

Table 30 displays the mean and median number of correct answers for each surrogate. With only two correct answers out of six, the distribution of potential values is small. Social cards score highest with a mean correct answer score of 1.75, followed by a tie between sc+t and sc^t at 1.70. The Archive-It Facsimile mean is the lowest at 1.30. The medians are 2.0 for all but the Archive-It Facsimile at 1.5. The Archive-It Facsimile paired with the social card comes closest to statistical significance at $p < 0.05$ with $p = 0.0569$. The next lowest p-values are for Archive-It vs. sc+t at $p = 0.0770$ and Archive-It vs. sc^t at $p = 0.108$. Within collection 2358, social cards, sc+t, and sc/t all fare better than the Archive-It Facsimile at $p = 0.0650$ in all cases. Within collection 2950, social cards and sc+t all fare better than the Archive-It Facsimile, both at $p = 0.0560$. Familiarity with some collections' topics may have influenced the results, which is why we had selected different collections for this evaluation. The close p-values indicate that our general results of social cards compared to the Archive-It surrogate are similar to those of Capra et al. [50], even though Capra focuses on information retrieval and not summarization.

The variation in the quality of Archive-It surrogates may also have shaped the results. Some of the Archive-It surrogates in the story for the *Egypt* collection contained as many as 12 additional metadata fields. In contrast, others from the same collection were minimal Archive-It surrogates. Almost all of the surrogates in the story for the *Occupy* collection contained only the additional metadata field `Group`. In those cases `Group` contained values like `Social Media` and `News Sites and Articles`, text that provides little information specific to the collection. In contrast, almost all Archive-It surrogates for stories from the *Russia* and *VATech* collections contained

TABLE 28

The Jaccard distance of named entities between the different collections in the dataset. Blue indicates the collection that is most distant from the corresponding collection on the left. Green indicates the collection second most distant.

Archive-It Collection	**694**	**1784**	**2017**	**2358**	**2535**	**2823**	**2950**	**3649**	**3936**	**4887**
694 – April 16 Archive	0.000	0.969	0.970	0.981	0.961	0.968	0.986	0.962	0.978	0.974
1784 – Earthquake in Haiti	0.969	0.000	0.959	0.971	0.960	0.975	0.983	0.967	0.972	0.961
2017 – Wikileaks 2010 Document Release Collection	0.970	0.959	0.000	0.962	0.953	0.977	0.965	0.959	0.956	0.966
2358 – Egypt Revolution and Politics	0.981	0.971	0.962	0.000	0.958	0.985	0.965	0.971	0.955	0.970
2535 – Brazilian School Shooting	0.961	0.960	0.953	0.958	0.000	0.974	0.967	0.955	0.952	0.961
2823 – Russia Plane Crash Sept 7,2011	0.968	0.975	0.977	0.985	0.974	0.000	0.992	0.978	0.987	0.977
2950 – Occupy Movement 2011/2012	0.986	0.983	0.965	0.965	0.967	0.992	0.000	0.974	0.942	0.981
3649 – 2013 Boston Marathon Bombing	0.962	0.967	0.959	0.971	0.955	0.978	0.974	0.000	0.961	0.968
3936 – United States Government Shutdowns	0.978	0.972	0.956	0.955	0.952	0.987	0.942	0.961	0.000	0.966
4887 – Global Health Events web archive	0.974	0.961	0.966	0.970	0.961	0.977	0.981	0.968	0.966	0.000

TABLE 29

Mean and median completion times in seconds for each surrogate type per collection and overall.

	Mean					**Median**				
Surrogate Type	**694**	**2358**	**2823**	**2950**	**Overall**	**694**	**2358**	**2823**	**2950**	**Overall**
	(VATech)	(Egypt)	(Russia)	(Occupy)		(VATech)	(Egypt)	(Russia)	(Occupy)	
Archive-It Facsimile	100.27	412.93	36.77	48.14	149.53	42.07	34.71	25.64	32.20	33.46
Browser Thumbnails	68.66	272.54	59.52	44.15	111.22	53.76	103.11	36.67	43.42	53.30
Social Cards	52.91	56.11	34.10	41.37	46.12	40.28	37.68	32.87	44.24	35.89
sc+t	62.19	106.80	26.42	56.03	62.86	43.90	52.18	28.82	39.04	40.07
sc/t	48.01	130.89	47.81	42.69	67.35	38.34	32.93	38.87	38.41	38.38
sc^t	51.67	55.50	27.35	116.29	62.70	53.14	46.35	21.43	36.63	38.07

TABLE 30

Mean and median completion correct answers for each surrogate type per collection and overall.

	Mean					**Median**				
Surrogate Type	**694**	**2358**	**2823**	**2950**	**Overall**	**694**	**2358**	**2823**	**2950**	**Overall**
	(VATech)	(Egypt)	(Russia)	(Occupy)		(VATech)	(Egypt)	(Russia)	(Occupy)	
Archive-It Facsimile	1.6	1.0	1.6	1.0	1.30	2.0	1.0	2.0	1.0	1.5
Browser Thumbnails	1.2	1.4	1.6	1.6	1.45	1.0	2.0	2.0	2.0	2.0
Social Cards	1.6	1.8	1.6	2.0	1.75	2.0	2.0	2.0	2.0	2.0
sc+t	1.2	1.8	1.8	2.0	1.70	1.0	2.0	2.0	2.0	2.0
sc/t	1.4	1.8	1.8	1.2	1.55	2.0	2.0	2.0	2.0	2.0
sc^t	2.0	1.2	2.0	1.6	1.70	2.0	2.0	2.0	2.0	2.0

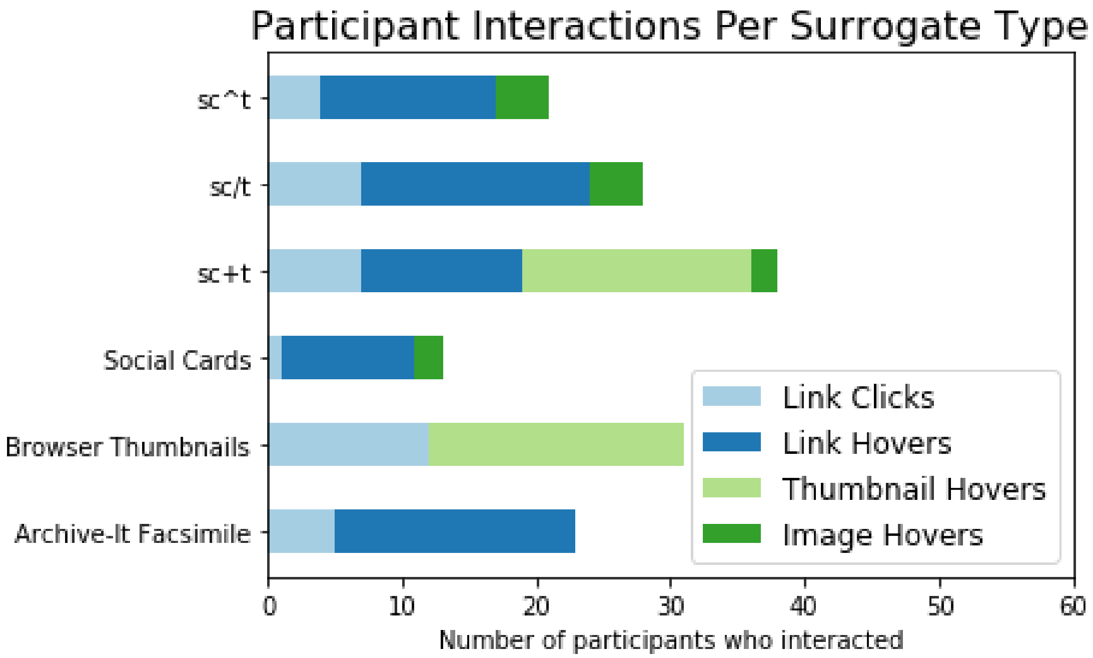


Fig. 151. The number of users interacting per surrogate type, broken down by image hovers, link hovers, and link clicks.

the additional `title` metadata field. For a story consisting of mostly minimal Archive-It surrogates, it is possible that a small number of metadata-rich surrogates provided enough information for the user to answer the question correctly.

Because each story has a different size, we cannot normalize the recorded user interactions across all stories. We chose to tally the number of users who hovered over images, hovered over links, and clicked links. With browser thumbnails, the image is the link's anchor; hence every hover over an image is also a hover over a link. To account for this, we created a separate category named "thumbnail hovers" combining link and image hovers for thumbnails. The results are shown in Figure 151. This engagement gives some insight into the amount of work each participant put into interacting with the story they viewed.

Social cards inspired the least user interactions and the least link clicks. Perhaps the social card inspired more confidence, and fewer participants needed to view the pages behind them. In contrast, most users clicked on thumbnails to open links. Perhaps they found the thumbnails harder to read and felt less confident about their content. Most of the participants interacted with the sc+t surrogate in some way. More link clicks occurred in all cases where thumbnails were present. This difference in behavior, coupled with the different response times and accuracy for sc/t compared to social cards suggests that including the thumbnail rather than a striking image drawn from the page may inspire more activity on the part of the user. Our survey may have measured users zooming in on thumbnails to see them better. Link hovers have a strong correlation with completion time at Pearson's $r = 0.562$, but other interactions, including link clicks, had much weaker correlations to completion time at $|r| > 0.20$. Link hovers have a weak negative correlation with answer accuracy at $r = -0.298$, but other interactions had much weaker correlations to accuracy at $|r| > 0.20$. Participants may have hovered over links to read the URIs in their browser status bar before making their choice.

5.2.3 SUMMARY

Surrogates are visualizations of the values of metadata concepts, like title, description, and striking image. Knowing the surrogate that works best for understanding web archive collections helps us understand which metadata we need to generate for storytelling. We examined the variation in metadata present in Archive-It surrogates. Despite more than half of Archive-It surrogates missing data, we found that a user could still potentially glean information from the URI present in a minimal surrogate. We asked participants from MT to view a story visualized using a given surrogate. We then gave them a question with six mementos visualized using the same surrogate and asked them to choose the two from the six that they believed belonged to the same collection as the story that they just viewed. The type of surrogate does not influence the time to complete the task. However, social cards and social cards side-by-side with thumbnails *probably* provide better collection understanding than the existing Archive-It interface at $p = 0.0569$, and $p = 0.0770$, respectively. These results are consistent with a study by Capra et al. [50] comparing the performance of social cards to text snippets in search results.

We also found that user interactions differ between surrogate types, with social cards having the fewest participants interact and a combination of social card side-by-side with thumbnail encouraging the most participants to interact. Because participants also hover and click more when thumbnails are present, we postulate that users engage more with browser thumbnails than other surrogate elements, possibly to zoom in and see details.

For collection summarization, the overall goal of surrogates is to convey aboutness without requiring the user to click on the underlying link. In this case, social cards require less interaction, provide higher accuracy, and allow the users to answer our questions in less time. These results are encouraging for users of social cards. Social cards require fewer resources to generate and store than thumbnails. Archive-It surrogates require humans to construct metadata, but social cards can be generated dynamically from existing web page content. Users also interact with social cards less, possibly indicating that they find them easier to use. We had initially hypothesized that surrogates with more information drawn from the source document would produce better results, but combining social cards with thumbnails produced more interaction than social cards alone. These features indicate that social cards may be the best surrogate for use in summarizing web archive collections. Thus the metadata of striking image, title, description, and source attribution better helps users understand web archive collections when compared to browser thumbnails or Archive-It surrogates.

Platforms require that web pages contain metadata before the platform can create a social card. Now that we have established social cards as the best surrogate for understanding, how does a system create them when no metadata exists to help it?

5.3 AUTOMATING THE CREATION OF SOCIAL CARD METADATA WHEN METADATA IS MISSING

In the last section we established social cards as our target surrogate because they performed best. As mentioned in Chapter 2, Facebook and Twitter have established standards that allow authors to supply values for the parts of social cards as part of HTML `META` elements published with the document. Because we are creating cards for past web resources, how prevalent is this metadata over time? What can we do if it is not present?

To quantify this problem, we analyzed a dataset of news articles — resources that have undergone editorial review and, presumably, received some care in their publication. News articles are also frequent sources for social cards on social media. Here we analyze [158, 163] the prevalence of this social card metadata in relation to other metadata standards by sampling from the NEWSROOM [110] dataset. Because 39.07% of our dataset does not have the metadata needed to create full cards, we explore the problems of automatically generating descriptions and automatically selecting striking images.

5.3.1 THE PREVALENCE OF SOCIAL CARD METADATA

For analysis over time and across news outlets, we need a relatively balanced dataset with respect to domain name and memento-datetime. Unfortunately, we discovered that the 1.3 million article NEWSROOM dataset was unbalanced in these dimensions. For example, NEWSROOM has 186,095 mementos from `nytimes.com` and 1,429 mementos from `economist.com`. In terms of the memento-datetime year, NEWSROOM contained 19 mementos from 1998 and 279,232 from 2016. Most (23%) of the mementos in the dataset come from 2016. The percentage of mementos decreases each year (e.g., 19% for 2015, 13% for 2014, 10% for 2012). Facebook released the Open Graph Protocol (OGP) standard in 2010 [114] and Twitter can also use that standard. Because of the high percentage of mementos captured in years closer to 2016, and because we wanted to contrast metadata usage before and after Facebook first introduced card standards in 2010, we added all 90,570 NEWSROOM articles with memento-datetimes from 2009 and before to our sample. For those after 2010, we created a bucket for each domain and memento-datetime year. We randomly chose URI-Ms until we filled each domain/year bucket to a size of 1,307 — the median size of all domain/year divisions after 2010. Our sample size after this process was 310,163 mementos.

We downloaded this sample in June 2020. We divided the URI-Ms in the sample into seven subsets to lessen the chance of being rate limited by the Internet Archive. We spread them across different servers in Amsterdam, Frankfurt, London, New York, Northern Virginia, San Francisco, and Toronto. We felt confident in this approach because the Internet Archive presents the content it recorded and does not alter it for different geographic locations. To address failed downloads caused by rate-limiting, we repeated the downloads once in July and again in August 2020. We discovered that the downloads with HTTP status codes of 400, 403, 404, and 405 were actual captures of pages with those status codes. Some mementos redirected to mementos captured after 2016, similar to the behavior that Ainsworth et al. [3] reported using a different dataset. We removed all mementos captured after 2016 because they were outside of the bounds of the original dataset. After resolving these issues, our downloaded NEWSROOM sample consisted of 277,724 mementos.

Table 31 lists the number of mementos in the NEWSROOM sample capable of creating different combinations of social card units. Because Facebook is more forgiving with missing fields, 59.56% of articles can create a full card on Facebook, while only 43.86% can do so with Twitter. We assigned each metadata field encountered to a category based on its corresponding standard or usage. We removed all instances where metadata fields were specific to a domain (e.g., only `nytimes.com` used the metadata field `byl`). Figure 152 demonstrates how these categories changed over time. There is a focus on HTML standard metadata over time throughout our sample because all mementos in the NEWSROOM dataset contain a textual summary in some form. Before the OGP or Twitter standards, these articles used the HTML standard `description` field. We observed that news articles rapidly adopted social card metadata fields, starting with 13.13% adoption of OGP in 2010 and reaching 93.05% by 2016. After 2010, there is a rise in all types of metadata usage, focusing on search engines, mobile apps, browser customization, and social media, showing that news articles leveraged promotion of their content once standard metadata fields became available.

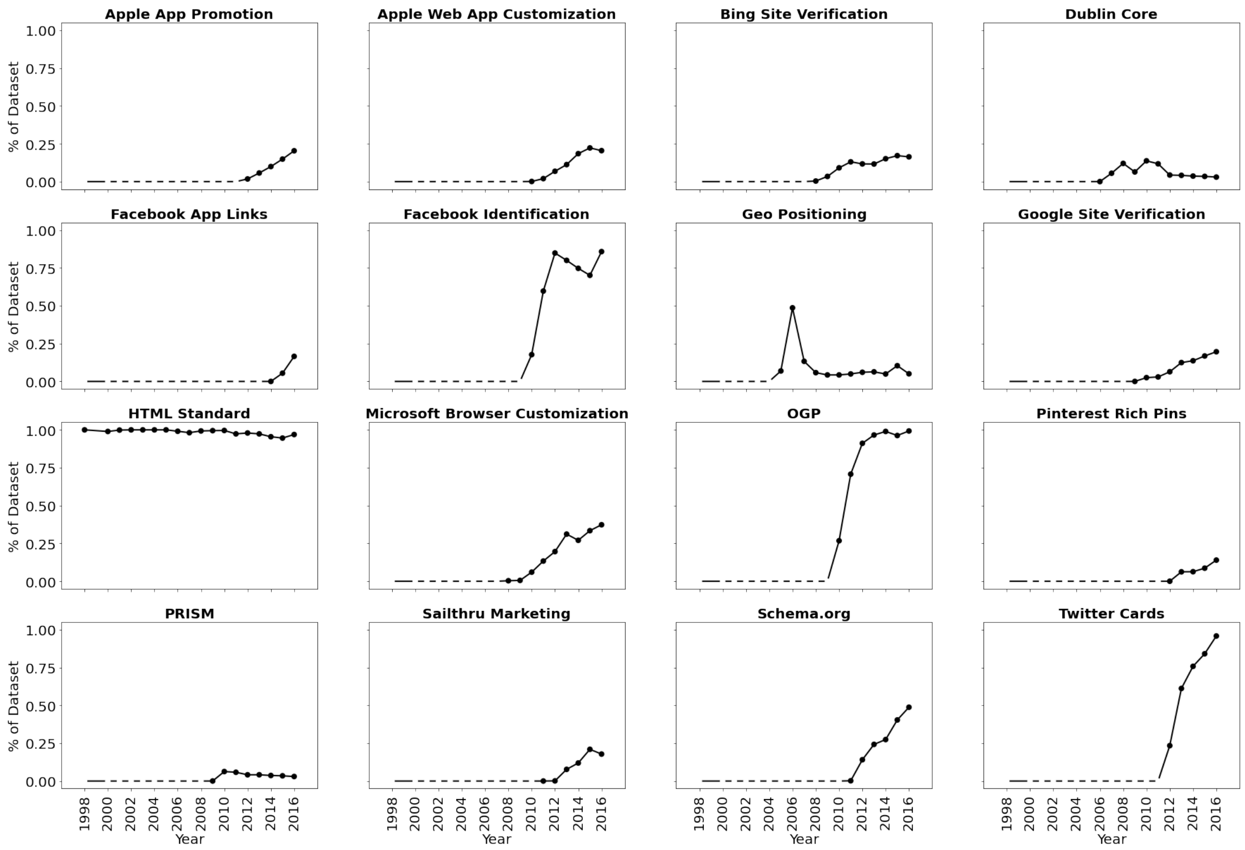


Fig. 152. By memento-datetime, the percentage of NEWSROOM articles each year that have adopted a given metadata category. Dashed lines indicate a y-axis value of 0. We see rapid growth in the adoption of OGP and Twitter Card metadata between 2010 and 2016. Other categories fail to reach this level of adoption; thus, authors and publishers desire social cards.

5.3.2 AUTOMATICALLY GENERATING DESCRIPTIONS

From Table 31 we see that 38.98% of news articles did not contain a social card description field. If no `og:description`, `twitter:description`, or `description` field exists, we can generate a description field using a variety of sentence ranking algorithms, such as TextRank [221], but we need to know an acceptable value for the length of the description field. Also, some sentence ranking algorithm implementations require an input argument of the number of words or sentences to return.

Figure 153 visualizes the changes in description mean character counts for different description fields over time. We note that there are different behaviors for the different description fields available to article authors. Twitter, represented by `twitter:description` demonstrates a

TABLE 31
The number of mementos in the NEWSROOM sample capable of generating different social card units.

	# of Capable Mementos for Platform and % of Dataset	
Social Card Units	**Twitter**	**Facebook**
title only	125,201 (45.08%)	277,724 (100%)*
title only from metadata	125,201 (45.08%)	189,275 (68.15%)
title and description only	123,386 (44.43%)	169,468 (61.02%)†
title, description, image	121,823 (43.86%)	165,426 (59.56%)

* if `og:title` is missing, Facebook will use the HTML's `title` element
† our testing shows that Facebook does not currently support this configuration, only displaying a card with a title

more consistent use of around 140 characters, starting at 145 characters in 2012, reaching a peak of 154 characters in 2014, and settling to 136 characters in 2016. Facebook's (`og:description`) mean description length rises and falls, starting at 151 characters in 2010, reaching a high of 228 characters in 2012, and finishing at 168 characters in 2016. The schema.org description field starts at 73 characters in 2011, peaks at 202 in 2015, and finishes at 144 in 2016. These different field lengths suggest that the choice of output platform affects the author's choice in character length. The HTML standard description field, however, has no such influences, and Figure 153 suggests that an author unencumbered by a target platform might choose a longer description length. In this case, the mean number of characters starts at 30 characters in 1998, reaches the first peak of 215 characters in 2002, the second peak of 237 characters in 2012, and finishes at 268 characters by 2016. This behavior implies that, without any platform influences, not only have descriptions gotten longer, but 268 characters is a good intended description length for description field metadata.

To address the second issues of input length argument for sentence ranking algorithms, we examined the word lengths (Figure 154) and sentence lengths (Figure 155) over time. The pattern is very similar to Figure 153. If we hold to our prior discussion of favoring the HTML standard description field, 2016's 52 word and two-sentence targets are likely good inputs.

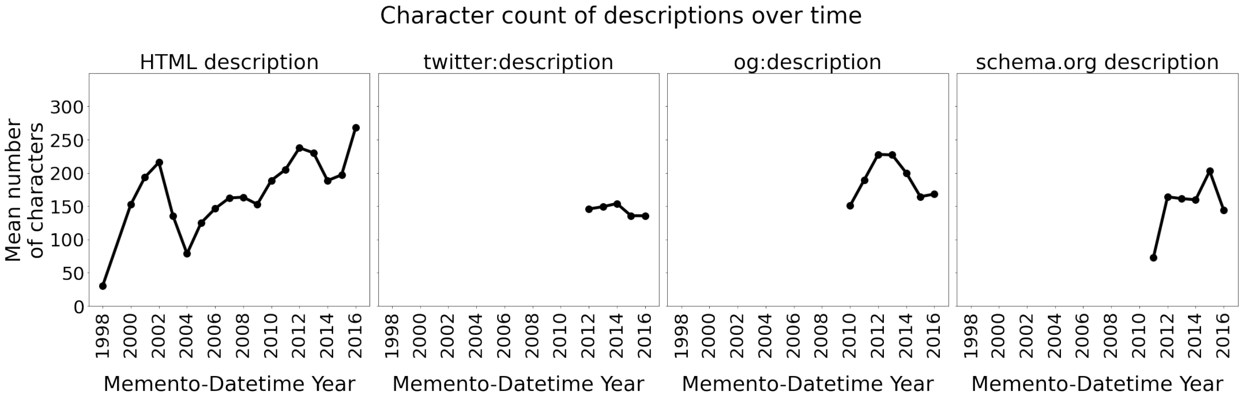


Fig. 153. The character count of descriptions by metadata field over time.

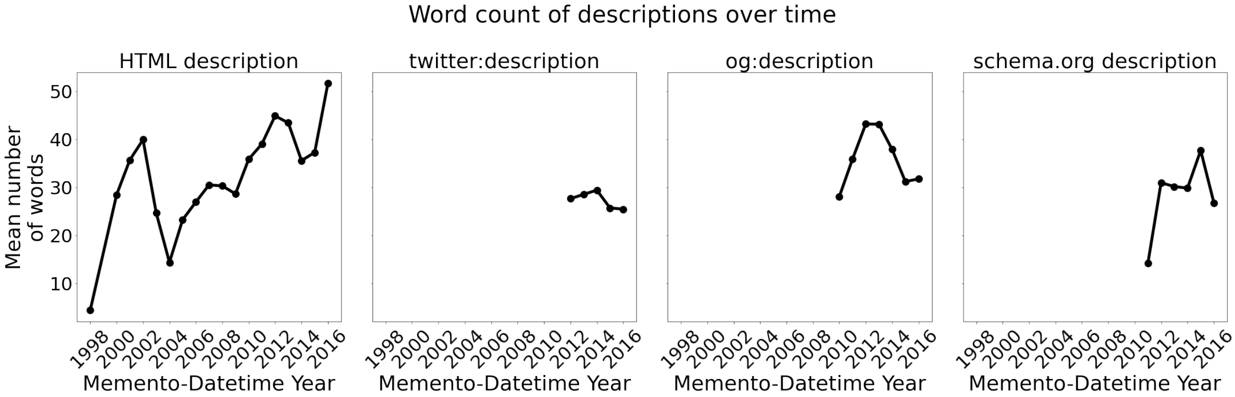


Fig. 154. The word count of descriptions by metadata field over time.

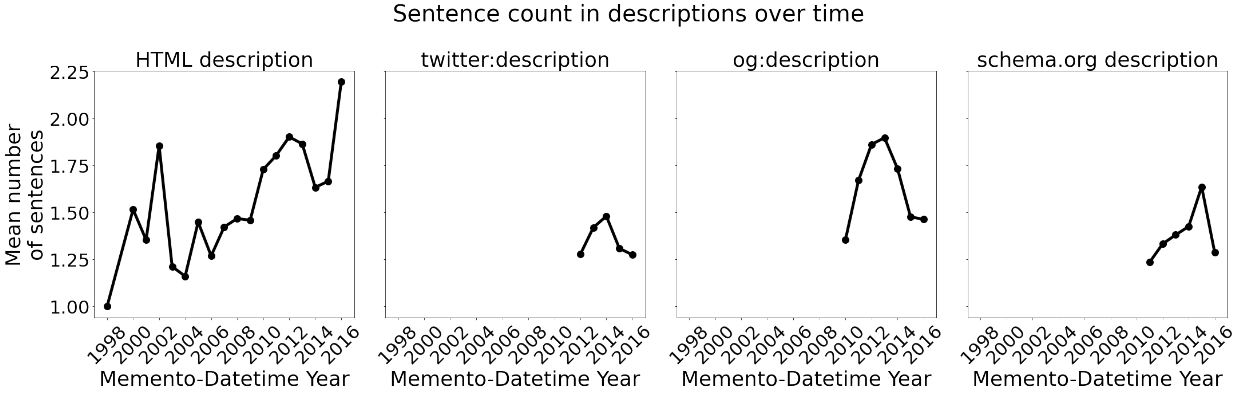


Fig. 155. The sentence count of descriptions by metadata field over time.

5.3.3 AUTOMATICALLY SELECTING STRIKING IMAGES FOR SOCIAL CARDS

With the automatic generation of descriptions addressed, we sought the best method for addressing missing striking images. Our overall goal was to find the approach that, given a set of images, will select the image closest to what a human selected, as found in the metadata, for the same document, regardless of whether that image exists outside of the image metadata. Thus, the striking images found in the metadata are our ground truth because the document's editorial process produced them.

Methods

Thus, any dataset applied to this endeavor requires documents where all images are available, and all documents must contain at least one striking image. This disqualifies documents that we cannot download or cannot parse with BeautifulSoup [277] and documents with images that we cannot process with Pillow [59] or ImageMagick [298]. We also disqualify documents with only one image because we have no prediction to make, and these easy wins may skew results. For each document, we consider image URIs found in the `META` elements from the `HEAD` as well as those provided by the `src` and `srcset` attributes of each `IMG` element found in the `BODY`.

For analyzing news articles, we started with our NEWSROOM sample of 310,163 mementos. The issues we encountered left us with 37,522 articles to evaluate. The fact that we lose much of our dataset to download issues further emphasizes that many mementos can benefit from automatic striking image prediction because some of their images may be missing or corrupted.

Once we had our datasets, we applied different prediction approaches. Each prediction approach required one or more image features (e.g., byte size) to be successful. Some approaches choose the image with the highest or lowest value for the feature (e.g., largest or smallest byte size). We also ran 20 trials where we randomly chose each document's striking image for comparison with all approaches.

Base features. Because social card creation tools must operate in real-time, we considered image features that could be calculated quickly by most image libraries across most image formats. The Pillow library provides the following features for all image formats:

- image byte size
- image width in pixels
- image height in pixels
- the number of columns in the image's histogram with a value of 0 (**negative space**)

- image size in pixels
- the image's aspect ratio
- the number of colors in the image

To achieve better results, we also supplied multiple features as input to various scikit-learn [261] classifiers. When training classifiers, we placed images found in the HTML metadata into the class of *present-in-metadata* and other images into the class of *other*. When testing classifiers, we considered each document to be its own test case. We supplied the features of each image in a document to the classifier and asked it to provide the probability that the image comes from the class *present-in-metadata*. From that set, we choose the image that has the highest probability of being in the class *present-in-metadata* as the predicted striking image for that document. We chose this probability method so that each document contains at least one striking image prediction. If we had merely evaluated how well the classifier predicted an image's class (*present-in-metadata* or *other*), there would be documents for which the classifier found no striking image. With our method, even a low probability of belonging to *present-in-metadata* still predicts a striking image because this process still compares all of the document's images' probabilities.

Social card creation tools have to contend with many different types of images. The values for some image features, like byte size, have no upper bound, making proper scaling challenging to estimate for the long term. Scaling can also remove precision from some measures, leading to poor results. Thus, we only considered the following classifiers because their scikit-learn implementations provide class probability scores and they do not require scaling:

- AdaBoost
- Decision Tree
- Gaussian Naive Bayes
- Linear Discriminant Analysis
- Logistic Regression
- Random Forest

We evaluate classifier operation with 10-fold cross-validation by document. Instead of evaluating how well each prediction approach predicts an image's class (*present-in-metadata* or *other*), we instead evaluate how well, given a set of images found in a document, the approach selects the ground truth striking image for that document. Evaluating this process leads us away from considering metrics like F_1 because they invite discussions about what recall means in this application. Instead, we consider other metrics commonly associated with information retrieval because we can consider each document a query and the output of a prediction approach as a set of ranked results where the striking images are the relevant results. Thus, we apply $Precision@1$ ($P@1$) to determine the given approach's level of success for predicting the striking image from the metadata. If an approach produces a relevant result as its first result for a document, the $P@1$ score for that document and approach is 1, and 0 otherwise. We take the mean of the $P@1$ scores for all documents, making $P@1$ a proxy for accuracy. We also supply Mean Reciprocal Rank (MRR) to evaluate how well an approach performed even if it failed to achieve $P@1 = 1.0$ for a document. For a given document, we determine the rank of the first relevant image ($rank_i$) as provided by the approach under test and compute the rank's reciprocal ($\frac{1}{rank_i}$). We then compute the mean of all of these reciprocal ranks across all documents for the same approach. For example, if ten images exist in a document and the approach ranks a relevant image in fourth place, then the reciprocal rank is $\frac{1}{4} = 0.25$. A score of 1.0 for $P@1$ and MRR is ideal.

We recognize that the same image may exist at different URIs, or the striking image may be a cropped or resized form of the same image elsewhere in the document. We apply a perceptive hash (pHash) distance to our evaluation as a proxy for human judgment of image similarity for these issues. If an approach selects image A, but the ground truth image is B, and if $pHashDistance(A,B) = 0$, then we consider A to be just as relevant when computing $P@1$ and MRR.

Different pHash implementations exist. We evaluated ImageHash's pHash [189], ImageMagick's pHash [292, 317], and Zauner's pHash [337]. We manually reviewed the intra- and inter-document similarity distances provided by each pHash implementation for the images found in ten news and ten scholarly articles. We concluded that ImageMagick's pHash provided the most intuitive similarity distances because it placed photographs at the same distance to each other, separate from logos and text. ImageMagick's pHash was more consistent with considering cropped or resized images to be similar to their original form. ImageMagick's pHash scores are not scaled and reached values as high as 6000 during this evaluation. While low scores intuitively indicated similar images, at higher values, there is a greater discrepancy between the scores and human perception; thus, we needed an upper bound for scaling that kept this discrepancy in mind. Our median score from this 20 article evaluation was 280.904. We scaled all distance scores such that scores above twice the median (561.808) became 1.0, and other scores result from dividing their value by 561.808. ImageMagick has an issue processing certain JPEGs [188], so we converted all

TABLE 32

Correlation values for features in the NEWSROOM dataset to image present in metadata.

feature category	feature	Spearman's ρ
base features	byte size	0.2146
	width	0.2191
	height	0.2537
	negative space	-0.1442
	size in pixels	0.2403
	aspect ratio	0.0225
	number of colors	0.1749

images to PNGs before computing the pHash distance.

Results

Figure 156 demonstrates the performance of different prediction approaches with the NEWSROOM sample. Each line demonstrates the increasing *MRR* or $P@1$ score, identified by its points' y-axis values, as produced by an evaluation that we performed at the corresponding pHash distance on the x-axis. When we evaluate at a pHash distance of 1.0, any image selected by the approach is equivalent to the ground truth; thus, *MRR* and $P@1$ are 1. At a pHash distance of 0, only images that are perceptively equal to the ground truth image are relevant. The best approach achieves the highest *MRR* and $P@1$ at the lowest pHash distance, resulting in lines that start in the upper left quadrant of each graph.

Randomly choosing images gives poor performance at a pHash distance of 0 with $MRR = 0.4016$, $P@1 = 0.1551$. Training Random Forest with all base features results in the best performance at a distance of 0 with $MRR = 0.8825$ and $P@1 = 0.8314$. To see if we could improve performance with fewer features, we applied Spearman's ρ to our features, as shown in Table 32. We see that aspect ratio and negative space have the lowest correlation. We retrained Random Forest with these features removed and achieved $MRR = 0.8782$, $P@1 = 0.8267$ at a pHash distance of 0. Removing additional features produced results with $P@1 < 0.8$.

5.3.4 SUMMARY

A social card summarizes each web page through the units of page title, striking image, domain name, and description. While the title and domain name are often easily extracted, automatically

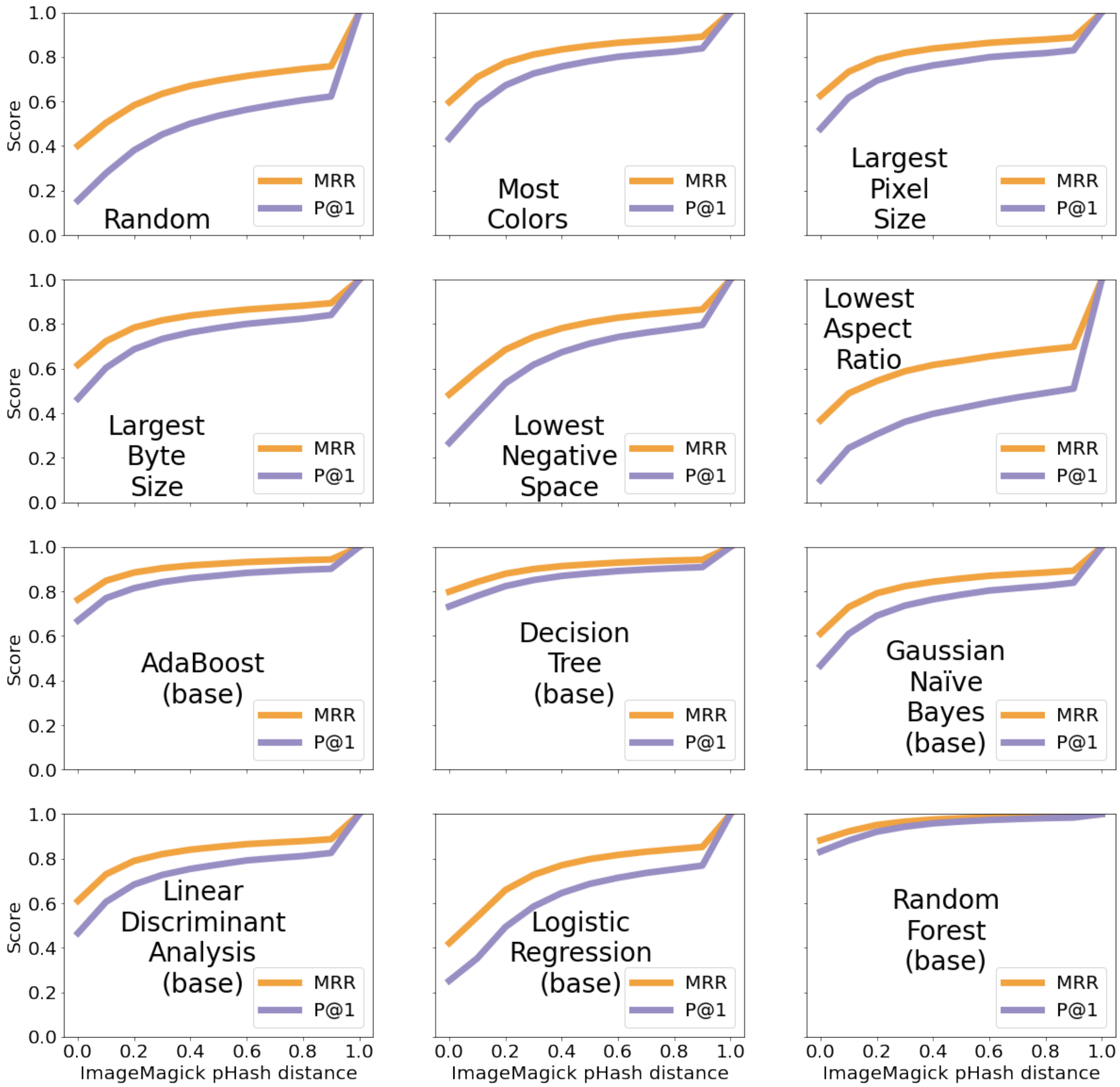


Fig. 156. This visualization demonstrates the *MRR* and *P@*1 results for different striking image prediction approaches as run against our sample. The best approach achieves the highest *MRR* and *P@*1 at the lowest pHash distance, making the ideal situation one where an approach's lines start higher into a graph's upper left corner. Items in parentheses are feature categories applied to the classifier, if applicable.

generating the description and a striking image metadata are more challenging. Social media platforms provide standards so that authors can insert their own values for these social card units into their HTML pages as metadata.

Our evaluation of the archived web pages (mementos) of 296,162 news articles from the NEWSROOM dataset revealed that card metadata fields are nonexistent for mementos captured before 2010. This translates into roughly 150 billion web pages at the Internet Archive for which social card creation tools will need to automatically generate descriptions and striking images. We found that news articles rapidly adopted social card metadata fields, with 13.13% adoption in 2010 and reaching 93.05% by 2016.

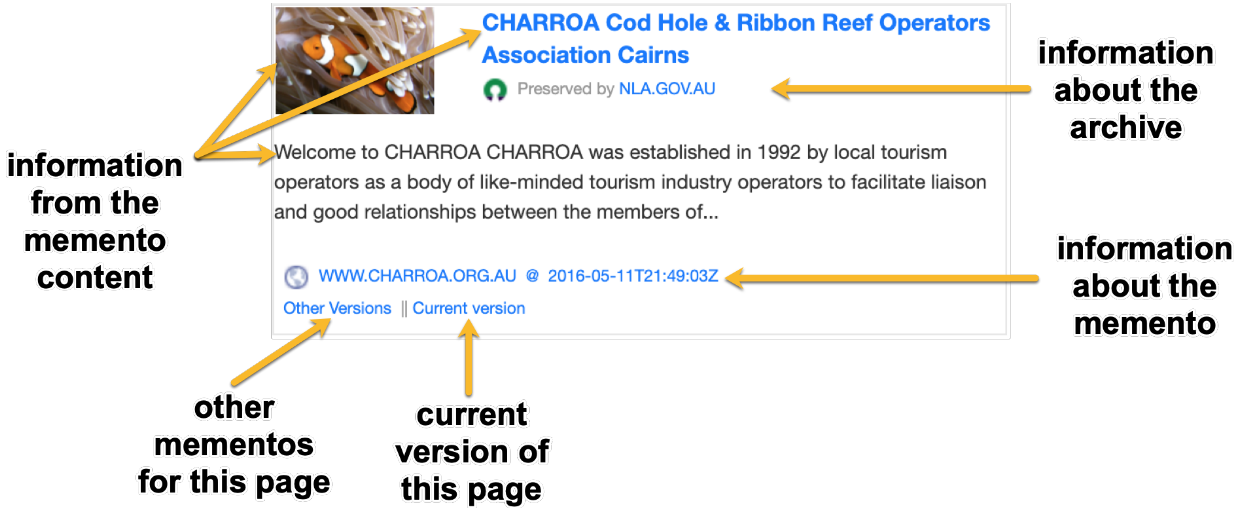
If metadata is missing, we can still generate its values from the content of the document. If the description is missing, we can apply one of the many text summarization algorithms that exist for choosing the best sentence from a document. These algorithms require an input in terms of number of characters, words, or sentences. After analyzing the behavior of authors with respect to description metadata fields, we find that authors generate a mean of 268 characters, 52 words, or 2 sentences for their own descriptions.

Selecting striking images when they are missing offers its own challenges. Like with the text summarization algorithms, our method chooses the best image from a document. Our goal was to select the same image as chosen by the author of the document, so we relied upon existing metadata as ground truth when analyzing different approaches. By analyzing 37,522 articles from the NEWSROOM dataset, we achieve $P@1 = 0.8314$, $MRR = 0.8825$ at a pHash distance of 0 with Random Forest and the base features of image width, height, byte size, pixel size, negative space, aspect ratio, and color count.

5.4 MEMENTOEMBED: ARCHIVE-AWARE SURROGATES AND DOCUMENT METADATA

In Section 5.1 we established that existing platforms do not reliably produce social cards for mementos. This result led us to explore which surrogate, and, by proxy, which metadata best helps users like Natasha, Rustam, and Olayinka understand the collection behind a story. Next, through user testing in Section 5.2, we established that social cards best support the understanding of collections. Finally, in Chapter 5.3, we discussed creating social cards when no metadata exists, which was the case for roughly 40% of the mementos in our dataset. With this motivation and the knowledge gained from these studies, we developed **MementoEmbed**, the first **archive-aware** metadata and surrogate service.

MementoEmbed is archive-aware, meaning it can include memento-specific information such



CHARROA Cod Hole & Ribbon Reef Operators Association Cairns

Preserved by NLA.GOV.AU

Welcome to CHARROA CHARROA was established in 1992 by local tourism operators as a body of like-minded tourism industry operators to facilitate liaison and good relationships between the members of...

WWW.CHARROA.ORG.AU @ 2016-05-11T21:49:03Z

Other Versions || Current version

information from the memento content
information about the archive
information about the memento
other mementos for this page
current version of this page

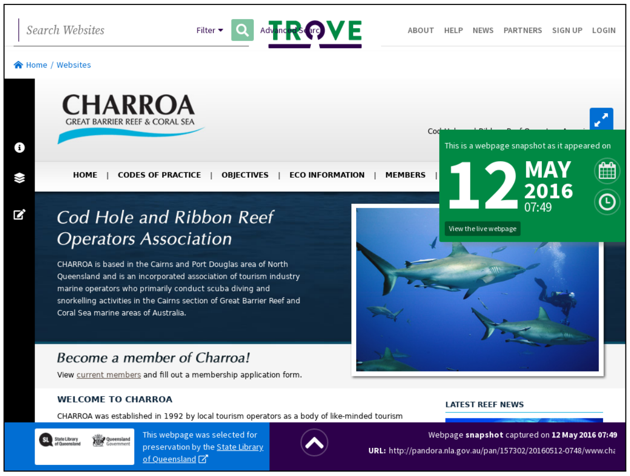
Fig. 157. A social card generated by MementoEmbed annotated to show how it keeps the information about the archive separate from information about the memento, avoiding misattribution.

as the memento-datetime, the archive from which a memento originates, and the memento's original resource domain name. This section provides an overview of MementoEmbed. Please consult the official documentation [147, 151] and work by Jones et al. [142, 156] for more details on MementoEmbed's internals.

As demonstrated in Figure 157, a social card generated by MementoEmbed separates information derived from the memento content from information about the archive containing it. Through the Memento Protocol, MementoEmbed can attribute this content to its original website, avoiding the confusion produced by other services, further allowing a user to understand the nature of the web page behind the surrogate. Table 33 lists the different metadata fields available in a MementoEmbed card and how those fields are not present or not accurate in a typical card service like Facebook.

MementoEmbed currently generates the following surrogate types:

- social cards (Figure 157)
- browser thumbnails (Figure 158)
- imagereels (Figure 159) – animated GIFs of the top five images
- word clouds (Figure 160)



This is a webpage snapshot as it appeared on 12 MAY 2016 07:49

Cod Hole and Ribbon Reef Operators Association

CHARROA is based in the Cairns and Port Douglas area of North Queensland and is an incorporated association of tourism industry marine operators who primarily conduct scuba diving and snorkelling activities in the Cairns section of Great Barrier Reef and Coral Sea marine areas of Australia.

Become a member of Charroa!

View current members and fill out a membership application form.

WELCOME TO CHARROA

CHARROA was established in 1992 by local tourism operators as a body of like-minded tourism

LATEST REEF NEWS

This webpage was selected for preservation by the State Library of Queensland

Webpage **snapshot** captured on **12 May 2016 07:49**

URL: http://pandora.nla.gov.au/pan/157302/20160512-0748/www.cha

Fig. 158. A browser thumbnail, a screenshot of a web page in a browser without the browser interface, generated by MementoEmbed based on the same memento as used for Figure 157.

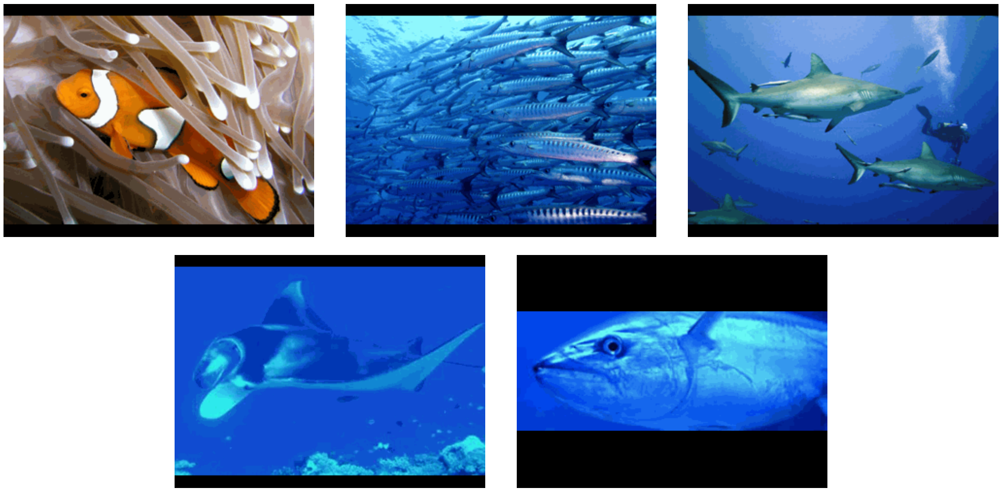


Fig. 159. The five images that make up MementoEmbed's imagereel created from the same memento as used for Figure 157. The imagereel starts black and fades to an image before returning to black and fading to another image. The full animated GIF contains 108 frames.

TABLE 33
Because it is archive-aware, MementoEmbed can provide more accurate metadata for mementos that typical social card services.

Metadata field	**Typical Card Service**	**MementoEmbed**
title	yes	yes
description	yes	yes
striking image	yes	yes
original resource domain	confuses archive and original resource	yes
archive domain	confuses archive and original resource	yes
archive favicon	confuses archive and original resource	yes
original resource favicon	confuses archive and original resource	yes
memento-datetime	no	yes
URI-R	no	yes
TimeMap	no	yes

MementoEmbed is more than a surrogate generation service. MementoEmbed provides an API capable of generating document metadata in addition to these surrogates to fulfill its role in the storytelling process. MementoEmbed's API provides the following endpoints that supply different types of information for a single memento:

- `/services/memento/contentdata/` - title, memento-datetime, and text description, Section 5.3 has more information on automatically generating a description

- `/services/memento/bestimage/` - the URI-M of the striking image, Section 5.3 and Jones et al. [156] has more information on how this process works

- `/services/memento/imagedata/` - an analysis of all images found in the memento, providing many derived values like pixel size, perceptive hash, and format, to be applied in the experiment run in Section 5.3

- `/services/memento/archivedata/` - information about the archive, collection names, domains, and archive favicon

- `/services/memento/originalresourcedata/` - information about the original resource, including domain name, URI-R, favicon, and if the URI-R is still accessible

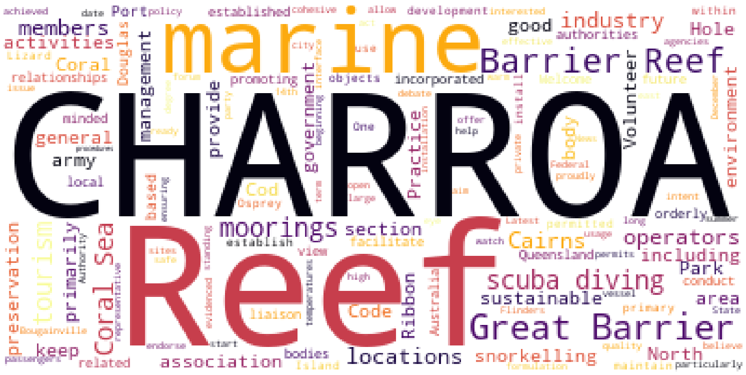


Fig. 160. A word cloud generated by MementoEmbed created from the same memento as used for Figure 157.

- `/services/memento/seeddata/` - information on the seed that was used to generate the memento, including its URI-T, the number of other mementos available for this memento, first and last mementos, and archivist's metadata for this same URI-R
- `/services/memento/paragraphrank/` - a listing of all paragraphs in the memento, scored by the readability algorithm [66]
- `/services/memento/sentencerank/` - a listing of sentences in the memento, scored by either Lede3 [110] or TextRank [221]
- `/services/memento/page-metadata` - all HTML metadata discovered in the memento's `META` elements

MementoEmbed's API also provides different endpoints at `/services/product/` for generating the different surrogates shown in Figures 157, 158, 159, and 160. For more detailed and current information consult the documentation [147].

As shown in Section 5.1, existing social media and card creation services fail to reliably generate surrogates for mementos. MementoEmbed is an archive-aware surrogate service that fulfills the storytelling process of generating document metadata for mementos. It is able to separate data about the archive from data about the memento, improving the accuracy of metadata for mementos. MementoEmbed can produce four surrogate types, including social cards and thumbnails, and provides an API that supplies metadata about mementos in a machine readable form.

5.5 CHAPTER SUMMARY

In this chapter, we focused on generating document metadata, primarily through analyzing surrogates. Surrogates give us a view of the metadata produced by different platforms for a given web resource. We started by evaluating 55 surrogate-producing platforms and found that all failed to generate the necessary metadata for mementos. Popular platforms like Facebook, Twitter, and Tumblr are poor targets for storytelling because they do not reliably create surrogates for mementos. Tumblr performed best but still conflated archive information with memento content, creating a potentially confusing experience for readers.

RQ3: What surrogates work best for understanding groups of mementos?

Surrogates are built on metadata. If we understand which surrogate works best to understand a collection, we know what document metadata fields we need to generate for storytelling. We conducted a user study with 120 Mechanical Turk participants. We showed each participant a story rendered using a specific surrogate type and generated from an Archive-It collection where the collection's archivist had chosen its exemplars. After 30 seconds, the participant was presented with six new surrogates of the same type and asked which two of the six came from the same collection. Participants who were shown social cards answered questions more accurately than those shown the Archive-It surrogate. We also instrumented their stories with JavaScript to measure their interactions and found that social cards required far fewer interactions. Thus, social cards probably provide for better understanding of web archive collections.

RQ4: What methods that automate the creation of surrogates produce results that best match humans' behavior?

Now that we have shown that social cards perform best, we sought to understand how they were constructed. As mentioned in Chapter 2 social cards representing a web page are built from metadata inserted by the author of that web page. We analyzed the metadata of 296,162 mementos of news articles and discovered that 40.44% could not produce a social card. Most of these corresponded to mementos captured before 2010, the year Facebook introduced the first social card standard. The Internet Archive captured roughly 150 billion pages before this year. To determine how to generate document metadata for these pages, we analyzed the behaviors of web page authors. First, to generate a description for the page, we can apply text selection algorithms to select the best sentences that describe the memento. We found that authors generate a mean of 268 characters, 52 words, or two sentences for their descriptions, making these good values as arguments for automatic summarization algorithms. Second, to address documents without striking image metadata, we can select striking images from among the images found in the document. With Random Forest and the features of image width, height, byte size, pixel size, negative space, aspect ratio, and color count, we achieve $P@1 = 0.8314$ when attempting to select the same image as the author.

The lack of support for mementos by existing platforms and the research into which surrogate worked best led us to develop MementoEmbed, an archive-aware metadata generation and surrogate service. MementoEmbed applies the Memento Protocol to ensure that archive metadata and memento content metadata stay separate. MementoEmbed can produce many different surrogate types, including social cards, browser thumbnails, word clouds, and imagereels. In addition, MementoEmbed supports an API that provides sentence ranking, image analysis, titles, and other metadata needed to support visualizing stories for mementos as part of its metadata generation duties.

In Chapter 4 we reviewed work on selecting exemplars and generating story metadata for stories. In this chapter, we tackled the problem of generating document metadata. Now we are ready to visualize our stories and distribute them for Natasha, Rustam, Olayinka, Elbert, and Ling.

CHAPTER 6

VISUALIZING AND DISTRIBUTING STORIES

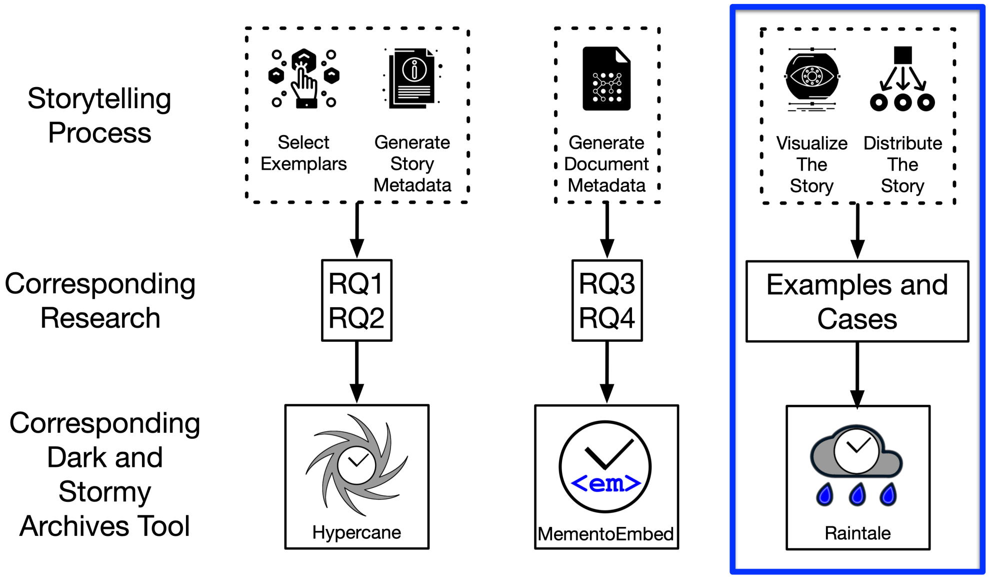


Fig. 161. The processes for storytelling that are covered in this chapter are indicated by the blue box.

Chapter 4 focused on selecting exemplars and generating story metadata. We introduced Hypercane as a tool that handles these storytelling processes. From these exemplars, we can generate document metadata, as covered in Chapter 5. MementoEmbed is the tool that best generates document metadata for mementos. In this chapter, we cover visualizing and distributing our stories. In this chapter 6.1 we introduce Raintale, a MementoEmbed client that accepts the story metadata and exemplars from Hypercane and generates stories by visualizing and distributing them (Figure 161). Raintale supports templates so that users can customize the presentation of their stories. Raintale can also produce output as files and social media posts, applying some of the ideas we discussed with Twitter and Facebook in Section 5.1. By combining Raintale with Hypercane and MementoEmbed, we address all of the storytelling processes in our model, allowing us to help our personas Natasha, Rustam, Olayinka, Elbert, and Ling (Table 1 on page 14). In Section 6.2 we cover several use cases with web archives, demonstrating all of the storytelling processes in action.

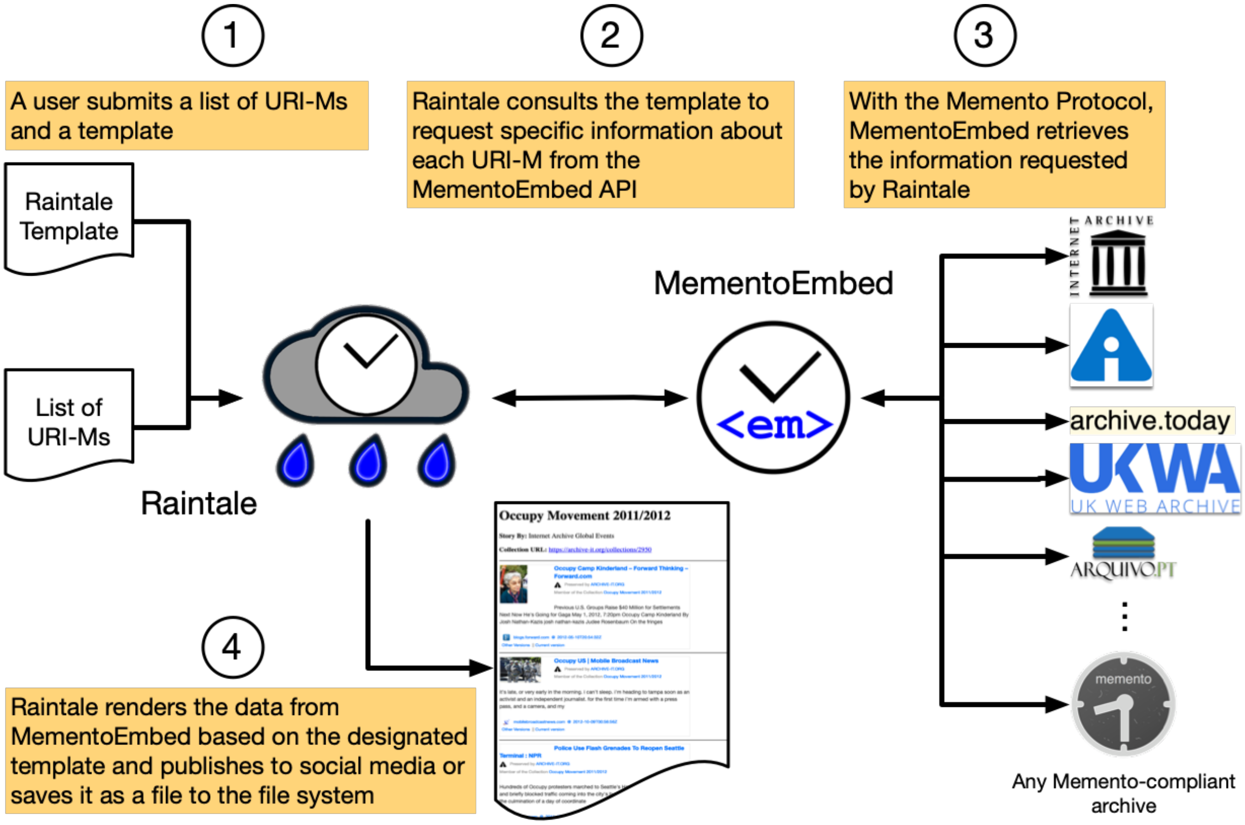

1
A user submits a list of URI-Ms and a template
2
Raintale consults the template to request specific information about each URI-M from the MementoEmbed API
3
With the Memento Protocol, MementoEmbed retrieves the information requested by Raintale
Raintale Template
List of URI-Ms
Raintale
MementoEmbed
<em>
INTERNET ARCHIVE
archive.today
UKWA
UK WEB ARCHIVE
ARQUIVO.PT
memento
Any Memento-compliant archive
Occupy Movement 2011/2012
4
Raintale renders the data from MementoEmbed based on the designated template and publishes to social media or saves it as a file to the file system

Fig. 162. The MementoEmbed-Raintale Architecture for Visualizing a Story.

6.1 RAINTALE: A STORYTELLING TOOL FOR WEB ARCHIVES

Raintale [143, 156] leverages MementoEmbed's API (Section 5.4) to generate complete stories containing the surrogates of many different mementos. More details on Raintale are covered by Jones et al. [143, 156], but we will summarize the tool here. It requires two forms of input: a story file containing a list of URI-Ms and a template. The template formats the output and contains variables that indicate what information the user wishes to display for each memento. Figure 162 demonstrates the relationship between MementoEmbed and Raintale. In step 1, the user provides a template and a list of URI-Ms to Raintale. In step 2, Raintale records all template variables. For each URI-M provided, Raintale consults MementoEmbed's API for each variable's value in the corresponding memento. In step 3, MementoEmbed downloads the memento from its web archive and performs natural language processing, image analysis, or extracts information via the Memento Protocol [312], as appropriate to the API request. In step 4, Raintale consolidates the data from these API responses and renders the template with the gathered data. It produces a story

```
# tellstory -i story-mementos.txt --storyteller template \
      --story-template mytemplate.tmpl -o mystory.html \
      --title "This is My Story Title"
```

Fig. 163. Example usage of Raintale's `tellstory` command.

constructed from both document and story metadata.

Raintale is a command-line application that provides various options for the end-user. To create a story using the mementos supplied in `story-mementos.txt` and a template file `mytemplate.tmpl` and save the output to `mystory.html`, a user would type the command shown in Figure 163.

We designed Raintale this way so that users can easily integrate it into shell scripts with other utilities, such as Hypercane (Section 4.5). Their scripts may contain utilities that provide Raintale with a list of mementos or an automated utility that might automatically generate templates for different use cases. We intend for Raintale to be part of an automated web archiving workflow, and we provide features that allow the archivist to customize its output to their needs.

Raintale supports different types of storytellers. File storytellers, such as `template` or `html` save content to a file. Service storytellers, such as `twitter` or `facebook`, write content to a specific social media service. A user can apply templates to different types of storytellers. Instead of supplying their own template, a Raintale user can apply one of its pre-packaged **presets** including surrogates rendered in formats like HTML, Markdown, and MediaWiki. Figure 164 shows an example of a story visualized with Raintale's default HTML template.

Raintale's story file input can take two forms, depending on the needs of the user. The simplest story file format is a newline-separate text file containing URI-Ms. A Raintale user applying this format *must* supply a title to their story via the `--title` command-line argument. Where the story file contains the URI-Ms and the values of some of the template variables, the Raintale template controls the formatting and which information Raintale will request and display for each URI-M. Figure 165 demonstrates a simplified example of how this works. Raintale first parses the template file and determines which variables exist. At this point in the template, Raintale encounters HTML formatting and the Raintale variable `element.surrogate.title`. Raintale uses its internal mapping to discover that the appropriate MementoEmbed API endpoint for `element.surrogate.title` is `/services/memento/contentdata`. Then Raintale iterates through all elements in the story file. When it reaches `https://archive.`

example.net/mymemento2, Raintale appends the current URI-M (https://archive.example.net/mymemento2) to /services/memento/contentdata and submits an HTTP request to MementoEmbed. It then extracts the title from MementoEmbed's JSON output and applies it to the resulting output file for the story.

Figure 166 demonstrates what we can do with a template designed for a collection from the National Library of Australia (NLA). The collection belongs to the Trove platform and covers Tourism. Rather than applying the default HTML template, we used a custom template incorporating the NLA's colors. Our template uses an interactive carousel that allows visitors to scroll left or right to see the various mementos from the collection. This interactive form of storytelling still has surrogates and incorporates story metadata gathered by Hypercane and document metadata generated by MementoEmbed.

Raintale is the endpoint of the DSA tool architecture, as shown in Figure 167. Hypercane (Section 4.5) focuses on intelligent sampling to select exemplars. Raintale generates the visualization. These two tools consult the others to produce document and story metadata and to render the final visualization.

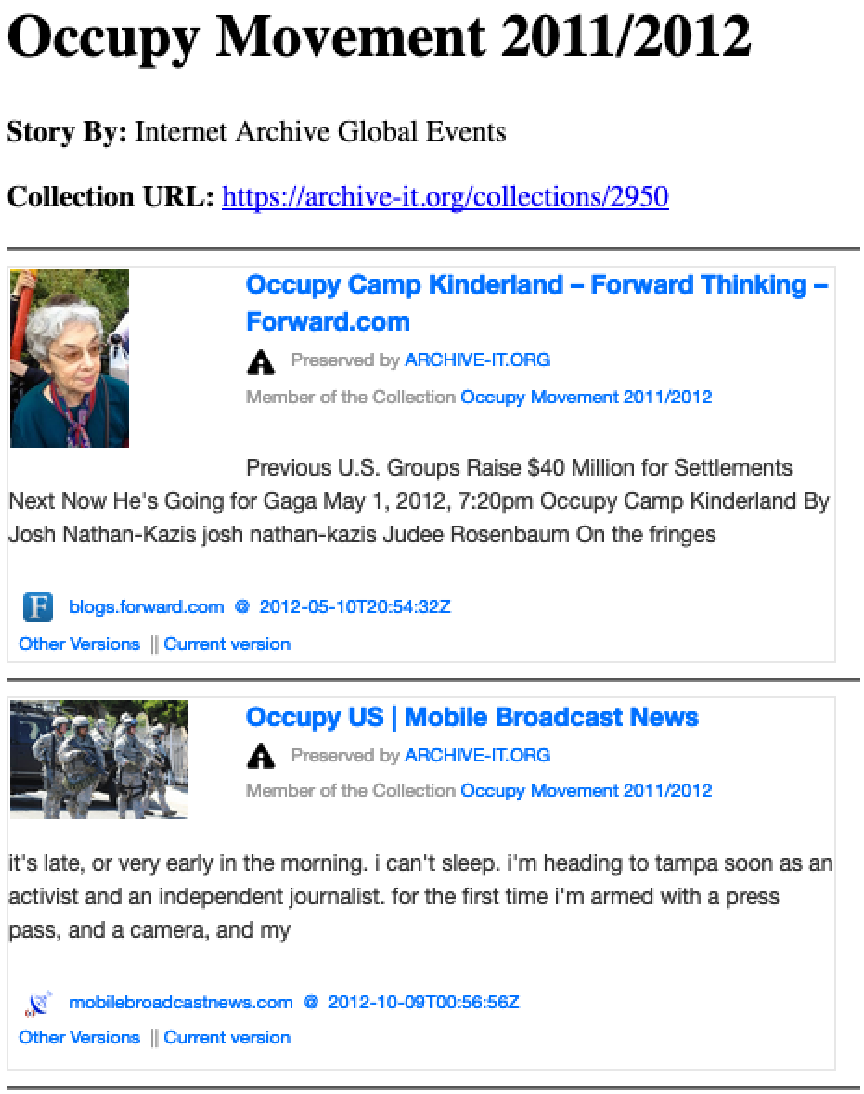
6.2 STORYTELLING WITH WEB ARCHIVE COLLECTIONS

We established the DSA Puddles web site[1] for sharing web visualizations from the Dark and Stormy Archives Project. DSA Puddles is written using Jekyll [212], a static site generator supported by GitHub Pages. Figure 168 shows the storytelling processes introduced in Chapter 1 and how using each DSA tool with the DSA Puddles website satisfies these processes. In this case, Raintale visualizes our story, and GitHub Pages handles distribution.

Figure 169 displays the Archive-It interface for collection 13529 *Novel Coronavirus (COVID-19)*. It consists of 13,200 seeds and 23,376 mementos. In 2020, we generated a story from this collection using the DSA tools. We built it via the following process, visualized in Figure 170:

1. Ask Hypercane select exemplars from this collection (13529) with the DSA1 algorithm
2. Ask Hypercane to produce the collection metadata (e.g., collection name, collecting organization) for this collection
3. Ask Hypercane to produce a report of named entities from these exemplars
4. Ask Hypercane to produce a report of terms from these exemplars

[1] https://oduwsdl.github.io/dsa-puddles/



Occupy Movement 2011/2012

Story By: Internet Archive Global Events

Collection URL: https://archive-it.org/collections/2950

Occupy Camp Kinderland – Forward Thinking – Forward.com

Preserved by ARCHIVE-IT.ORG

Member of the Collection Occupy Movement 2011/2012

Previous U.S. Groups Raise $40 Million for Settlements Next Now He's Going for Gaga May 1, 2012, 7:20pm Occupy Camp Kinderland By Josh Nathan-Kazis josh nathan-kazis Judee Rosenbaum On the fringes

blogs.forward.com @ 2012-05-10T20:54:32Z

Other Versions || Current version

Occupy US | Mobile Broadcast News

Preserved by ARCHIVE-IT.ORG

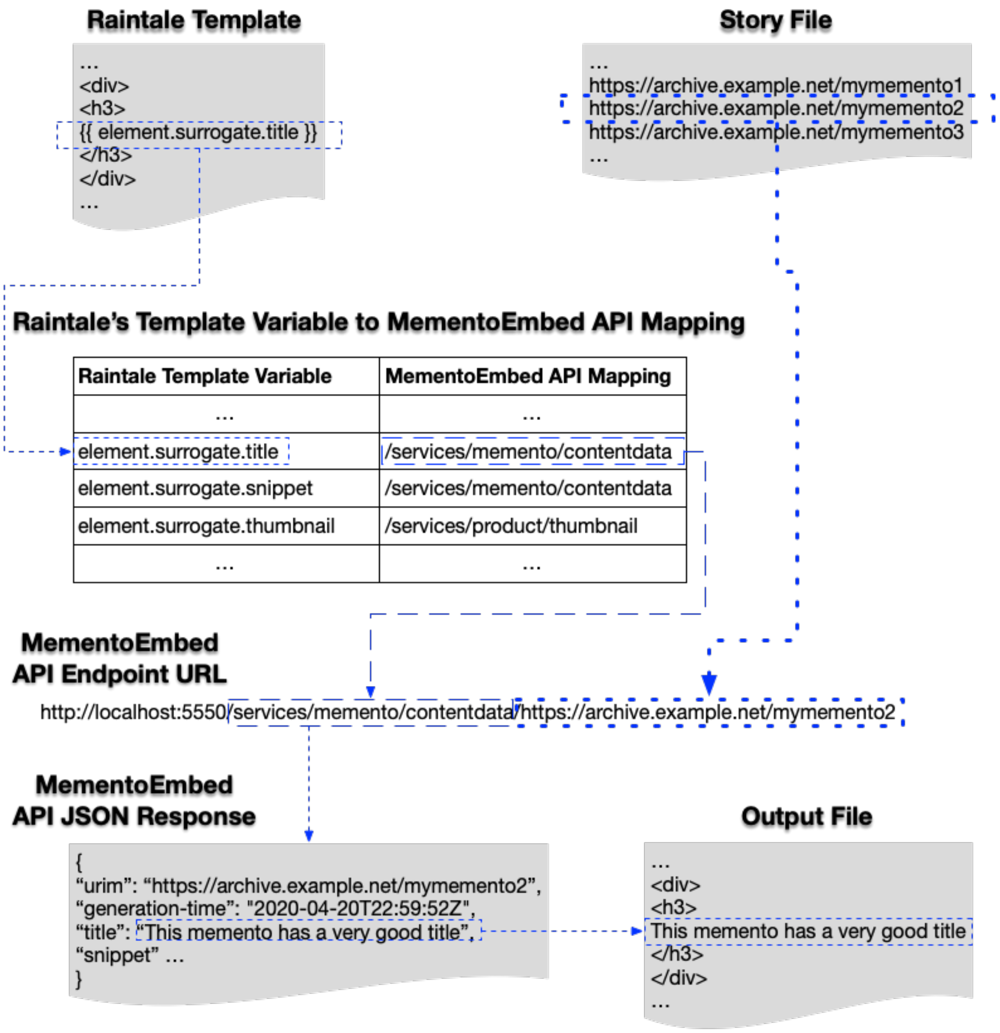
Member of the Collection Occupy Movement 2011/2012

it's late, or very early in the morning. i can't sleep. i'm heading to tampa soon as an activist and an independent journalist. for the first time i'm armed with a press pass, and a camera, and my

mobilebroadcastnews.com @ 2012-10-09T00:56:56Z

Other Versions || Current version

Fig. 164. An example using Raintale's default HTML preset, with MementoEmbed cards visualizing the exemplars and story metadata (title, link to source collection, collection creator) shown at the top.



```
...
<div>
<h3>
{{ element.surrogate.title }}
</h3>
</div>
...
```

```
...
https://archive.example.net/mymemento1
https://archive.example.net/mymemento2
https://archive.example.net/mymemento3
...
```

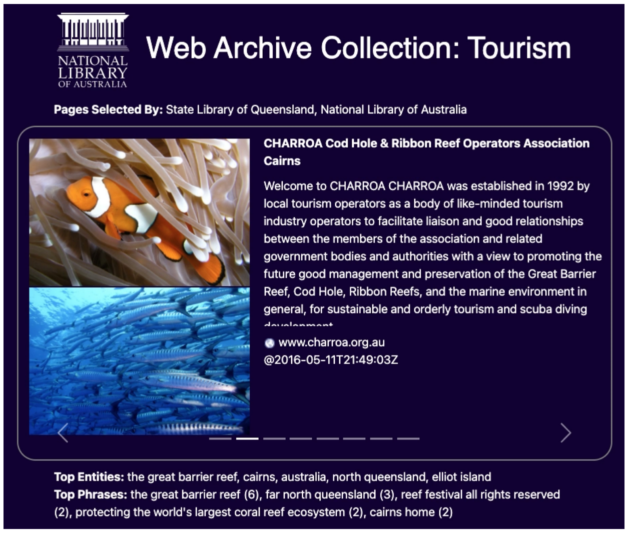
Raintale Template Variable	MementoEmbed API Mapping
...	...
element.surrogate.title	/services/memento/contentdata
element.surrogate.snippet	/services/memento/contentdata
element.surrogate.thumbnail	/services/product/thumbnail
...	...

```
http://localhost:5550/services/memento/contentdata/https://archive.example.net/mymemento2
```

```
{
“urim”: “https://archive.example.net/mymemento2”,
“generation-time”: "2020-04-20T22:59:52Z",
“title”: “This memento has a very good title”,
“snippet” ...
}
```

```
...
<div>
<h3>
This memento has a very good title
</h3>
</div>
...
```

Fig. 165. How Raintale combines the template and the story file to insert memento data into the desired output.



Web Archive Collection: Tourism

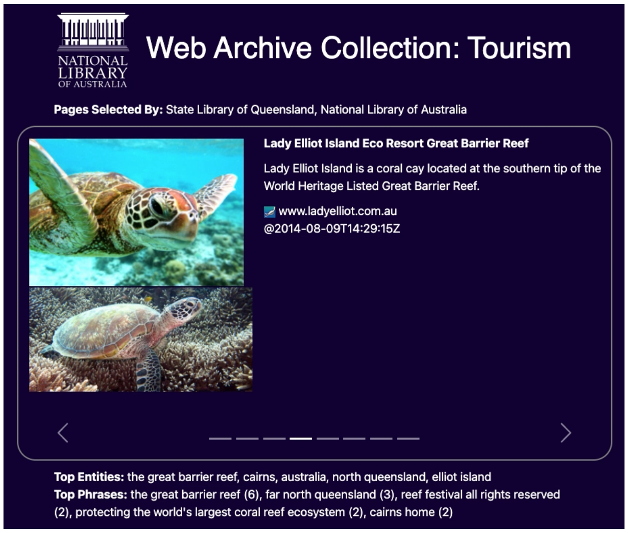
Pages Selected By: State Library of Queensland, National Library of Australia

CHARROA Cod Hole & Ribbon Reef Operators Association Cairns

Welcome to CHARROA CHARROA was established in 1992 by local tourism operators as a body of like-minded tourism industry operators to facilitate liaison and good relationships between the members of the association and related government bodies and authorities with a view to promoting the future good management and preservation of the Great Barrier Reef, Cod Hole, Ribbon Reefs, and the marine environment in general, for sustainable and orderly tourism and scuba diving development

www.charroa.org.au
@2016-05-11T21:49:03Z

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)



Web Archive Collection: Tourism

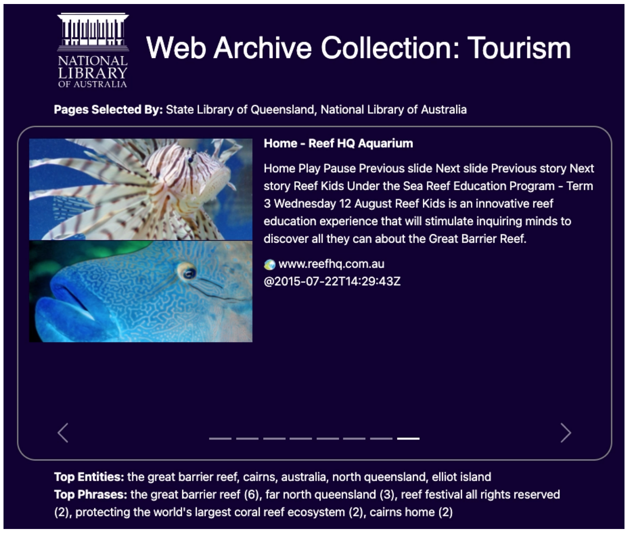
Pages Selected By: State Library of Queensland, National Library of Australia

Lady Elliot Island Eco Resort Great Barrier Reef

Lady Elliot Island is a coral cay located at the southern tip of the World Heritage Listed Great Barrier Reef.

www.ladyelliot.com.au
@2014-08-09T14:29:15Z

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)



Web Archive Collection: Tourism

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

Home - Reef HQ Aquarium

Home Play Pause Previous slide Next slide Previous story Next story Reef Kids Under the Sea Reef Education Program - Term 3 Wednesday 12 August Reef Kids is an innovative reef education experience that will stimulate inquiring minds to discover all they can about the Great Barrier Reef.

www.reefhq.com.au
@2015-07-22T14:29:43Z

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)

Fig. 166. Three frames of an HTML carousel story created using a custom template. This represents exemplars from the *Tourism* collection created by the National Library of Australia.

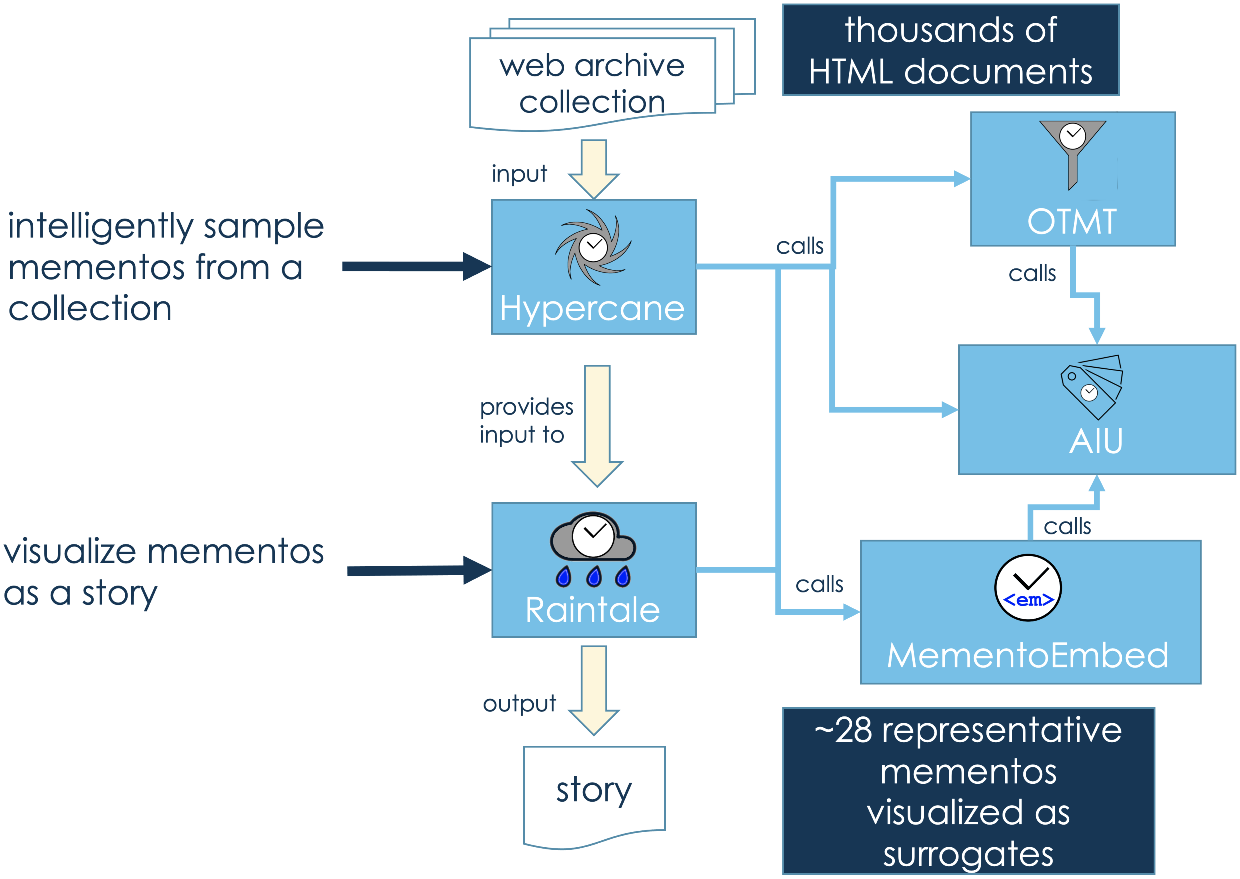

web archive collection
thousands of HTML documents
input
intelligently sample mementos from a collection
Hypercane
calls
OTMT
calls
AIU
provides input to
visualize mementos as a story
Raintale
calls
calls
<em>
MementoEmbed
output
story
~28 representative mementos visualized as surrogates

Fig. 167. The entire DSA architecture for storytelling.

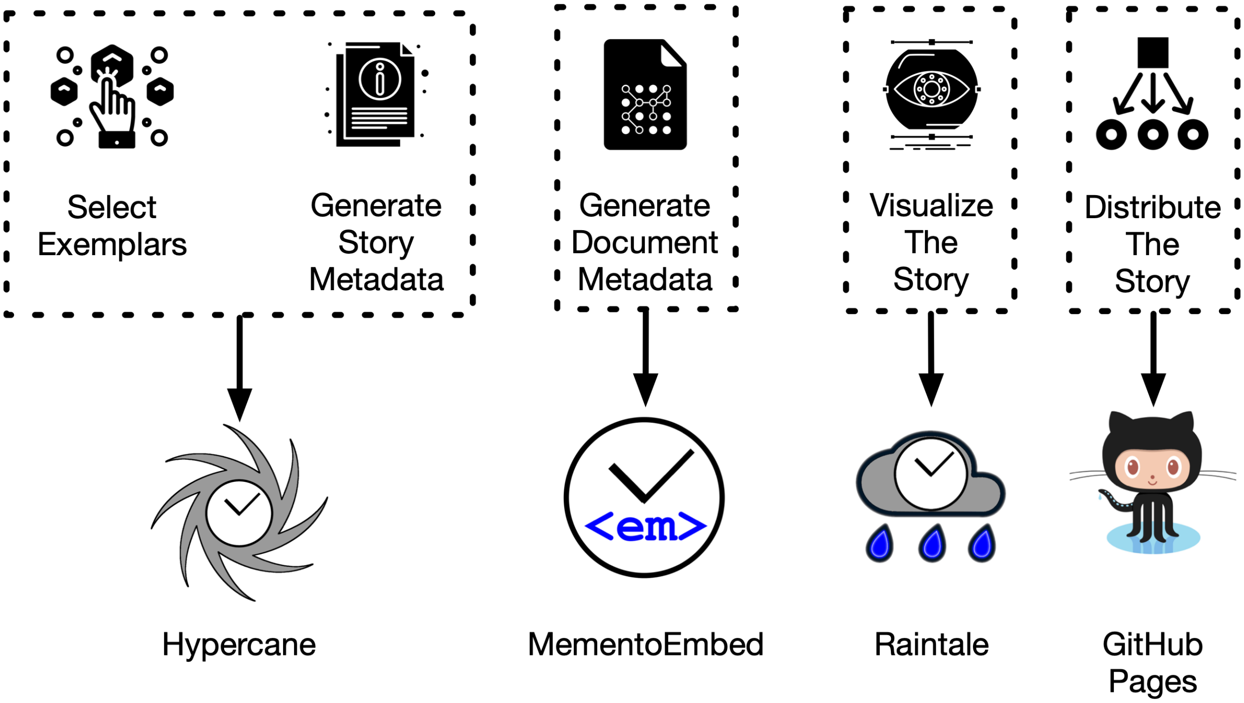
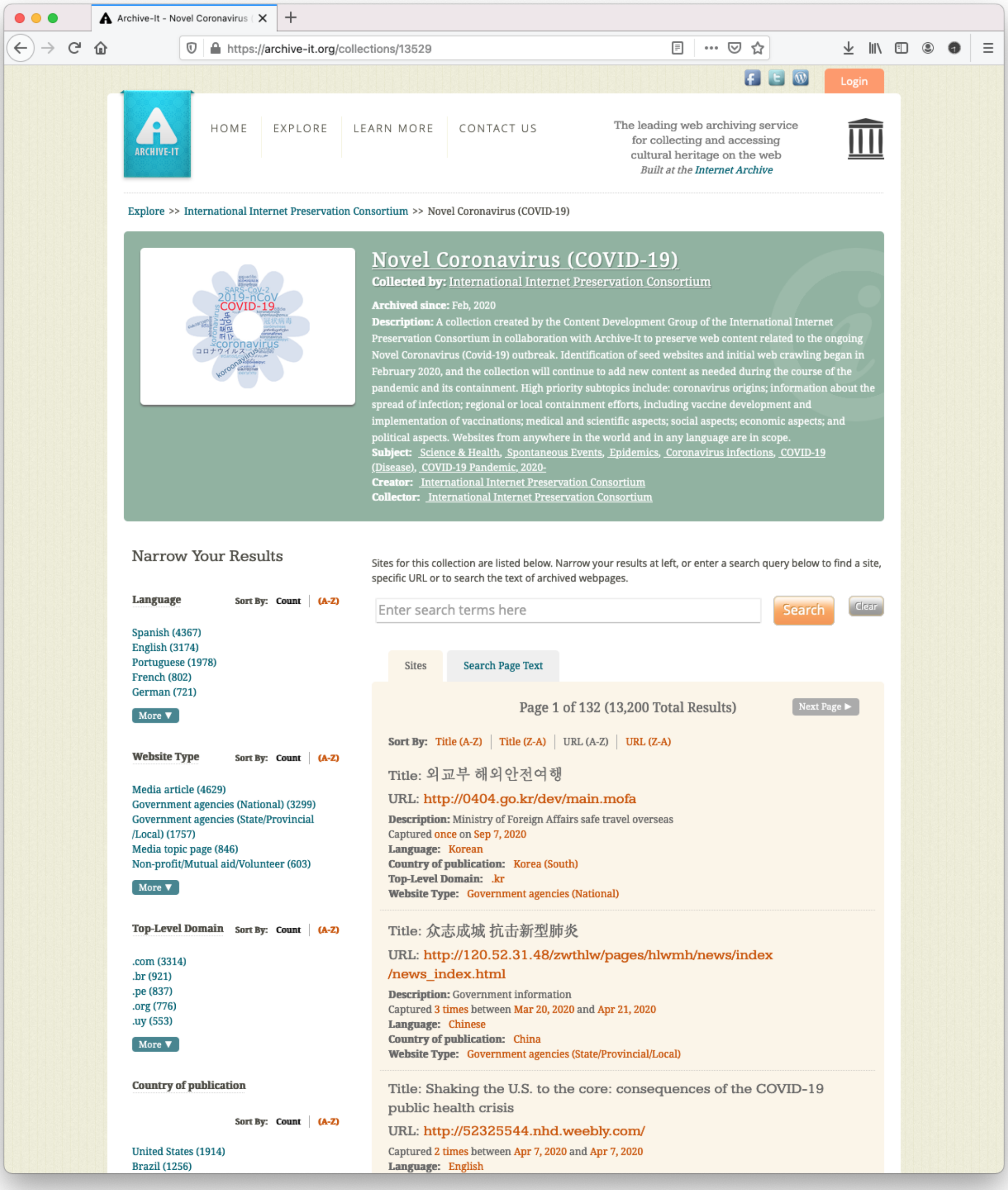


Fig. 168. Storytelling processes incorporating the DSA Puddles web site.

5. Ask Hypercane to produce a report about the images found in these exemplars, and select the best striking image to represent this story
6. Ask Hypercane to synthesize the exemplars and these reports into a JSON Raintale story file
7. Ask Raintale to generate the story using a custom template that provides variables for all story metadata and formats exemplars
   (a) Raintale then asks MementoEmbed to generate the document metadata for each exemplar based on the variables from the template
   (b) Raintale visualizes the story by applying the template and this metadata
8. We commit the story to GitHub, and GitHub Pages distributes the story for us

In Figure 169 we see seeds, metadata, facets, and other tools to help us search the collection. In Figure 171 we see people in masks, pictures of the virus, maps of the pandemic spreading across the world, terms, named entities, sources, dates, and links back to the Archive-It collection from which it came. This story is not only a multimedia experience; it has a higher information



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)

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)

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

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

Country of publication

Sort By: Count | (A-Z)

United States (1914)
Brazil (1256)

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

Page 1 of 132 (13,200 Total Results)

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
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

Fig. 169. The International Internet Preservation Consortium’s Archive-It collection *Novel Coronavirus (COVID-19)* seen in its Archive-It interface.

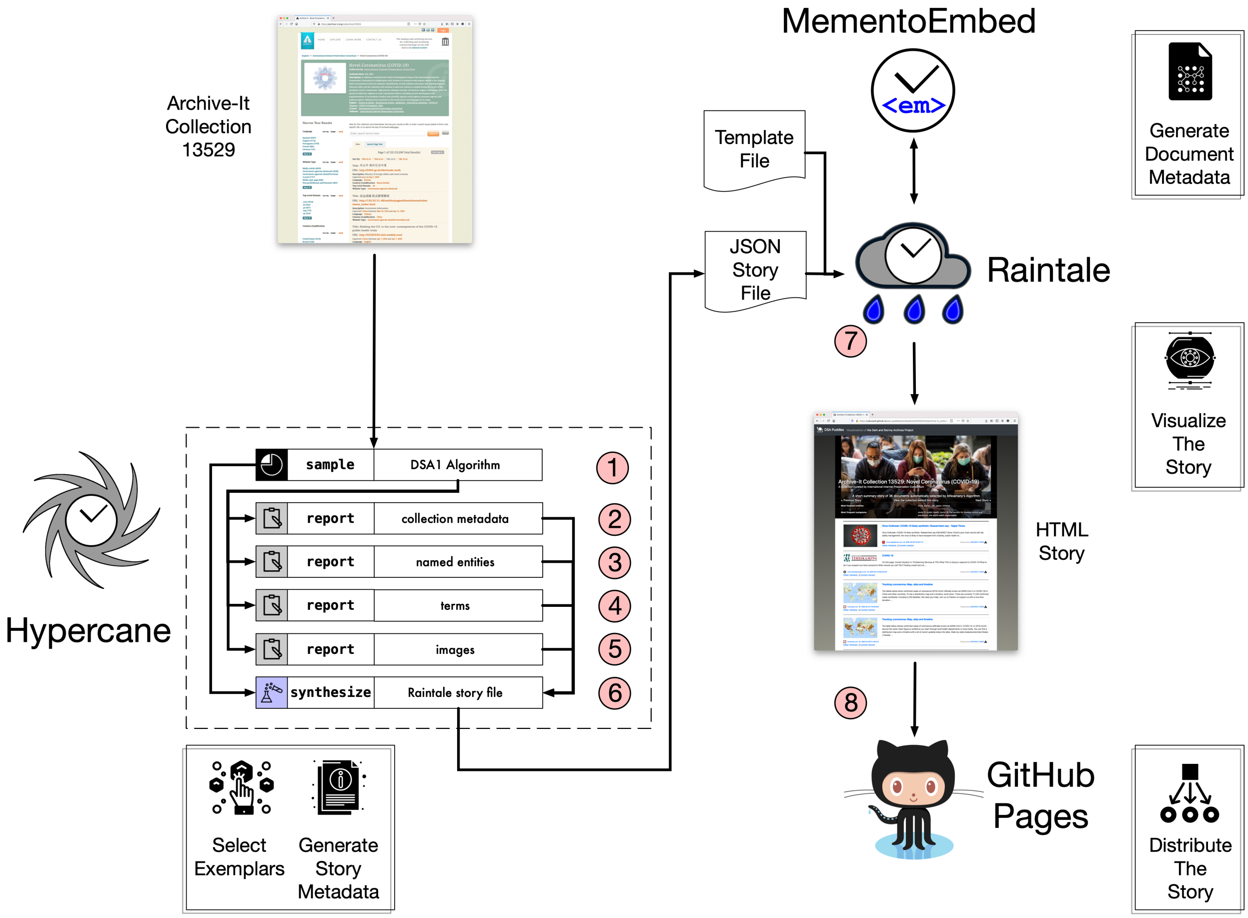
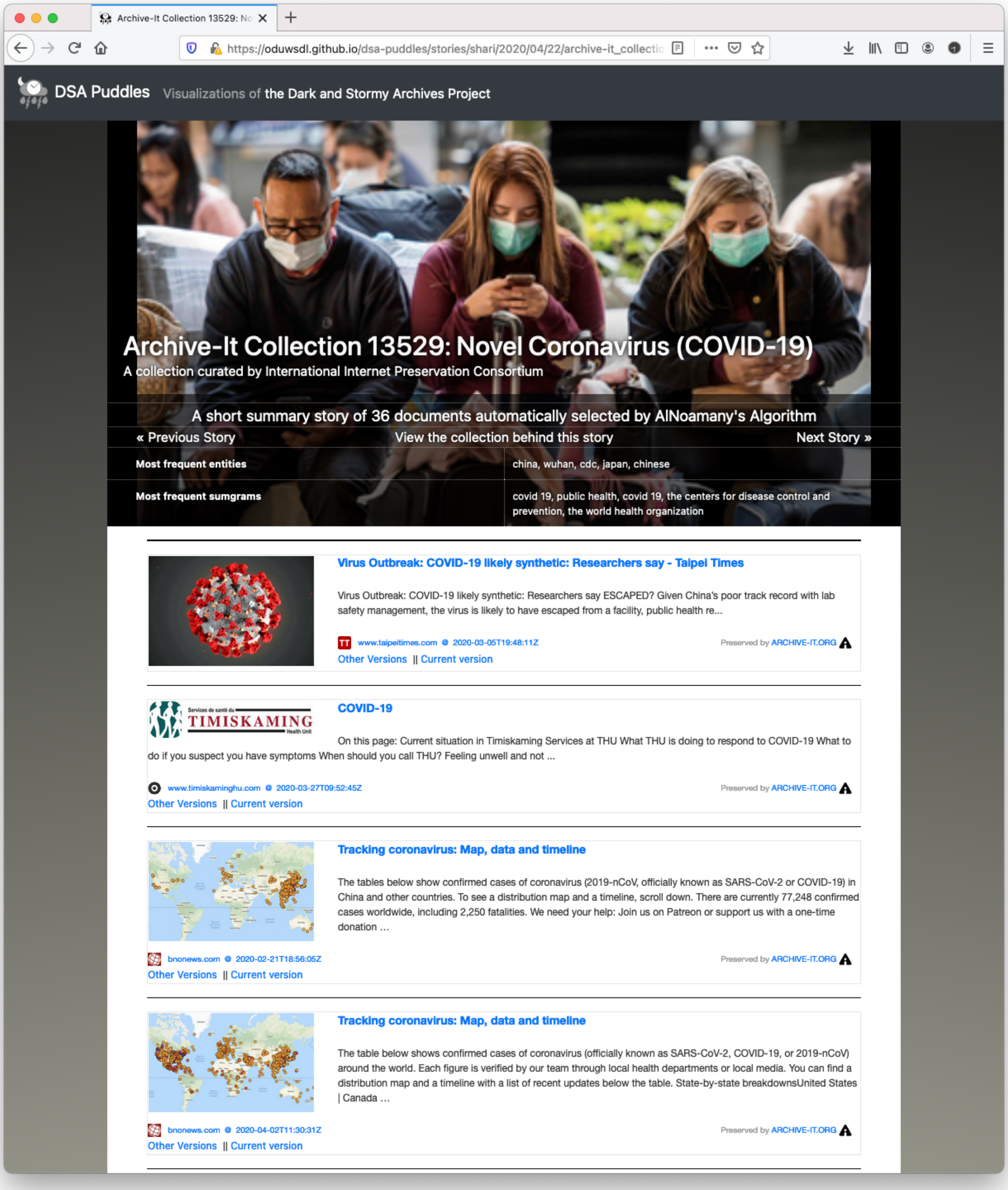


Fig. 170. A flowchart visualizing the process for producing a story from an Archive-It collection for the DSA Puddles web site.

scent than the Archive-It interface. Recall from Chapter 1 that Elbert was trying to promote his collection and make it more enticing so that people like Natasha would be able to understand its contents at a glance. The example in Figure 171 demonstrates a story with a high degree of information scent. Thus, this process meets Elbert's needs for collection promotion and Natasha's needs for understanding at a glance. Elbert can further customize this template to ensure that Raintale renders the document and story metadata in line with the needs of his institution.

To further address Natasha's need to review collections quickly, we can use this heightened information scent to allow us to compare collections, like the three collections in Figure 172 about different shooting events. Based on their Archive-It interfaces, we see some metadata and facets and may search the collections, but we must primarily rely on their titles or descriptions to know what each contains, and this may not be apparent at a glance. For example, the metadata for both


Archive-It Collection 13529: No
https://oduwsdl.github.io/dsa-puddles/stories/shari/2020/04/22/archive-it_collectio

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...

TT 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

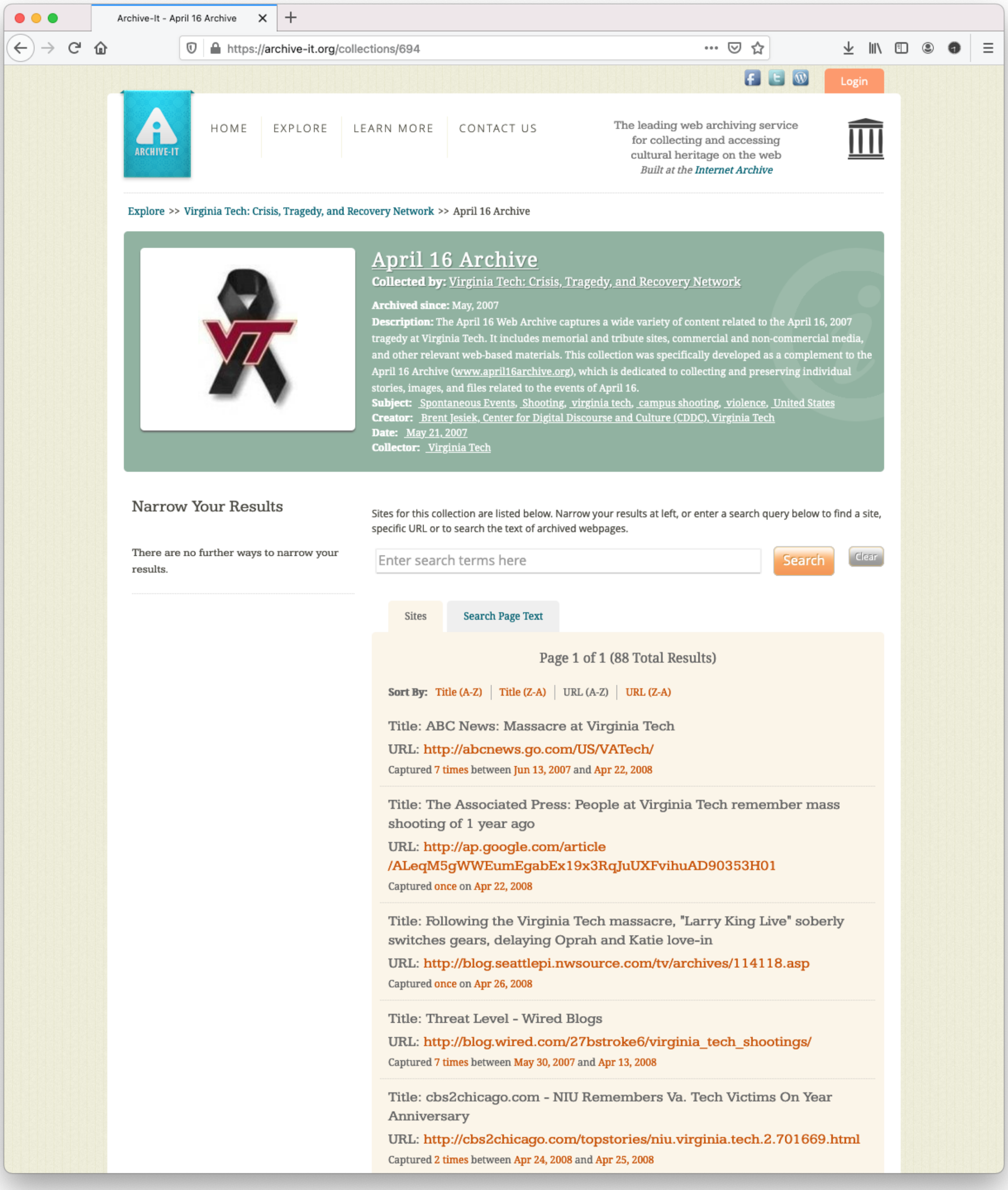
Tracking coronavirus: Map, data and timeline

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 Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Fig. 171. The International Internet Preservation Consortium's Archive-It collection *Novel Coronavirus (COVID-19)*, visualized by Raintale as a story on the DSA Puddles web site with exemplars selected by Hypercane, story metadata provided by Hypercane, and document metadata generated by MementoEmbed.



Archive-It - April 16 Archive
https://archive-it.org/collections/694
Login
ARCHIVE-IT
HOME
EXPLORE
LEARN MORE
CONTACT US

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

Explore >> Virginia Tech: Crisis, Tragedy, and Recovery Network >> April 16 Archive

April 16 Archive

Collected by: Virginia Tech: Crisis, Tragedy, and Recovery Network

Archived since: May, 2007

Description: The April 16 Web Archive captures a wide variety of content related to the April 16, 2007 tragedy at Virginia Tech. It includes memorial and tribute sites, commercial and non-commercial media, and other relevant web-based materials. This collection was specifically developed as a complement to the April 16 Archive (www.april16archive.org), which is dedicated to collecting and preserving individual stories, images, and files related to the events of April 16.

Subject: Spontaneous Events, Shooting, virginia tech, campus shooting, violence, United States
Creator: Brent Jesiek, Center for Digital Discourse and Culture (CDDC), Virginia Tech
Date: May 21, 2007
Collector: Virginia Tech

Narrow Your Results

There are no further ways to narrow your results.

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.

Page 1 of 1 (88 Total Results)

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

Title: ABC News: Massacre at Virginia Tech
URL: http://abcnews.go.com/US/VATech/
Captured 7 times between Jun 13, 2007 and Apr 22, 2008

Title: The Associated Press: People at Virginia Tech remember mass shooting of 1 year ago
URL: http://ap.google.com/article/ALeqM5gWWEumEgabEx19x3RqJuUXFvihuAD90353H01
Captured once on Apr 22, 2008

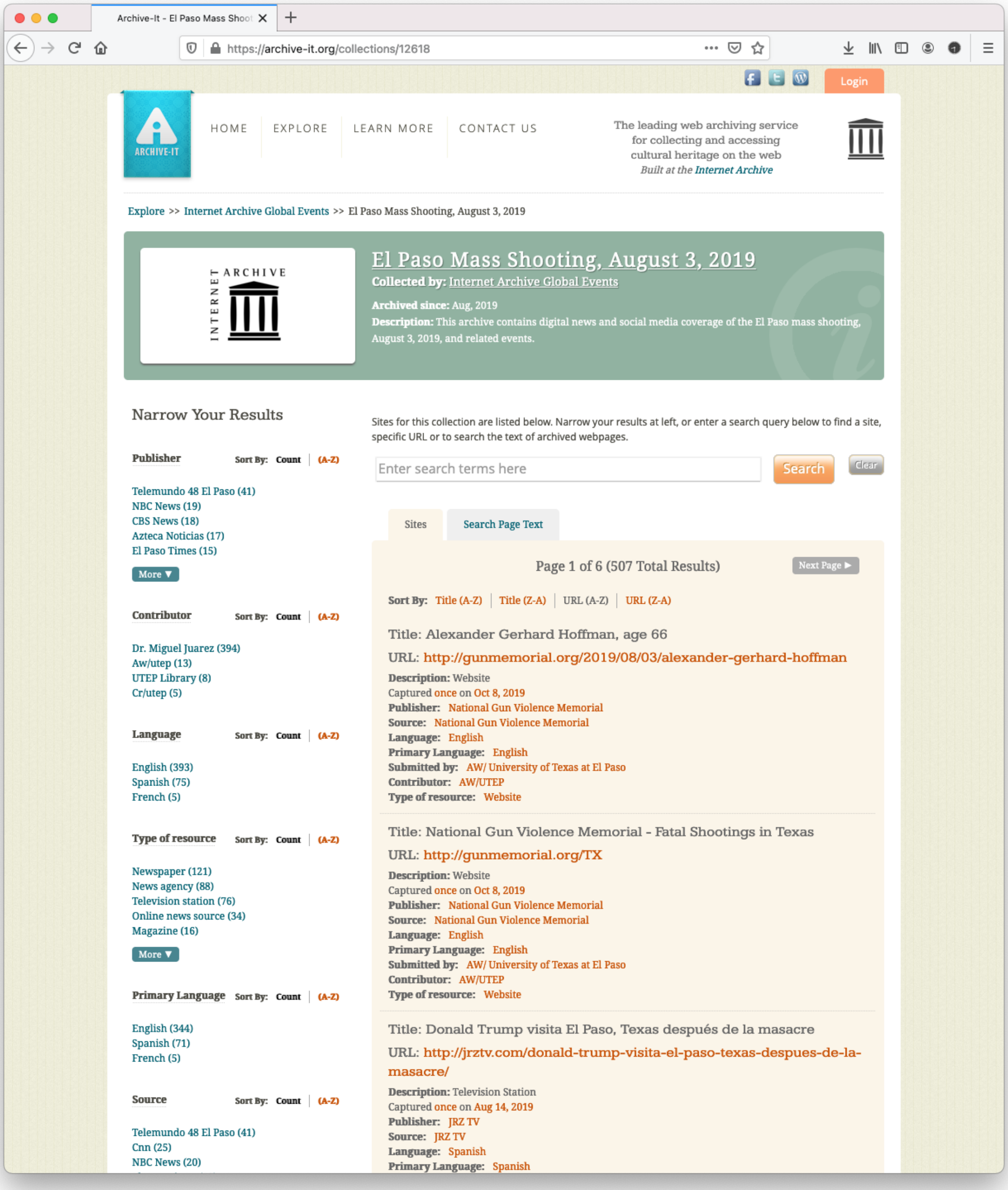
Title: Following the Virginia Tech massacre, "Larry King Live" soberly switches gears, delaying Oprah and Katie love-in
URL: http://blog.seattlepi.nwsource.com/tv/archives/114118.asp
Captured once on Apr 26, 2008

Title: Threat Level - Wired Blogs
URL: http://blog.wired.com/27bstroke6/virginia_tech_shootings/
Captured 7 times between May 30, 2007 and Apr 13, 2008

Title: cbs2chicago.com - NIU Remembers Va. Tech Victims On Year Anniversary
URL: http://cbs2chicago.com/topstories/niu.virginia.tech.2.701669.html
Captured 2 times between Apr 24, 2008 and Apr 25, 2008

(a) Virginia Tech Shooting

Fig. 172. Screenshots of three different Archive-It collections on shootings.



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

Explore >> Internet Archive Global Events >> El Paso Mass Shooting, August 3, 2019

El Paso Mass Shooting, August 3, 2019

Collected by: Internet Archive Global Events

Archived since: Aug, 2019

Description: This archive contains digital news and social media coverage of the El Paso mass shooting, August 3, 2019, and related events.

Narrow Your Results

Publisher Sort By: Count | (A-Z)

Telemundo 48 El Paso (41)
NBC News (19)
CBS News (18)
Azteca Noticias (17)
El Paso Times (15)

Contributor Sort By: Count | (A-Z)

Dr. Miguel Juarez (394)
Aw/utep (13)
UTEP Library (8)
Cr/utep (5)

Language Sort By: Count | (A-Z)

English (393)
Spanish (75)
French (5)

Type of resource Sort By: Count | (A-Z)

Newspaper (121)
News agency (88)
Television station (76)
Online news source (34)
Magazine (16)

Primary Language Sort By: Count | (A-Z)

English (344)
Spanish (71)
French (5)

Source Sort By: Count | (A-Z)

Telemundo 48 El Paso (41)
Cnn (25)
NBC News (20)

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

Page 1 of 6 (507 Total Results)

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

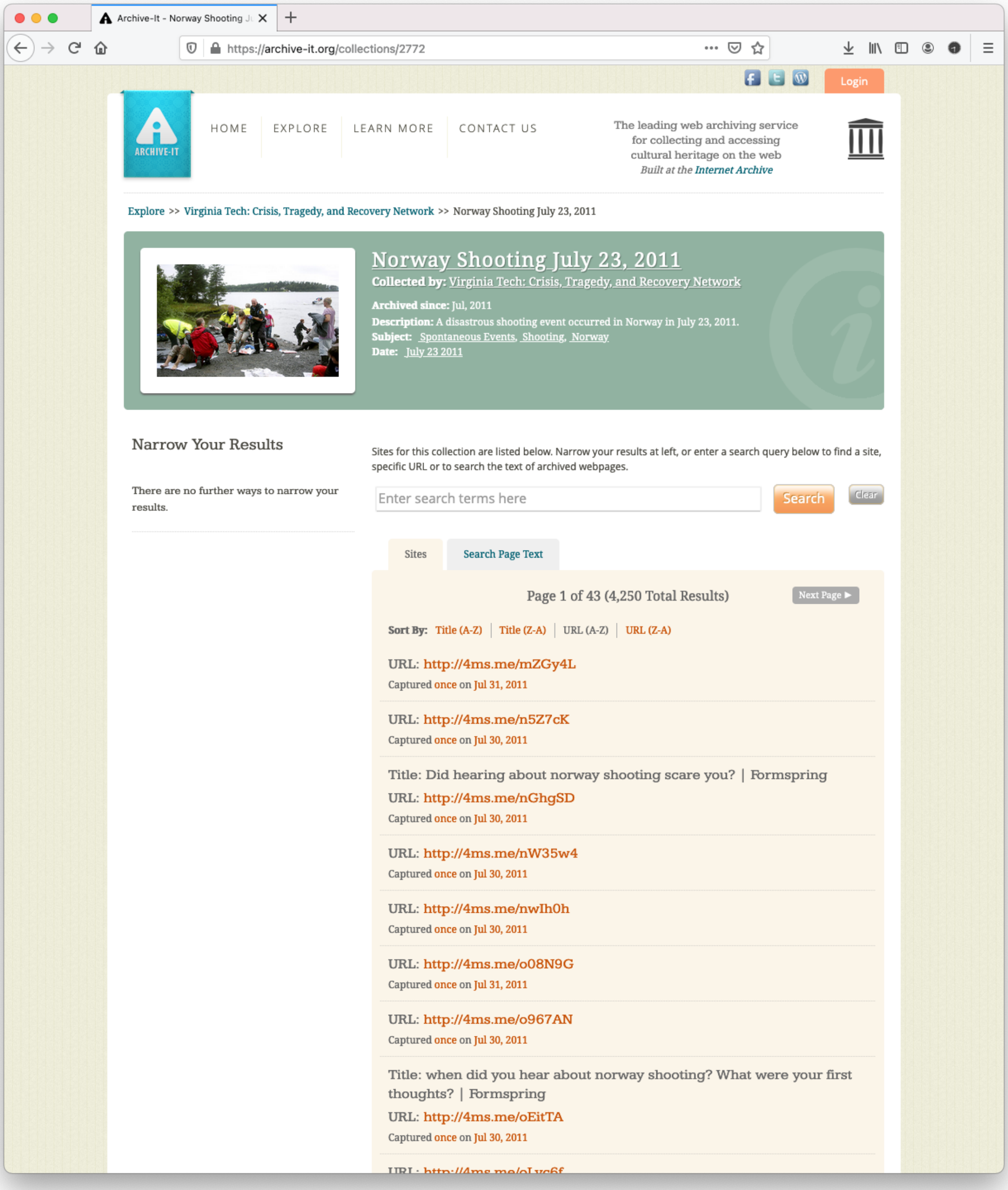
Title: Alexander Gerhard Hoffman, age 66
URL: http://gunmemorial.org/2019/08/03/alexander-gerhard-hoffman
Description: Website
Captured once on Oct 8, 2019
Publisher: National Gun Violence Memorial
Source: National Gun Violence Memorial
Language: English
Primary Language: English
Submitted by: AW/ University of Texas at El Paso
Contributor: AW/UTEP
Type of resource: Website

Title: National Gun Violence Memorial - Fatal Shootings in Texas
URL: http://gunmemorial.org/TX
Description: Website
Captured once on Oct 8, 2019
Publisher: National Gun Violence Memorial
Source: National Gun Violence Memorial
Language: English
Primary Language: English
Submitted by: AW/ University of Texas at El Paso
Contributor: AW/UTEP
Type of resource: Website

Title: Donald Trump visita El Paso, Texas después de la masacre
URL: http://jrztv.com/donald-trump-visita-el-paso-texas-despues-de-la-masacre/
Description: Television Station
Captured once on Aug 14, 2019
Publisher: JRZ TV
Source: JRZ TV
Language: Spanish
Primary Language: Spanish

(b) El Paso Shooting

Fig. 172. (Continued) Screenshots of three different Archive-It collections on shootings.


Archive-It - Norway Shooting J
https://archive-it.org/collections/2772
Login
ARCHIVE-IT
HOME
EXPLORE
LEARN MORE
CONTACT US

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

Explore >> Virginia Tech: Crisis, Tragedy, and Recovery Network >> Norway Shooting July 23, 2011

Norway Shooting July 23, 2011

Collected by: Virginia Tech: Crisis, Tragedy, and Recovery Network

Archived since: Jul, 2011

Description: A disastrous shooting event occurred in Norway in July 23, 2011.

Subject: Spontaneous Events, Shooting, Norway

Date: July 23 2011

Narrow Your Results

There are no further ways to narrow your results.

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.

Page 1 of 43 (4,250 Total Results)

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

URL: http://4ms.me/mZGy4L
Captured once on Jul 31, 2011

URL: http://4ms.me/n5Z7cK
Captured once on Jul 30, 2011

Title: Did hearing about norway shooting scare you? | Formspring
URL: http://4ms.me/nGhgSD
Captured once on Jul 30, 2011

URL: http://4ms.me/nW35w4
Captured once on Jul 30, 2011

URL: http://4ms.me/nwIh0h
Captured once on Jul 30, 2011

URL: http://4ms.me/o08N9G
Captured once on Jul 31, 2011

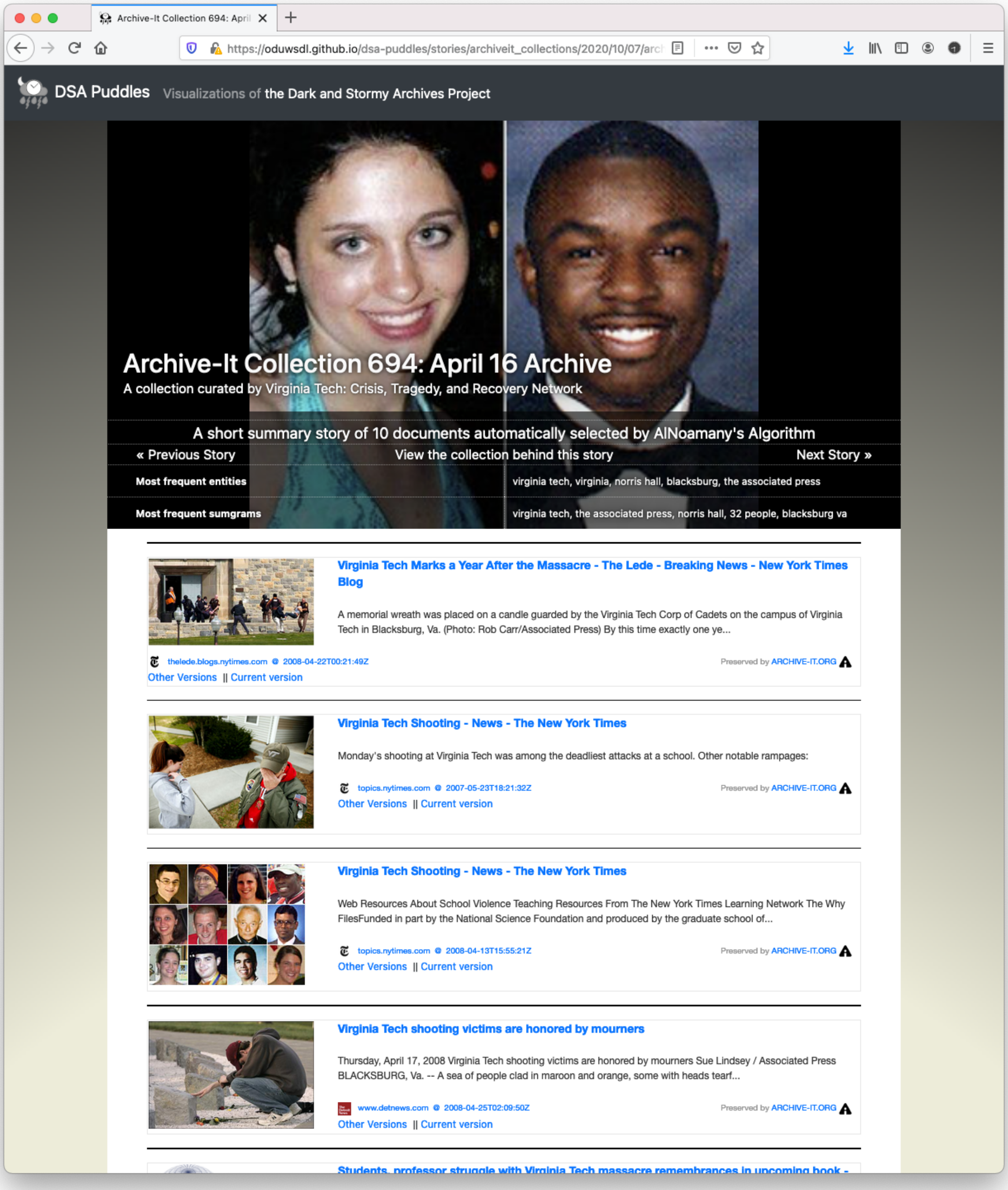
URL: http://4ms.me/o967AN
Captured once on Jul 30, 2011

Title: when did you hear about norway shooting? What were your first thoughts? | Formspring
URL: http://4ms.me/oEitTA
Captured once on Jul 30, 2011

URL: http://4ms.me/oLvc6f

(c) Norway Shooting

Fig. 172. (Continued) Screenshots of three different Archive-It collections on shootings.



Archive-It Collection 694: April
https://oduwsdl.github.io/dsa-puddles/stories/archiveit_collections/2020/10/07/arc
DSA Puddles
Visualizations of the Dark and Stormy Archives Project

Archive-It Collection 694: April 16 Archive

A collection curated by Virginia Tech: Crisis, Tragedy, and Recovery Network

A short summary story of 10 documents automatically selected by AlNoamany's Algorithm

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

Most frequent entities	virginia tech, virginia, norris hall, blacksburg, the associated press
Most frequent sumgrams	virginia tech, the associated press, norris hall, 32 people, blacksburg va

Virginia Tech Marks a Year After the Massacre - The Lede - Breaking News - New York Times Blog

A memorial wreath was placed on a candle guarded by the Virginia Tech Corp of Cadets on the campus of Virginia Tech in Blacksburg, Va. (Photo: Rob Carr/Associated Press) By this time exactly one ye...

thelede.blogs.nytimes.com @ 2008-04-22T00:21:49Z Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Virginia Tech Shooting - News - The New York Times

Monday's shooting at Virginia Tech was among the deadliest attacks at a school. Other notable rampages:

topics.nytimes.com @ 2007-05-23T18:21:32Z Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Virginia Tech Shooting - News - The New York Times

Web Resources About School Violence Teaching Resources From The New York Times Learning Network The Why FilesFunded in part by the National Science Foundation and produced by the graduate school of...

topics.nytimes.com @ 2008-04-13T15:55:21Z Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Virginia Tech shooting victims are honored by mourners

Thursday, April 17, 2008 Virginia Tech shooting victims are honored by mourners Sue Lindsey / Associated Press BLACKSBURG, Va. -- A sea of people clad in maroon and orange, some with heads tearf...

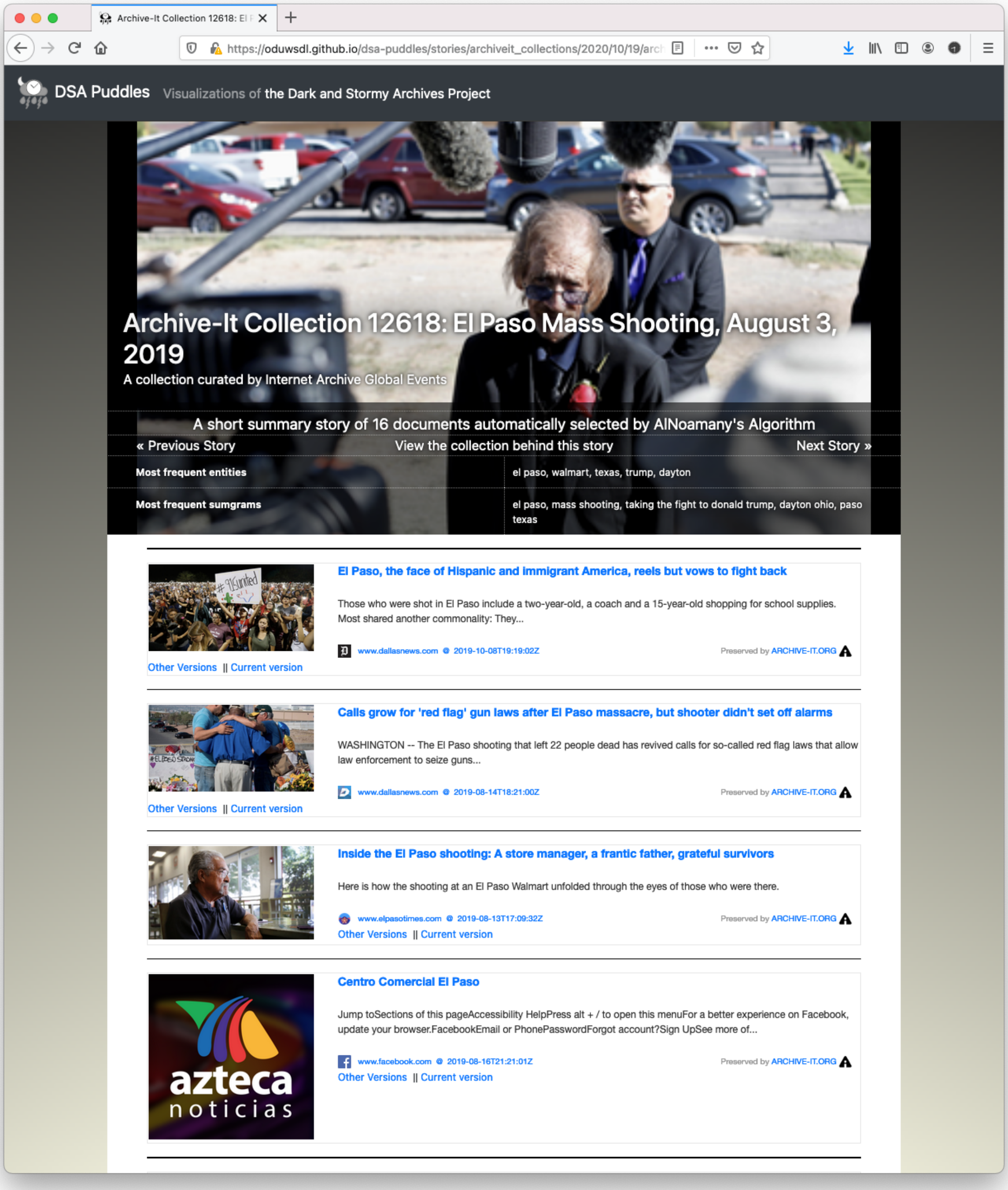
www.detnews.com @ 2008-04-25T02:09:50Z Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Students, professor struggle with Virginia Tech massacre remembrances in upcoming book -

(a) Virginia Tech Shooting

Fig. 173. We can compare different Archive-It collections from Figure 172 about shootings through their stories.


Archive-It Collection 12618: El P
https://oduwsdl.github.io/dsa-puddles/stories/archiveit_collections/2020/10/19/arch

DSA Puddles Visualizations of the Dark and Stormy Archives Project

Archive-It Collection 12618: El Paso Mass Shooting, August 3, 2019

A collection curated by Internet Archive Global Events

A short summary story of 16 documents automatically selected by AlNoamany's Algorithm

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

Most frequent entities	el paso, walmart, texas, trump, dayton
Most frequent sumgrams	el paso, mass shooting, taking the fight to donald trump, dayton ohio, paso texas

El Paso, the face of Hispanic and immigrant America, reels but vows to fight back

Those who were shot in El Paso include a two-year-old, a coach and a 15-year-old shopping for school supplies. Most shared another commonality: They...

www.dallasnews.com @ 2019-10-08T19:19:02Z Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Calls grow for 'red flag' gun laws after El Paso massacre, but shooter didn't set off alarms

WASHINGTON -- The El Paso shooting that left 22 people dead has revived calls for so-called red flag laws that allow law enforcement to seize guns...

www.dallasnews.com @ 2019-08-14T18:21:00Z Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Inside the El Paso shooting: A store manager, a frantic father, grateful survivors

Here is how the shooting at an El Paso Walmart unfolded through the eyes of those who were there.

www.elpasotimes.com @ 2019-08-13T17:09:32Z Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Centro Comercial El Paso

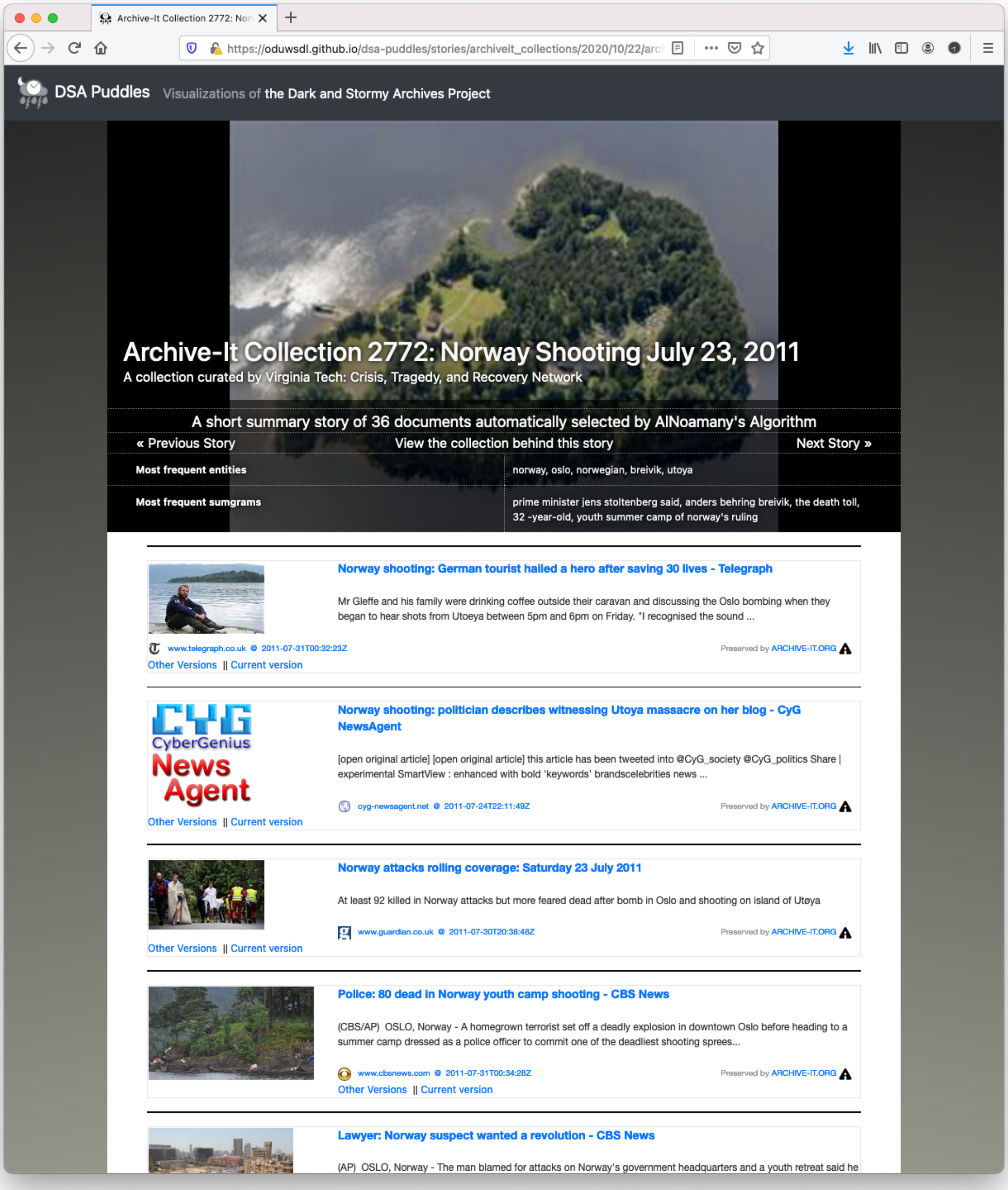
Jump toSections of this pageAccessibility HelpPress alt + / to open this menuFor a better experience on Facebook, update your browser.FacebookEmail or PhonePasswordForgot account?Sign UpSee more of...

www.facebook.com @ 2019-08-16T21:21:01Z Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

(a) El Paso Shooting

Fig. 174. (Continued) We can compare different Archive-It collections from Figure 172 about shootings through their stories.



Archive-It Collection 2772: Nor
https://oduwsdl.github.io/dsa-puddles/stories/archiveit_collections/2020/10/22/arc
DSA Puddles
Visualizations of the Dark and Stormy Archives Project

Archive-It Collection 2772: Norway Shooting July 23, 2011

A collection curated by Virginia Tech: Crisis, Tragedy, and Recovery Network

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	norway, oslo, norwegian, breivik, utoya
Most frequent sumgrams	prime minister jens stoltenberg said, anders behring breivik, the death toll, 32 -year-old, youth summer camp of norway's ruling

Norway shooting: German tourist hailed a hero after saving 30 lives - Telegraph

Mr Gleffe and his family were drinking coffee outside their caravan and discussing the Oslo bombing when they began to hear shots from Utoeya between 5pm and 6pm on Friday. "I recognised the sound ...

www.telegraph.co.uk @ 2011-07-31T00:32:23Z Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Norway shooting: politician describes witnessing Utoya massacre on her blog - CyG NewsAgent

[open original article] [open original article] this article has been tweeted into @CyG_society @CyG_politics Share | experimental SmartView : enhanced with bold 'keywords' brandscelebrities news ...

cyg-newsagent.net @ 2011-07-24T22:11:49Z Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Norway attacks rolling coverage: Saturday 23 July 2011

At least 92 killed in Norway attacks but more feared dead after bomb in Oslo and shooting on island of Utøya

www.guardian.co.uk @ 2011-07-30T20:38:48Z Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Police: 80 dead in Norway youth camp shooting - CBS News

(CBS/AP) OSLO, Norway - A homegrown terrorist set off a deadly explosion in downtown Oslo before heading to a summer camp dressed as a police officer to commit one of the deadliest shooting sprees...

www.cbsnews.com @ 2011-07-31T00:34:28Z Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Lawyer: Norway suspect wanted a revolution - CBS News

(AP) OSLO, Norway - The man blamed for attacks on Norway's government headquarters and a youth retreat said he

(a) Norway Shooting

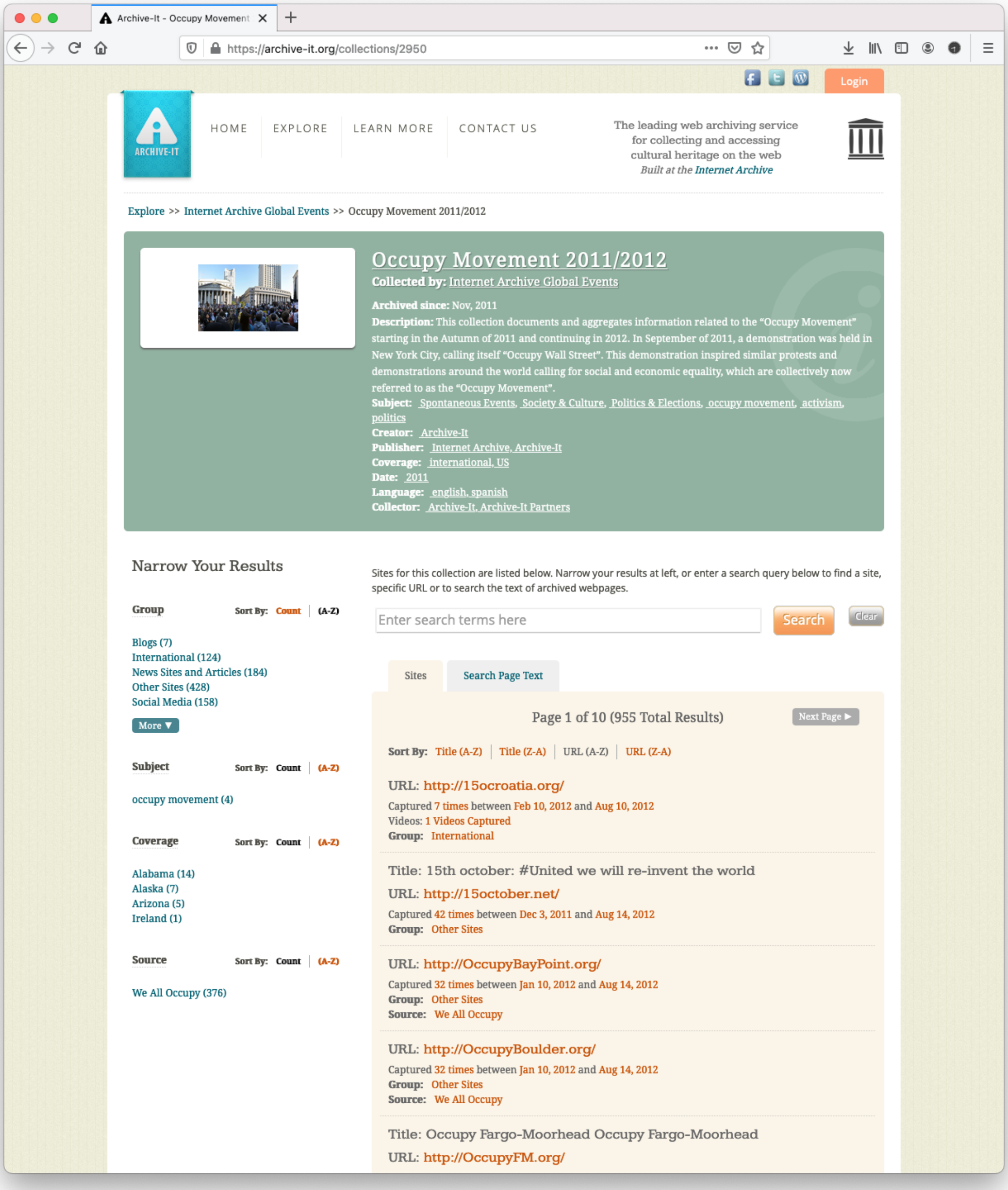
Fig. 175. (Continued) We can compare different Archive-It collections from Figure 172 about shootings through their stories.

the El Paso and the Norway shooting collections gives us a date and place but differing levels of seed metadata. The Virginia Tech collection's title does not describe a place, but its lengthy description gives us this information. In the related stories shown in Figure 175 we see victims, locations, survivors, protestors, dates, sources, headlines, and content that helps us compare, at a glance, that collection's aboutness.

GitHub Pages is not the only distribution platform we can use. We can substitute it for a social media platform. Raintale can leverage the Twitter API to create Tweets and distribute stories via that platform, as shown in Figure 176. Originally, Storify was both the visualization and distribution platform, but Storify's deprecation illustrates that we need a more flexible solution for storytelling with web archives. In this case, Raintale still manages the visualization — within the confines of Twitter — but Twitter distributes the story. By separating the concept of visualization from distribution, Raintale's templates guide the formulation of tweets. The first Tweet in the thread contains the collection's name, its collector, and its URI. Each subsequent tweet represents an exemplar memento selected from the collection. These tweets contain the generated document metadata of the memento's title, its memento-datetime, its URI-M, a browser thumbnail, and the three highest scoring images from the memento. Twitter users can add to this thread with comments and extend it further, potentially enriching the story with additional context not found in the original collection. Elbert would also be happy with this solution because it can drive engagement with his exemplars and encourage users to visit the collection.

Moreover, Elbert is not the only archivist that can benefit from these storytelling processes. Recall that Ling was trying to understand a collection that she inherited. Assume for this example that Archive-It Collection 6848 *Congress* is the collection she has inherited. Because archivists have different needs, their stories require different information scent than stories presented to the general public. Figure 177 shows how Ling created her story:

1. Ask Hypercane to select exemplars from this collection (6848) with the DSA2 algorithm
2. Ask Hypercane to produce the collection metadata (e.g., collection name, collecting organization) for this collection
3. Ask Hypercane to generate seed statistics for this collection – this includes the seed features mentioned in Chapter 4.1
4. Ask Hypercane to produce a growth curve statistics and the growth curve for this collection – as mentioned in Chapter 4.1
5. Ask Hypercane to generate metadata statistics for the collection


Archive-It - Occupy Movement
https://archive-it.org/collections/2950
Login
ARCHIVE-IT
HOME
EXPLORE
LEARN MORE
CONTACT US
The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive

Explore >> Internet Archive Global Events >> Occupy Movement 2011/2012

Occupy Movement 2011/2012

Collected by: Internet Archive Global Events

Archived since: Nov, 2011

Description: This collection documents and aggregates information related to the "Occupy Movement" starting in the Autumn of 2011 and continuing in 2012. In September of 2011, a demonstration was held in New York City, calling itself "Occupy Wall Street". This demonstration inspired similar protests and demonstrations around the world calling for social and economic equality, which are collectively now referred to as the "Occupy Movement".

Subject: Spontaneous Events, Society & Culture, Politics & Elections, occupy movement, activism, politics

Creator: Archive-It

Publisher: Internet Archive, Archive-It

Coverage: international, US

Date: 2011

Language: english, spanish

Collector: Archive-It, Archive-It Partners

Narrow Your Results

Group Sort By: Count | (A-Z)

Blogs (7)
International (124)
News Sites and Articles (184)
Other Sites (428)
Social Media (158)

More ▼

Subject Sort By: Count | (A-Z)

occupy movement (4)

Coverage Sort By: Count | (A-Z)

Alabama (14)
Alaska (7)
Arizona (5)
Ireland (1)

Source Sort By: Count | (A-Z)

We All Occupy (376)

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 10 (955 Total Results) Next Page ▶

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

URL: http://15ocroatia.org/
Captured 7 times between Feb 10, 2012 and Aug 10, 2012
Videos: 1 Videos Captured
Group: International

Title: 15th october: #United we will re-invent the world
URL: http://15october.net/
Captured 42 times between Dec 3, 2011 and Aug 14, 2012
Group: Other Sites

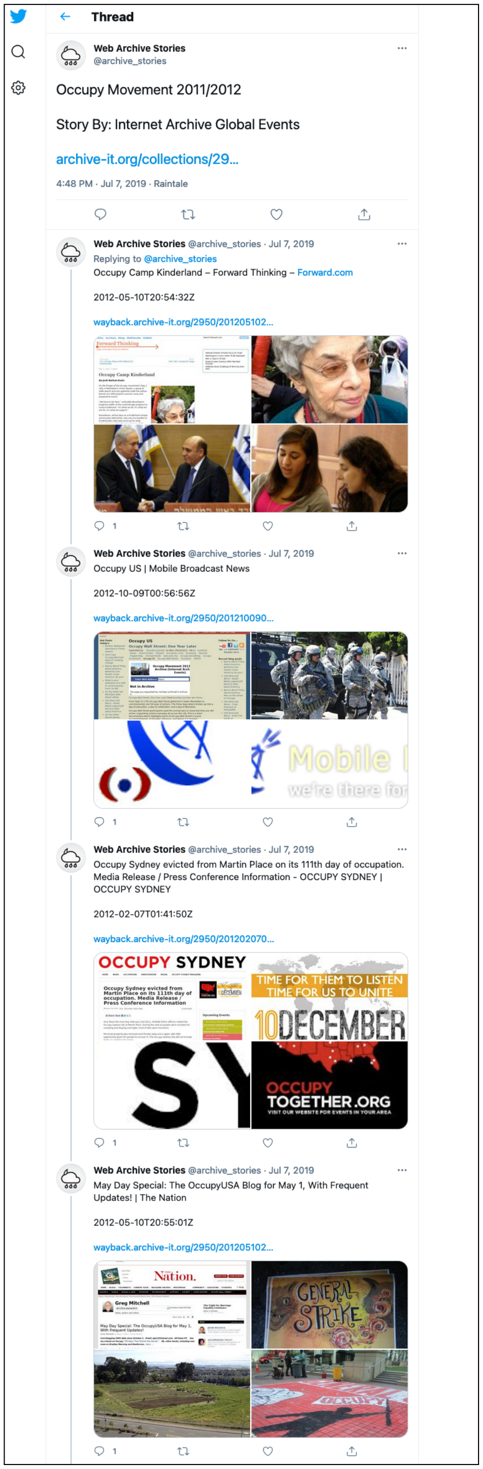
URL: http://OccupyBayPoint.org/
Captured 32 times between Jan 10, 2012 and Aug 14, 2012
Group: Other Sites
Source: We All Occupy

URL: http://OccupyBoulder.org/
Captured 32 times between Jan 10, 2012 and Aug 14, 2012
Group: Other Sites
Source: We All Occupy

Title: Occupy Fargo-Moorhead Occupy Fargo-Moorhead
URL: http://OccupyFM.org/

(a) Archive-It Interface

Fig. 176. Archive-It Collection 2950: *Occupy Movement 2011/2012* represented by its regular interface and as a Twitter story.



Web Archive Stories
@archive_stories

Occupy Movement 2011/2012

Story By: Internet Archive Global Events

archive-it.org/collections/29...

4:48 PM · Jul 7, 2019 · Raintale

Web Archive Stories @archive_stories · Jul 7, 2019
Replying to @archive_stories
Occupy Camp Kinderland – Forward Thinking – Forward.com

2012-05-10T20:54:32Z

wayback.archive-it.org/2950/201205102...

Web Archive Stories @archive_stories · Jul 7, 2019
Occupy US | Mobile Broadcast News

2012-10-09T00:56:56Z

wayback.archive-it.org/2950/201210090...

Web Archive Stories @archive_stories · Jul 7, 2019
Occupy Sydney evicted from Martin Place on its 111th day of occupation. Media Release / Press Conference Information - OCCUPY SYDNEY | OCCUPY SYDNEY

2012-02-07T01:41:50Z

wayback.archive-it.org/2950/201202070...

Web Archive Stories @archive_stories · Jul 7, 2019
May Day Special: The OccupyUSA Blog for May 1, With Frequent Updates! | The Nation

2012-05-10T20:55:01Z

wayback.archive-it.org/2950/201205102...

(b) Twitter story

Fig. 176. (Continued) Archive-It Collection 2950: *Occupy Movement 2011/2012* represented by its regular interface and as a Twitter story.

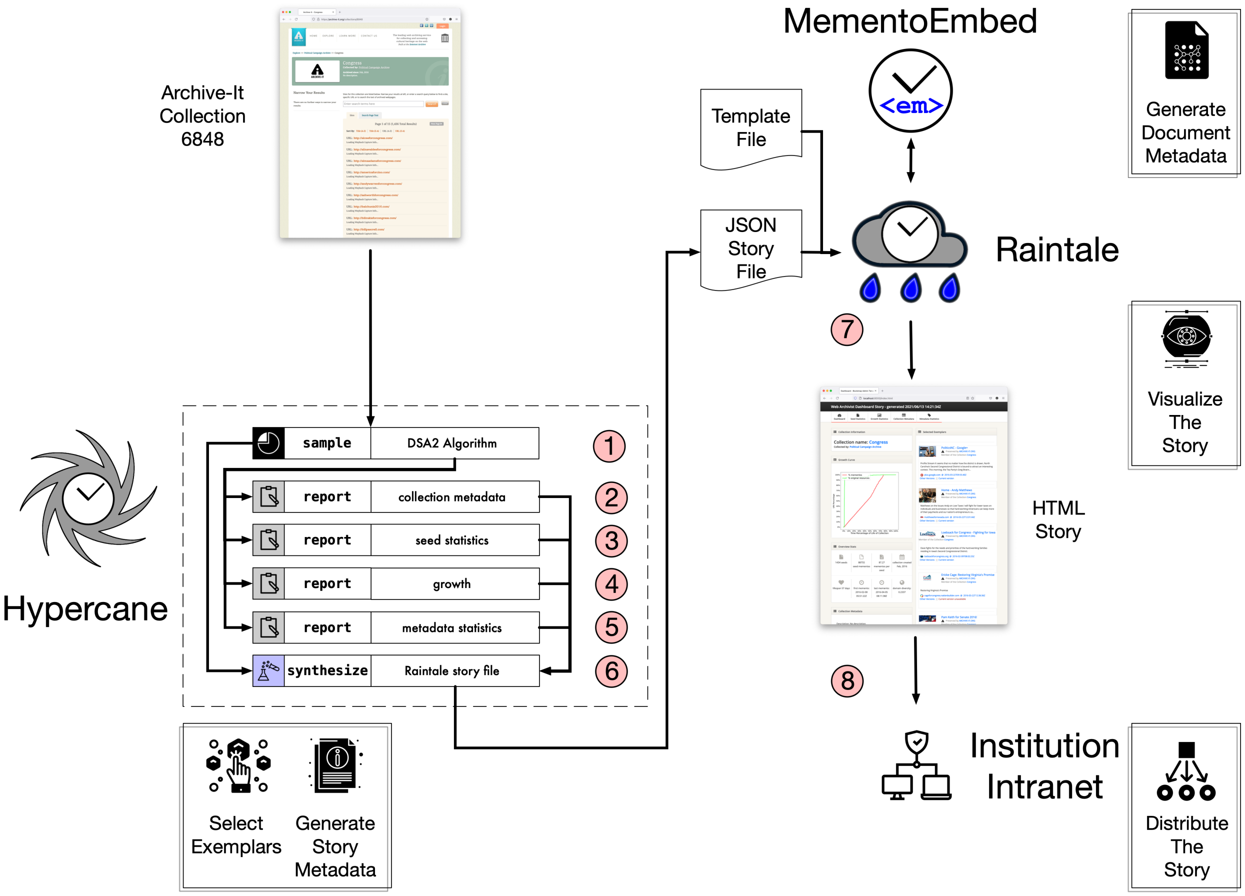


Fig. 177. The process for creating an archivist dashboard story.

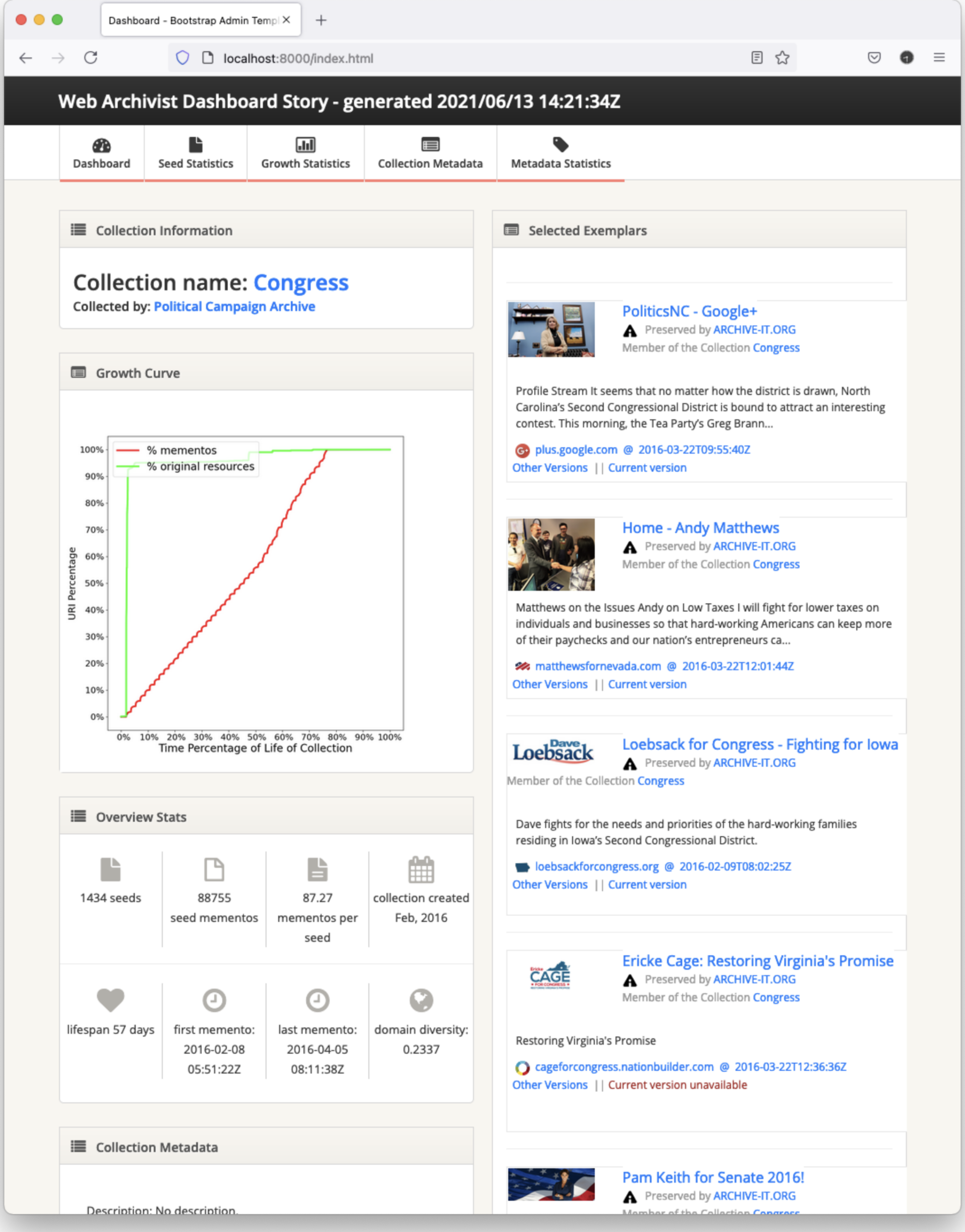
6. Synthesize the exemplars and these statistics into a JSON Raintale story file
7. Create a custom template for the need of archivists that includes the variables for these statistics and ask Raintale to generate a story using this custom template and the JSON Raintale story file
8. Distribute the story locally on a web server for other archivists at the same institution so others can benefit from this story's knowledge

Here we combine the structural features and other statistics that archivists might find to be beneficial along with selected exemplars. This way the archivist themselves gets the gist of a collection under their care so that they can make decisions about it. In Figure 178 we see the result of her work. This story contains the collection name and the collecting organization. We have links back to Archive-It collection 6848 and the organization that created it. This story includes the collection growth curve so that an archivist understands the collection's curatorial history. We also have overview stats, like the number of seeds, number of mementos, the number of mementos-per-seed (not mentioned in Chapter 4 but introduced in Chapter 7), when the collection was created, its lifespan, the dates of its first and last mementos, and its domain diversity. Further down we have the other seed, growth and metadata statistics generated by various Hypercane reports. On the right, we have a set of exemplars represented by MementoEmbed cards. With this information, an archivist gets a snapshot of the history, contents, and an overview of the collection.

Through this process, Ling can create a story that is distributed locally within her institution. Other archivists can learn from this story and understand the collection that Ling now maintains. Based on this story, Ling can understand how the collection grew, the diversity of its sources, additional statistics from its structural features, and also view a list of exemplars to give her some idea of the collection at a glance. For example, the collection's name is *Congress*. When she inherited the collection, she thought it was about news coverage and official documentation from members of the US Congress. However, now, thanks to the exemplars, she understands that the goal of this collection was to preserve congressional candidate web pages.

Rustam wanted to view how a specific page changed through time, an FPST story. Rustam is using storytelling to explore a collection. He is willing to be a human-in-the-loop who makes decisions about which exemplars are part of his story. He still needs Hypercane for its identification and filtering capabilities and will use Raintale to create a visualization of the output. Here is how Rustam created his story (Figure 179):

1. Rustam asks Hypercane to `identify` all URI-Rs from a collection about the Boston Marathon Bombing (3649)



Web Archivist Dashboard Story - generated 2021/06/13 14:21:34Z

Collection Information

Collection name: Congress

Collected by: Political Campaign Archive

Growth Curve

Overview Stats

1434 seeds	88755 seed mementos	87.27 mementos per seed	collection created Feb, 2016
lifespan 57 days	first memento: 2016-02-08 05:51:22Z	last memento: 2016-04-05 08:11:38Z	domain diversity: 0.2337

Collection Metadata

Description: No description

Selected Exemplars

PoliticsNC - Google+
Preserved by ARCHIVE-IT.ORG
Member of the Collection Congress

Profile Stream It seems that no matter how the district is drawn, North Carolina's Second Congressional District is bound to attract an interesting contest. This morning, the Tea Party's Greg Brann...

plus.google.com @ 2016-03-22T09:55:40Z
Other Versions || Current version

Home - Andy Matthews
Preserved by ARCHIVE-IT.ORG
Member of the Collection Congress

Matthews on the Issues Andy on Low Taxes I will fight for lower taxes on individuals and businesses so that hard-working Americans can keep more of their paychecks and our nation's entrepreneurs ca...

matthewsfornevada.com @ 2016-03-22T12:01:44Z
Other Versions || Current version

Loebsack for Congress - Fighting for Iowa
Preserved by ARCHIVE-IT.ORG
Member of the Collection Congress

Dave fights for the needs and priorities of the hard-working families residing in Iowa's Second Congressional District.

loebsackforcongress.org @ 2016-02-09T08:02:25Z
Other Versions || Current version

Ericke Cage: Restoring Virginia's Promise
Preserved by ARCHIVE-IT.ORG
Member of the Collection Congress

Restoring Virginia's Promise

cageforcongress.nationbuilder.com @ 2016-03-22T12:36:36Z
Other Versions || Current version unavailable

Pam Keith for Senate 2016!
Preserved by ARCHIVE-IT.ORG
Member of the Collection Congress

Fig. 178. Stories can be customized to meet the needs of archivists, with collection statistics in addition to exemplar surrogates.

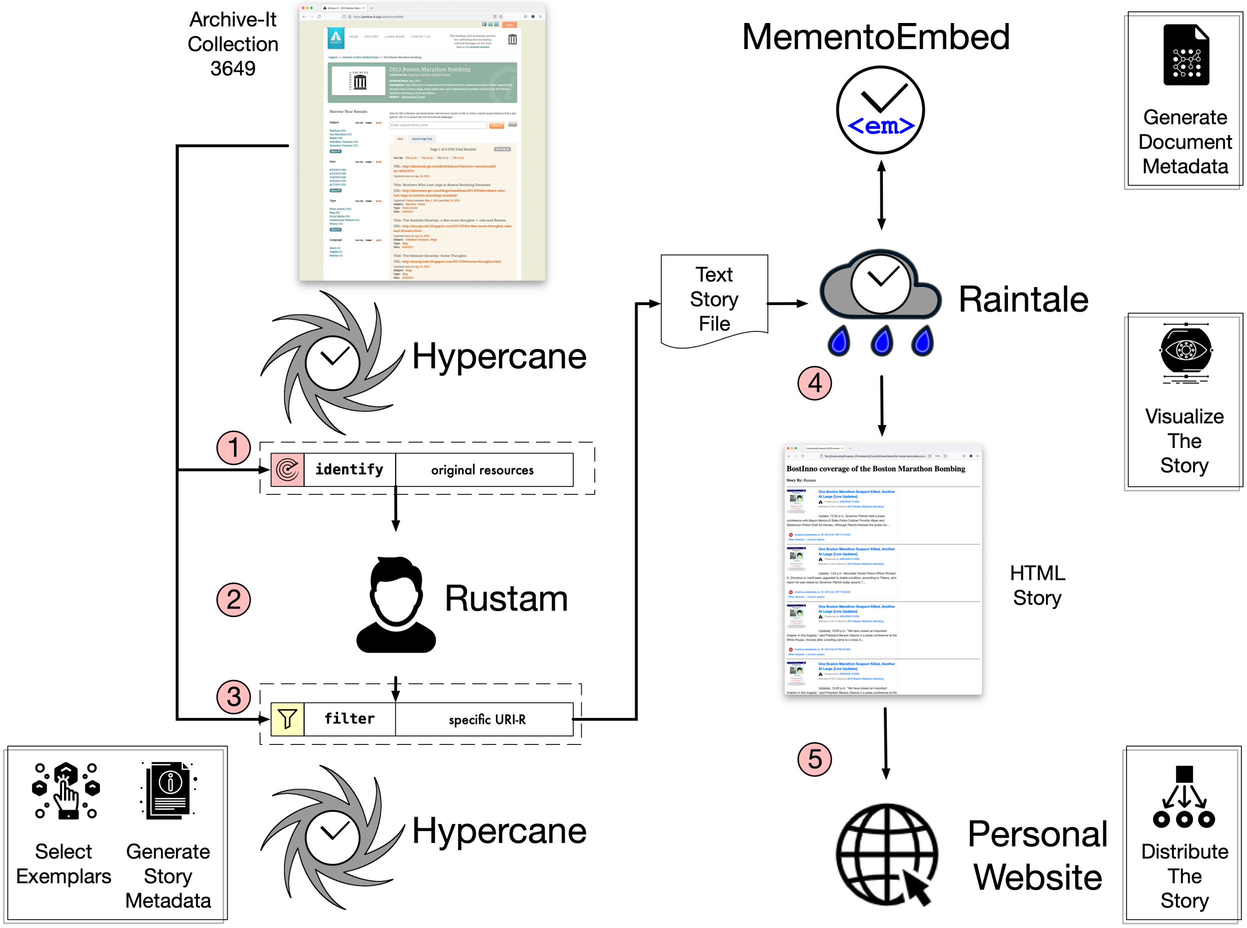
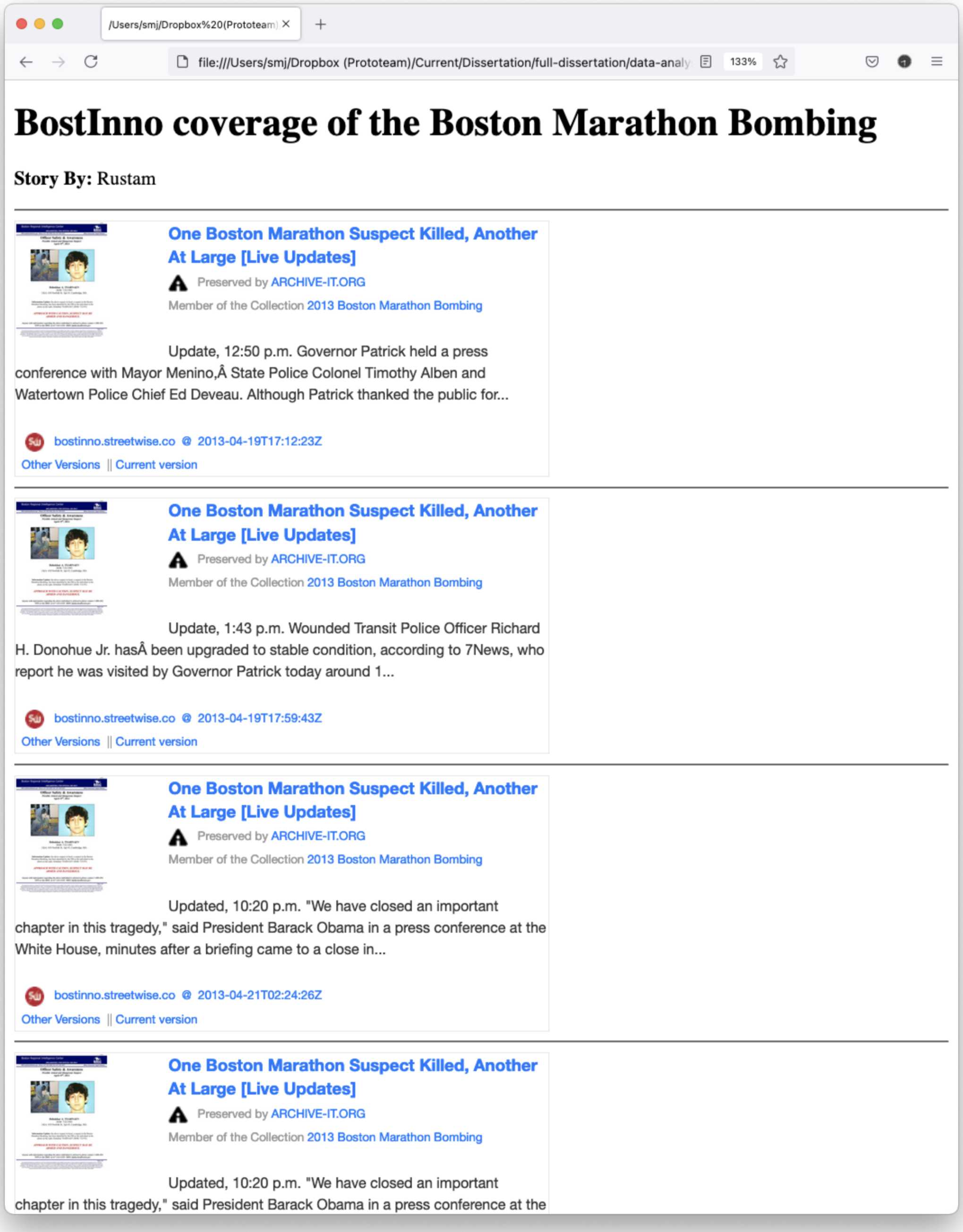


Fig. 179. Visitors can generate their own stories to explore aspects of public web archive collections. Here Rustam has applied the DSA tools to study how BostInno covered the Boston Marathon Bombing. Rustam applies Hypercane twice, but serves as a human-in-the-loop and decides which URI-R the FPST story includes.


/Users/smj/Dropbox%20(Prototeam)
file:///Users/smj/Dropbox (Prototeam)/Current/Dissertation/full-dissertation/data-analy
133%

BostInno coverage of the Boston Marathon Bombing

Story By: Rustam

One Boston Marathon Suspect Killed, Another At Large [Live Updates]

Preserved by ARCHIVE-IT.ORG

Member of the Collection 2013 Boston Marathon Bombing

Update, 12:50 p.m. Governor Patrick held a press conference with Mayor Menino,Â State Police Colonel Timothy Alben and Watertown Police Chief Ed Deveau. Although Patrick thanked the public for...

bostinno.streetwise.co @ 2013-04-19T17:12:23Z

Other Versions || Current version

One Boston Marathon Suspect Killed, Another At Large [Live Updates]

Preserved by ARCHIVE-IT.ORG

Member of the Collection 2013 Boston Marathon Bombing

Update, 1:43 p.m. Wounded Transit Police Officer Richard H. Donohue Jr. hasÂ been upgraded to stable condition, according to 7News, who report he was visited by Governor Patrick today around 1...

bostinno.streetwise.co @ 2013-04-19T17:59:43Z

Other Versions || Current version

One Boston Marathon Suspect Killed, Another At Large [Live Updates]

Preserved by ARCHIVE-IT.ORG

Member of the Collection 2013 Boston Marathon Bombing

Updated, 10:20 p.m. "We have closed an important chapter in this tragedy," said President Barack Obama in a press conference at the White House, minutes after a briefing came to a close in...

bostinno.streetwise.co @ 2013-04-21T02:24:26Z

Other Versions || Current version

One Boston Marathon Suspect Killed, Another At Large [Live Updates]

Preserved by ARCHIVE-IT.ORG

Member of the Collection 2013 Boston Marathon Bombing

Updated, 10:20 p.m. "We have closed an important chapter in this tragedy," said President Barack Obama in a press conference at the

Fig. 180. This screenshot displays Rustam's story applying the DSA tools to study how BostInno covered the Boston Marathon Bombing. Note how, even though the titles and striking images are the same, the description and memento-datetimes is different in each card.

2. Rustam manually reviewed the list of URI-Rs and decided that `http://bostinno.streetwise.co/2013/04/19/boston-marathon-suspect-dead-another-at-large-after-standoff/` may be of interest

3. Rustam filters collection 3649 so that the output would only include mementos for that URI-R – these are our exemplars

4. Rustam asks Raintale to generate a story using its default HTML template from Rustam's exemplars so he does not need to specify a template

5. Rustam shares this HTML story on his personal web site

Rustam is interested in a simple visualization of the document metadata. Because he his applying storytelling more toward learning about a specific aspect of the collection, he does not need the additional entities and other data that we have generated for other stories. This is an example of how storytelling can be used to expose a specific aspect of a collection for that story's creator and others. In Figure 180, we see a story that presents a single resource over time. With each card, we get an update of what the public knew about the Boston Marathon Bombing at that particular moment.

6.3 STORYTELLING WITH GENERAL WEB ARCHIVES

Remember that we are not limited to summarizing a web archive collection from Archive-It. These tools work with any Memento-compliant web archive. Recall that Olayinka wanted to understand the news on specific dates and times. There is not a single web archive collection that specializes in the news for the long term. However, we can apply other tools to select exemplars and leverage news articles from the Internet Archive and Archive.Today. StoryGraph [239, 240, 241, 242, 247] analyzes the news every 10 minutes to determine the biggest news stories at that moment in time. Figure 181 shows how we still satisfy the storytelling processes by including other tools. In this case, StoryGraph selects exemplar URI-Rs, and Hypercane preserves them as exemplar URI-Ms before providing story metadata. Raintale and MementoEmbed execute as before. Because of the addition of StoryGraph and because Hypercane uses the ArchiveNow library [22] to preserve original resources (URI-Rs) as mementos (URI-Ms), we refer to the daily news summary process as SHARI (StoryGraph Hypercane ArchiveNow Raintale Integration) [161].

A more detailed form of the SHARI process, shown in Figure 182, is:

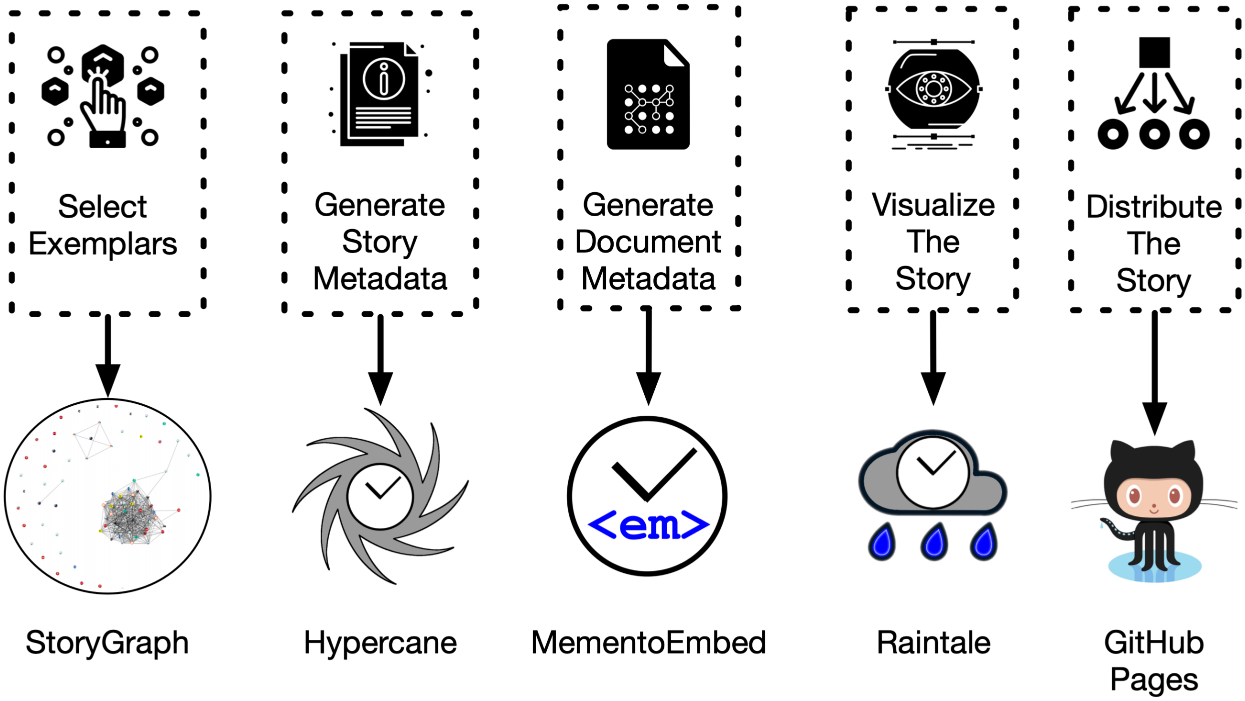


Fig. 181. Storytelling to visualize the biggest news story of the day — how the SHARI process maps to our five storytelling processes.

1. Extract URI-Rs from StoryGraph for the given date based on the longest living news story for that day

2. Ask Hypercane to preserve the original resoruce news article URI-Rs as mementos and return URI-Ms through its `identify` action to produce our exemplars

3. Ask Hypercane to produce a report of named entities from these exemplars

4. Ask Hypercane to produce a report of terms from these exemplars

5. Ask Hypercane to produce a report about the images found in these exemplars, and select the best striking image to represent this story

6. Ask Hypercane to synthesize the exemplars and these reports into a JSON Raintale story file

7. Ask Raintale to generate the story using SHARI's custom template that provides variables for all story metadata and formats exemplars

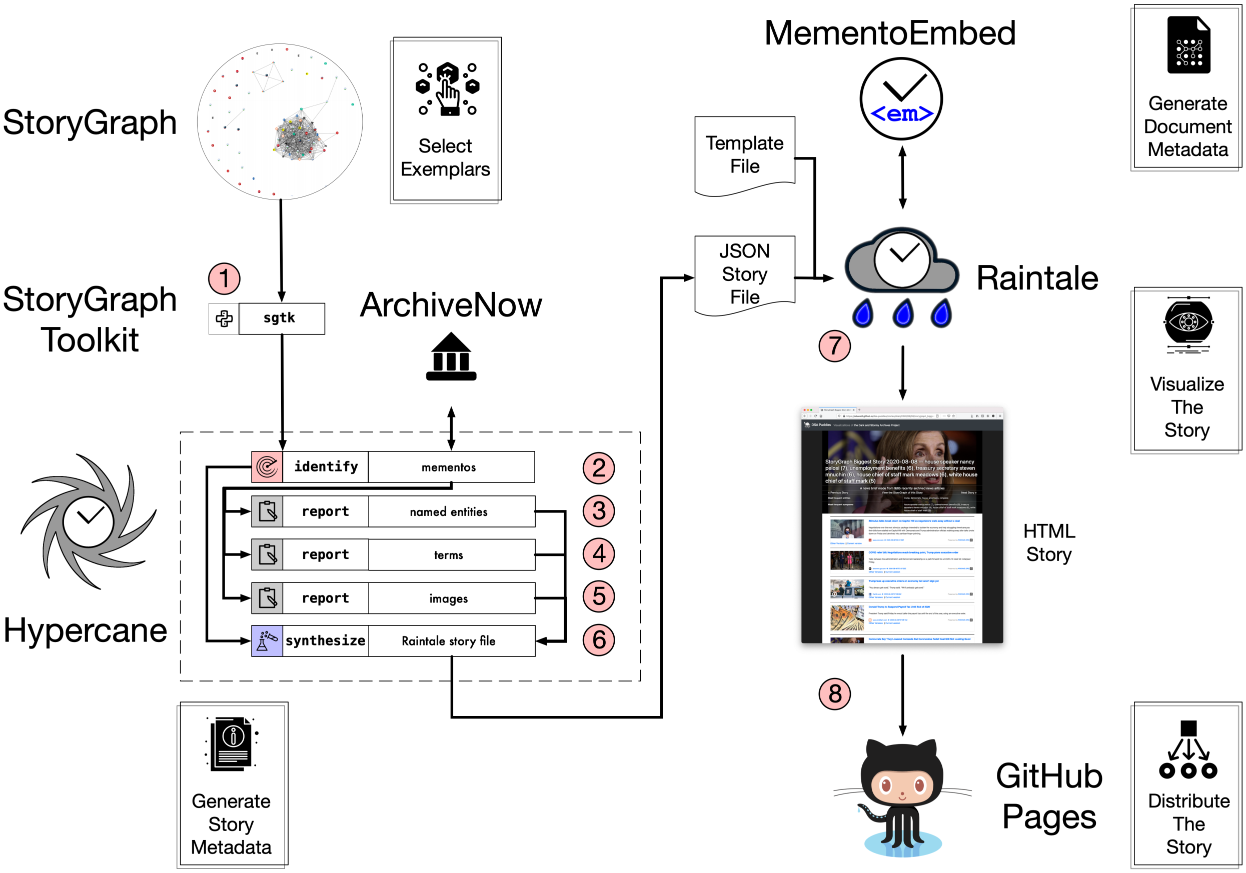


Fig. 182. The SHARI process for creating a story for the biggest news story of the day.

   (a) Raintale then asks MementoEmbed to generate the document metadata for each exemplar based on the variables from the template

   (b) Raintale visualizes the story by applying the template and this metadata

8. We commit the story to GitHub, and GitHub Pages distributes the story for us

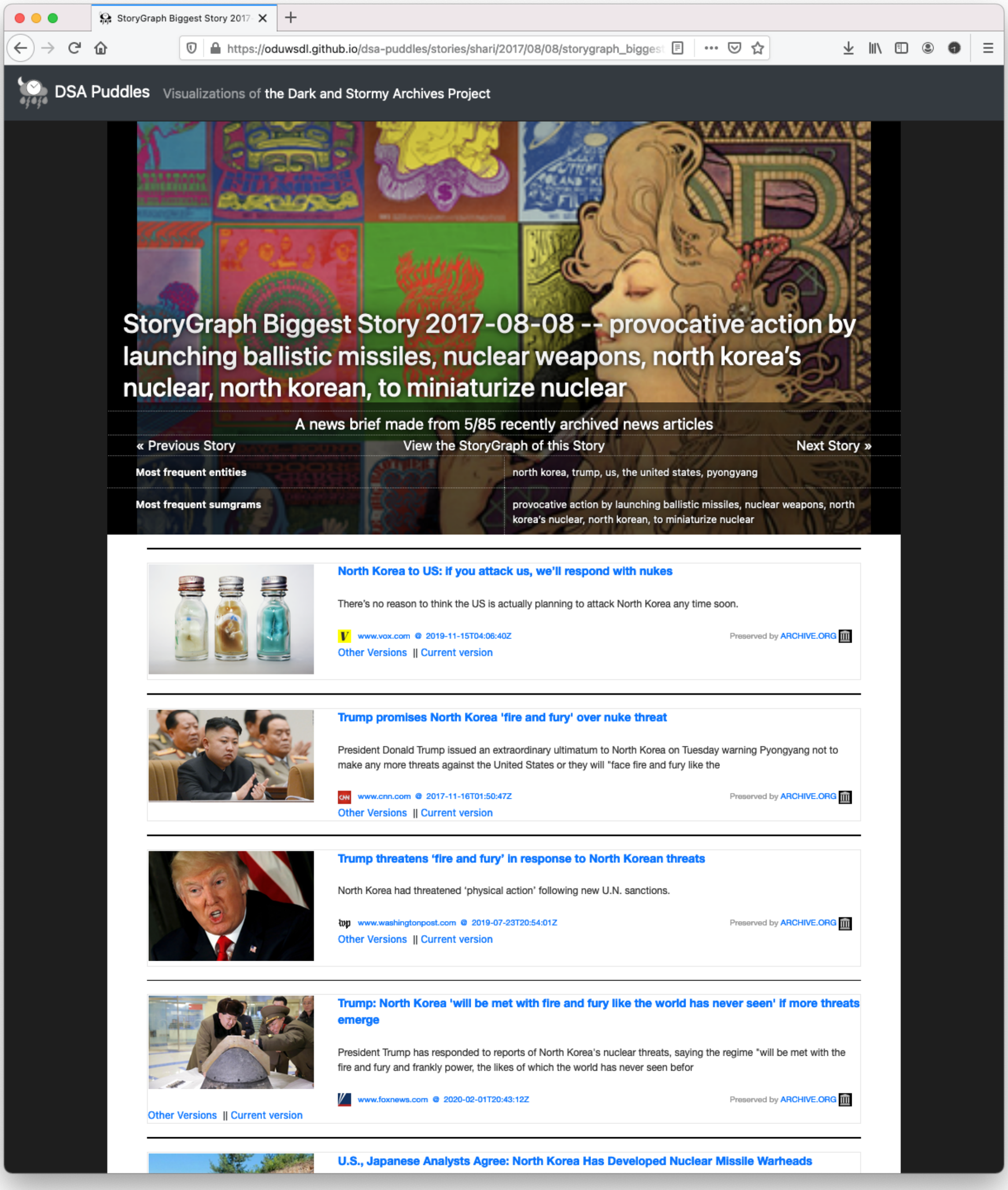
Thanks to StoryGraph, we can produce SHARI stories as far back as 2017, when Alexander Nwala first brought StoryGraph online. Figure 186 shows screenshots of four SHARI stories, each from August 8 of different years. By comparing them like this, Olayinka can see how things that seemed eternal at the time were forgotten by the following year. In 2017 on that date, the news was covering nuclear provocation by North Korea. By 2018, the discussion was about US Congressional elections. In 2019, the US was reeling from the events of dual shootings at El Paso and Dayton. By 2020, the world was reeling from COVID-19, and Congress was debating economic stimulus for the United States. Olayinka and other historians can apply these stories to understand what was happening on a given date, and what information we had at the time compared to what we know now. SHARI does not summarize web archive collections but adapts the collections of StoryGraph instead and is one of many additional possibilities for storytelling with web archives.

6.4 CHAPTER SUMMARY

In this chapter, we covered the final processes of our model of storytelling with web archives. We introduced Raintale, a tool that visualizes the story and document metadata needed for storytelling. Raintale is a client of MementoEmbed. It requires a story file containing URI-Ms. It also needs a template dictating how it should render those URI-Ms as a story. Raintale can save its stories to files in different text formats, like HTML or Markdown, depending on the template. Raintale also leverages its templating capability to format stories for social media to share them on Facebook or Twitter.

Raintale is the final tool that we will cover. By combining it with others, we satisfy all of the storytelling processes. Hypercane selects exemplars and generates story metadata. MementoEmbed generates document metadata. Raintale visualizes our stories and helps us distribute them with other services. We covered several examples of these storytelling processes with GitHub Pages, Twitter, and personal and institutional websites.

We covered several examples of these storytelling processes in use with both GitHub pages and Twitter. We illustrated how stories have a higher information scent than the standard Archive-It interface and thus allow us to compare collections to one another more easily. We also mentioned


StoryGraph Biggest Story 2017
https://oduwsdl.github.io/dsa-puddles/stories/shari/2017/08/08/storygraph_biggest

DSA Puddles Visualizations of the Dark and Stormy Archives Project

StoryGraph Biggest Story 2017-08-08 -- provocative action by launching ballistic missiles, nuclear weapons, north korea's nuclear, north korean, to miniaturize nuclear

A news brief made from 5/85 recently archived news articles

« Previous Story View the StoryGraph of this Story Next Story »

Most frequent entities	north korea, trump, us, the united states, pyongyang
Most frequent sumgrams	provocative action by launching ballistic missiles, nuclear weapons, north korea's nuclear, north korean, to miniaturize nuclear

North Korea to US: if you attack us, we'll respond with nukes

There's no reason to think the US is actually planning to attack North Korea any time soon.

www.vox.com @ 2019-11-15T04:06:40Z Preserved by ARCHIVE.ORG

Other Versions || Current version

Trump promises North Korea 'fire and fury' over nuke threat

President Donald Trump issued an extraordinary ultimatum to North Korea on Tuesday warning Pyongyang not to make any more threats against the United States or they will "face fire and fury like the

www.cnn.com @ 2017-11-16T01:50:47Z Preserved by ARCHIVE.ORG

Other Versions || Current version

Trump threatens 'fire and fury' in response to North Korean threats

North Korea had threatened 'physical action' following new U.N. sanctions.

www.washingtonpost.com @ 2019-07-23T20:54:01Z Preserved by ARCHIVE.ORG

Other Versions || Current version

Trump: North Korea 'will be met with fire and fury like the world has never seen' if more threats emerge

President Trump has responded to reports of North Korea's nuclear threats, saying the regime "will be met with the fire and fury and frankly power, the likes of which the world has never seen befor

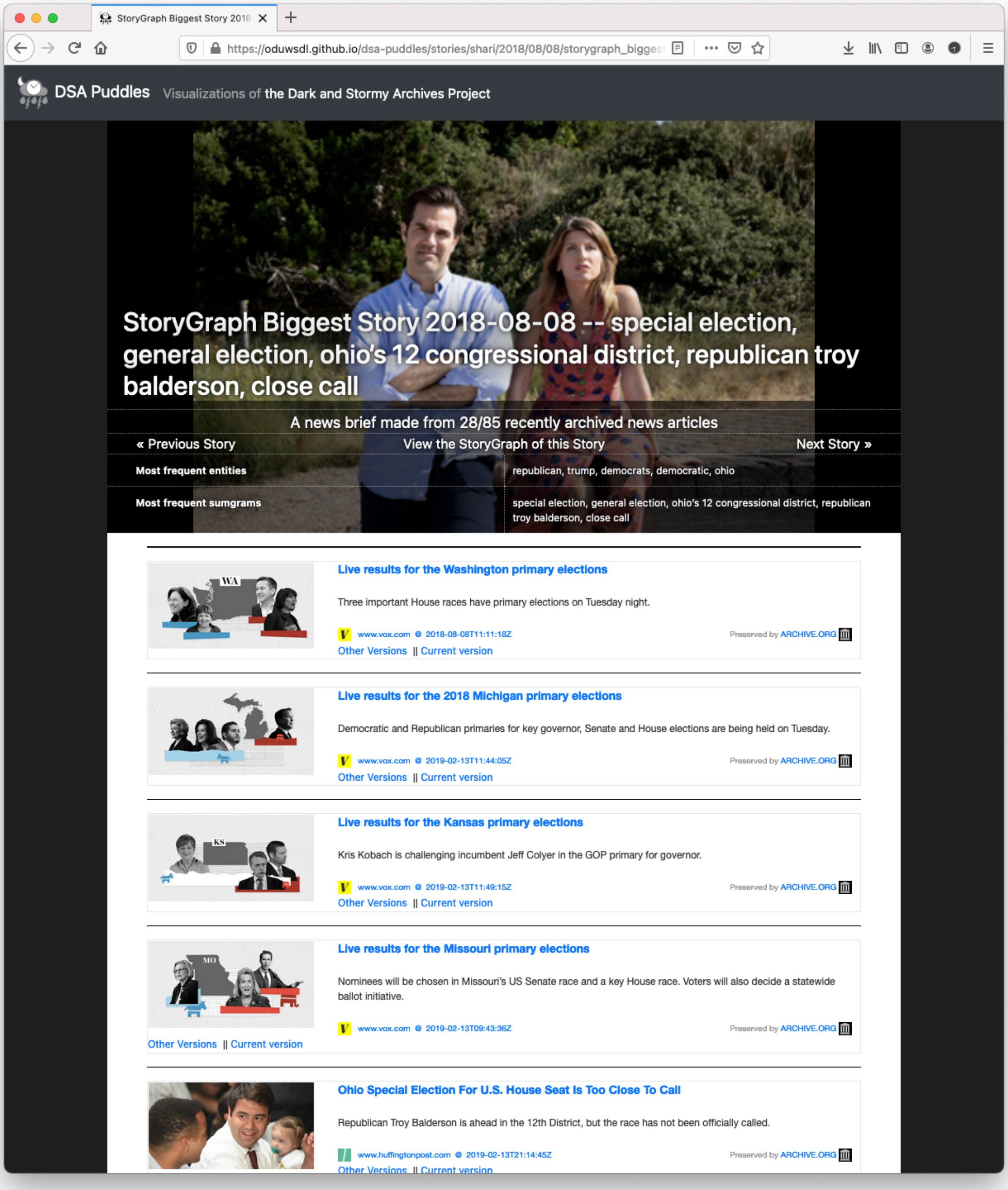
www.foxnews.com @ 2020-02-01T20:43:12Z Preserved by ARCHIVE.ORG

Other Versions || Current version

U.S., Japanese Analysts Agree: North Korea Has Developed Nuclear Missile Warheads

(a) 2017

Fig. 183. With the SHARI process, we can compare the news across the same day in different years


StoryGraph Biggest Story 2018
https://oduwsdl.github.io/dsa-puddles/stories/shari/2018/08/08/storygraph_bigges
DSA Puddles
Visualizations of the Dark and Stormy Archives Project

StoryGraph Biggest Story 2018-08-08 -- special election, general election, ohio's 12 congressional district, republican troy balderson, close call

A news brief made from 28/85 recently archived news articles

« Previous Story | View the StoryGraph of this Story | Next Story »

Most frequent entities	republican, trump, democrats, democratic, ohio
Most frequent sumgrams	special election, general election, ohio's 12 congressional district, republican troy balderson, close call

Live results for the Washington primary elections

Three important House races have primary elections on Tuesday night.

www.vox.com @ 2018-08-08T11:11:18Z Preserved by ARCHIVE.ORG

Other Versions || Current version

Live results for the 2018 Michigan primary elections

Democratic and Republican primaries for key governor, Senate and House elections are being held on Tuesday.

www.vox.com @ 2019-02-13T11:44:05Z Preserved by ARCHIVE.ORG

Other Versions || Current version

Live results for the Kansas primary elections

Kris Kobach is challenging incumbent Jeff Colyer in the GOP primary for governor.

www.vox.com @ 2019-02-13T11:49:15Z Preserved by ARCHIVE.ORG

Other Versions || Current version

Live results for the Missouri primary elections

Nominees will be chosen in Missouri's US Senate race and a key House race. Voters will also decide a statewide ballot initiative.

www.vox.com @ 2019-02-13T09:43:36Z Preserved by ARCHIVE.ORG

Other Versions || Current version

Ohio Special Election For U.S. House Seat Is Too Close To Call

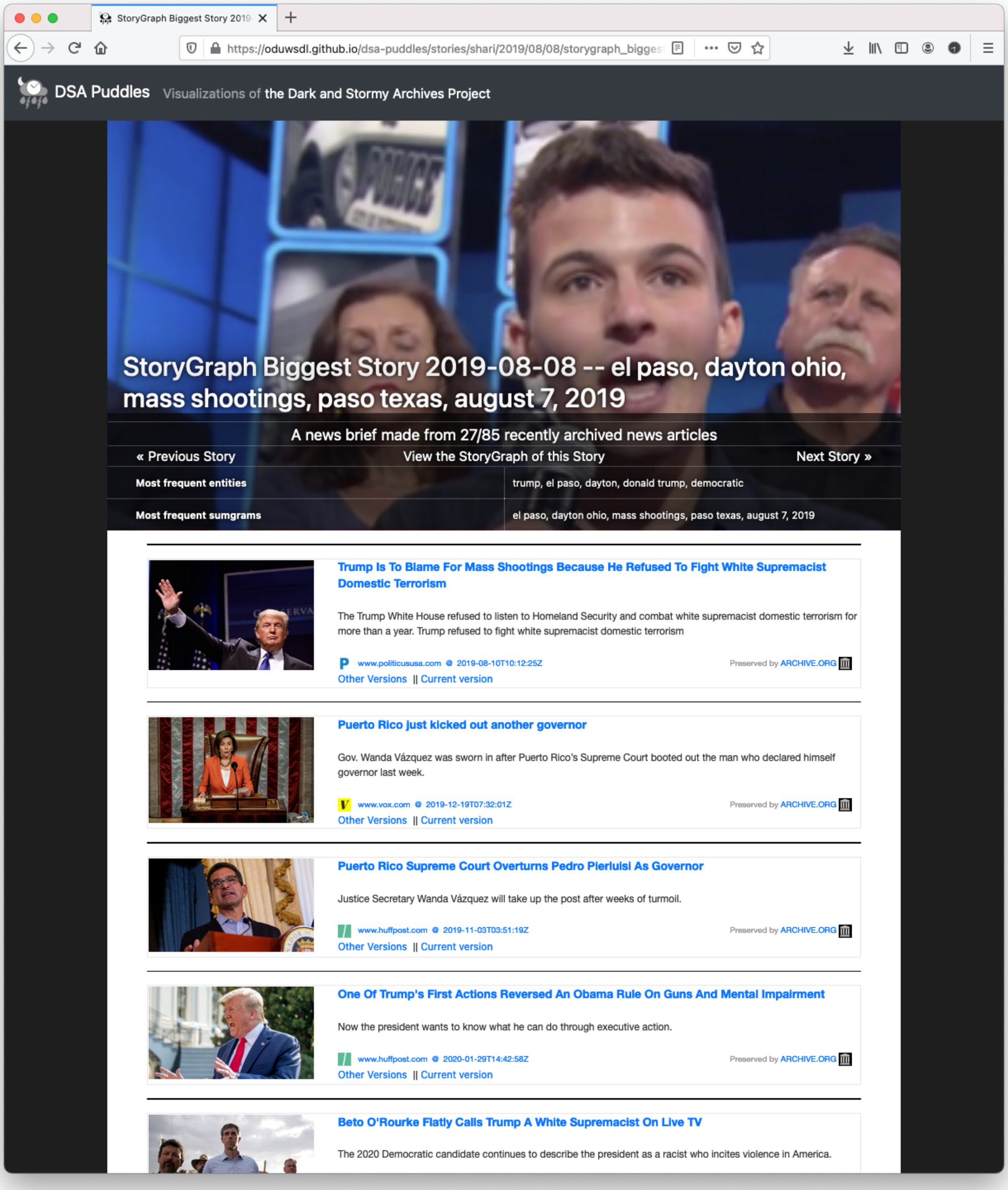
Republican Troy Balderson is ahead in the 12th District, but the race has not been officially called.

www.huffingtonpost.com @ 2019-02-13T21:14:45Z Preserved by ARCHIVE.ORG

Other Versions || Current version

(a) 2018

Fig. 184. (Continued) With the SHARI process, we can compare the news across the same day in different years


StoryGraph Biggest Story 2019
https://oduwsdl.github.io/dsa-puddles/stories/shari/2019/08/08/storygraph_bigges
DSA Puddles
Visualizations of the Dark and Stormy Archives Project

StoryGraph Biggest Story 2019-08-08 -- el paso, dayton ohio, mass shootings, paso texas, august 7, 2019

A news brief made from 27/85 recently archived news articles

« Previous Story | View the StoryGraph of this Story | Next Story »

Most frequent entities	trump, el paso, dayton, donald trump, democratic
Most frequent sumgrams	el paso, dayton ohio, mass shootings, paso texas, august 7, 2019

Trump Is To Blame For Mass Shootings Because He Refused To Fight White Supremacist Domestic Terrorism

The Trump White House refused to listen to Homeland Security and combat white supremacist domestic terrorism for more than a year. Trump refused to fight white supremacist domestic terrorism

www.politicususa.com @ 2019-08-10T10:12:25Z Preserved by ARCHIVE.ORG
Other Versions || Current version

Puerto Rico just kicked out another governor

Gov. Wanda Vázquez was sworn in after Puerto Rico's Supreme Court booted out the man who declared himself governor last week.

www.vox.com @ 2019-12-19T07:32:01Z Preserved by ARCHIVE.ORG
Other Versions || Current version

Puerto Rico Supreme Court Overturns Pedro Pierluisi As Governor

Justice Secretary Wanda Vázquez will take up the post after weeks of turmoil.

www.huffpost.com @ 2019-11-03T03:51:19Z Preserved by ARCHIVE.ORG
Other Versions || Current version

One Of Trump's First Actions Reversed An Obama Rule On Guns And Mental Impairment

Now the president wants to know what he can do through executive action.

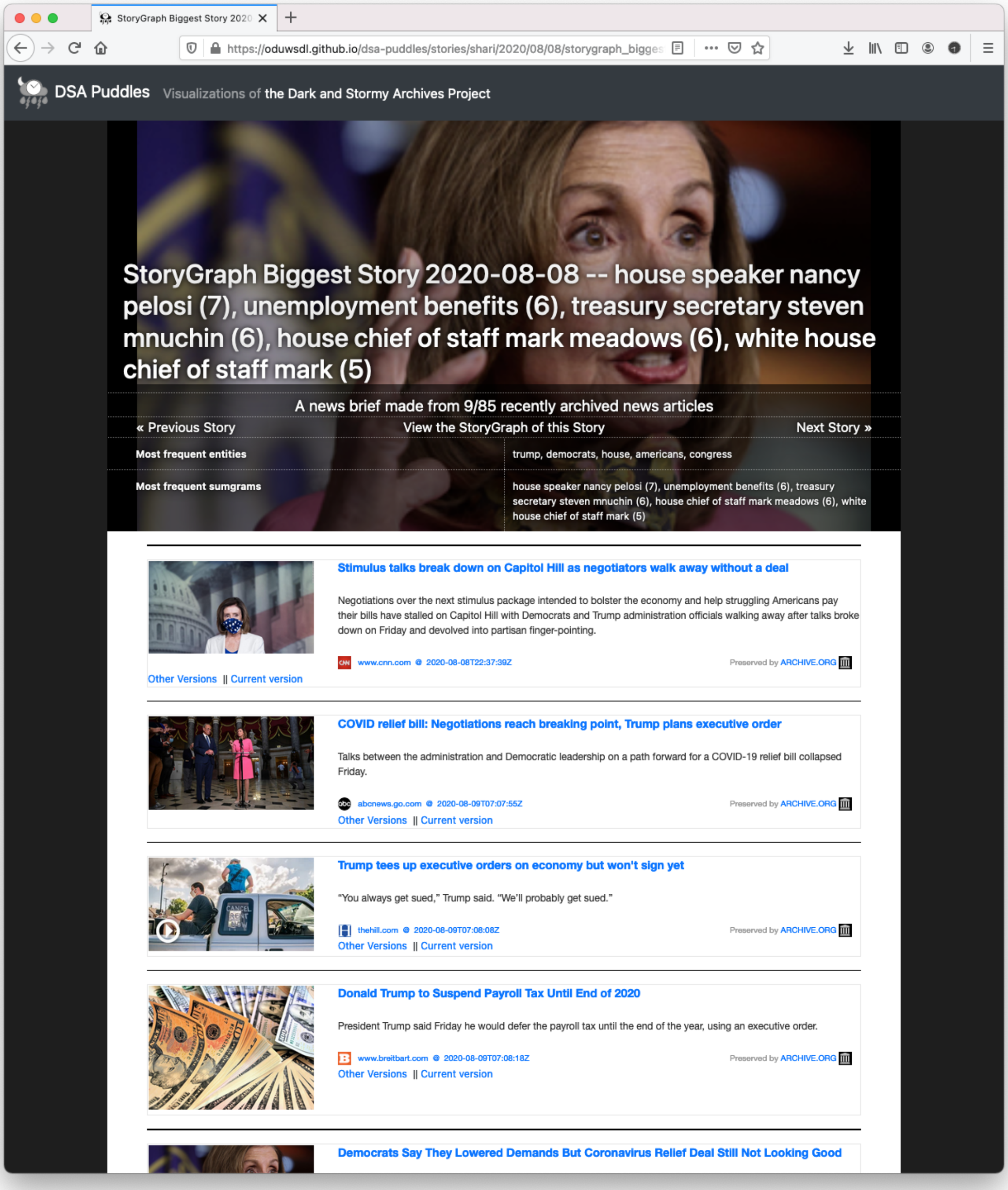
www.huffpost.com @ 2020-01-29T14:42:58Z Preserved by ARCHIVE.ORG
Other Versions || Current version

Beto O'Rourke Flatly Calls Trump A White Supremacist On Live TV

The 2020 Democratic candidate continues to describe the president as a racist who incites violence in America.

(a) 2019

Fig. 185. (Continued) With the SHARI process, we can compare the news across the same day in different years



StoryGraph Biggest Story 2020
https://oduwsdl.github.io/dsa-puddles/stories/shari/2020/08/08/storygraph_bigges
DSA Puddles
Visualizations of the Dark and Stormy Archives Project

StoryGraph Biggest Story 2020-08-08 -- house speaker nancy pelosi (7), unemployment benefits (6), treasury secretary steven mnuchin (6), house chief of staff mark meadows (6), white house chief of staff mark (5)

A news brief made from 9/85 recently archived news articles

« Previous Story | View the StoryGraph of this Story | Next Story »

Most frequent entities	trump, democrats, house, americans, congress
Most frequent sumgrams	house speaker nancy pelosi (7), unemployment benefits (6), treasury secretary steven mnuchin (6), house chief of staff mark meadows (6), white house chief of staff mark (5)

Stimulus talks break down on Capitol Hill as negotiators walk away without a deal

Negotiations over the next stimulus package intended to bolster the economy and help struggling Americans pay their bills have stalled on Capitol Hill with Democrats and Trump administration officials walking away after talks broke down on Friday and devolved into partisan finger-pointing.

www.cnn.com @ 2020-08-08T22:37:39Z Preserved by ARCHIVE.ORG

Other Versions || Current version

COVID relief bill: Negotiations reach breaking point, Trump plans executive order

Talks between the administration and Democratic leadership on a path forward for a COVID-19 relief bill collapsed Friday.

abcnews.go.com @ 2020-08-09T07:07:55Z Preserved by ARCHIVE.ORG

Other Versions || Current version

Trump tees up executive orders on economy but won't sign yet

"You always get sued," Trump said. "We'll probably get sued."

thehill.com @ 2020-08-09T07:08:08Z Preserved by ARCHIVE.ORG

Other Versions || Current version

Donald Trump to Suspend Payroll Tax Until End of 2020

President Trump said Friday he would defer the payroll tax until the end of the year, using an executive order.

www.breitbart.com @ 2020-08-09T07:08:18Z Preserved by ARCHIVE.ORG

Other Versions || Current version

Democrats Say They Lowered Demands But Coronavirus Relief Deal Still Not Looking Good

(a) 2020

Fig. 186. (Continued) With the SHARI process, we can compare the news across the same day in different years

how the storytelling processes need not be carried out solely with DSA tools. We closed with the SHARI process, where StoryGraph selects exemplars and Hypercane merely generates story metadata. The exemplars and story metadata are passed to Raintale as before, but this time we are not summarizing individual collections but summarizing the exemplars selected by StoryGraph. What other tools might satisfy these processes? How else might we apply storytelling with web archives? In the next chapter, we discuss some future directions.

CHAPTER 7

FUTURE WORK

A natural outcome of collecting things is to place some structure on them. We can manually explore a small number of items (e.g., 28) and do not need to apply much formal order to understand them. As the collection grows, it is not easy to comprehend the complexity of the whole, and we need a summary. Our storytelling processes discover the innate structure of a collection before visualizing it as a summary. We have selected exemplars, generated story metadata, generated document metadata, visualized the story, and distributed it. This model applies to many different storytelling ventures. The work to improve individual processes is ongoing. We have mentioned before that this method can be applied to other web archives and is by no means limited to Archive-It. This chapter discusses how we can improve this solution for web archives and potentially extend it beyond them.

7.1 MORE STORYTELLING WITH WEB ARCHIVES

Where does the archivist Elbert go from here? He can use our DSA algorithms to create stories for visitors like Natasha who are trying to to compare collections; he can create unique stories for visitors like Rustam and Olayinka who want to explore an aspect of a collection for their research. He can help out fellow archivists like Ling who wants to understand a collection that she inherited. There are still use cases that we have not addressed. We can improve on our processes for selecting exemplars. We can determine which visualizations work better for understanding.

7.1.1 IMPROVEMENTS FOR SELECTING EXEMPLARS

We still have work to do in selecting exemplars. As noted in Chapter 2, Latent Semantic Indexing assigns a single topic to each document, which is not realistic. Latent Dirichlet Allocation (LDA), however, assigns multiple topics to a document. We employed LDA for clustering in the DSA2 algorithm, but we could leverage it for finding off-topic mementos as well. Perhaps it would be better at determining which mementos are off-topic within a TimeMap. Those mementos that correlate poorly with their topics could be considered off-topic. We could also evaluate other intra-TimeMap measures, like Spamsum [187], which functions similarly to Simhash. The UK Web Archive has leveraged Spamsum to measure content drift [130] among mementos in a TimeMap.

Sometimes the page starts as off-topic. We need a better method than provided by our TimeMap measures and our algorithm from Chapter 4.2. Patel et al. [259] may offer a solution. Rather than focusing on identifying off-topic mementos, they consider the task to be a binary classification problem where they divide the collection into "in-scope" and not. They compare the results using a bag-of-words generated from the whole document against one from the the start and end of a document to see if they get better results with less text. They also apply a CNN classifier to see if it improves performance. They concluded that text from specific portions of documents is more effective than processing the whole document. They did not, however, process web pages, focusing solely on PDFs. We would need to extend their efforts to see if we can apply their solution to the cornucopia of different resource types (e.g., social media posts, news articles, blog posts) in most web archive collections.

We need to evaluate how well these methods work for collections that contain multiple languages. Most of our experiments have centered on English. Some portions of the DSA algorithms are language independent (e.g., clustering) but others may need to be modified to better support specific languages.

There are many more structural features than those that we have not yet explored. Consider the following:

mementos-per-seed — This can take the form of a percentage or a ratio. As a percentage it would just be (number of mementos / number of seeds). As a ratio it would need to be reduced (e.g., 8:4 becomes 2:1) so we can compare collections to each other. This simple metric may provide additional curation insight. If this number is close to 1:1, we likely have a collection with the growth curve shape of seed mementos growing continuously. A value where the numerator is less than the denominator indicates that many seeds were registered but not crawled.

seeds with metadata — This feature, though helpful to archivists, is likely not helpful in predicting the semantic type of a collection or which DSA algorithm would be the best fit for it. Too many collections lack seed metadata for this to be a significant differentiator.

scheme usage — This feature would make more sense in the web of the 1990s than it would today. Most seeds use *http* or *https* as their scheme. It could be interesting to archivists to see this as a percentage. If we consider *http* and *https* to be the same scheme, then it might be helpful to identify collections with seeds that use *ftp*. From our 2017 study, only *Sun Microsystem documents* (ID 7000) applied *ftp* as its most frequent scheme.

valid seeds — This feature is another that might make more sense to archivists. Sometimes an archivist creates a seed that is an invalid URI. Unfortunately, the seed remains on the Archive-It collection page. It is unlikely to differentiate one collection from another but may be helpful for

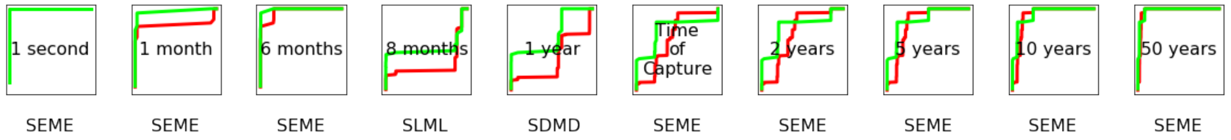


Fig. 187. With this mock collection, we see how collection growth curve shapes change as archivist activity grows dormant. (SEME = seeds early, mementos early; SLML = seeds late, mementos late; SDMD = seeds continuously, mementos continuously) See Chapter 4.1 for more information on growth curves and how to interpret them.

collection administration.

% of seeds in given page category – We are uncertain how to turn this into a specific metric, but it may be of interest to know how many seeds fit into a given web page category. For example, some *Self-Archiving* collections are a mix of social media and *.gov* or *.edu* domains combined with social media domains. Measuring the percentage of seeds linking to a social media domain may help differentiate collections that fit into this semantic category from those in other categories. The method employed by Padia et al. [257] requires updating a list of domains for each category. Abrahamson et al. [2] provide a more adaptable genre classification algorithm.

In Chapter 4.1, we normalized all growth curve features to the life of the collection. It may be interesting to study how growth curves change over time without this normalization (Figure 187). As time goes on, all unmaintained collections approach the growth curve shape of seeds early, seed mementos early. Can we determine if a collection is dormant based on its growth curve and help users recognize that it no longer being maintained?

We also divided different growth curves into categories based on the area under their curves. Maybe there are other ways to evaluate these curves. If we assume that archivists are inspired to add to collections when news events occur, perhaps correlating local minima or maxima to world events provided by Wikipedia could provide additional insight into these collections. Perhaps we could leverage this external data to improve the selection of exemplars or even augment stories with additional metadata.

We did not evaluate if a particular DSA algorithm works best for a particular semantic type of collection. DSA3 may produce better stories for *Self-Archiving* collections, and DSA2 may create better stories for *Time Bounded - Expected*. Perhaps our semantic categories are not a good fit for this problem. Someone instead might classify Archive-It collections based on which algorithms work best for a given collection.

How well do our exemplars cover the content of their collections? We could leverage methods like Kilgarriff's [179] word frequency method to compare corpora. This method considers each corpus a probability distribution and applies the χ^2-test to quantify the statistically significant difference between distributions. Alternatively, we could leverage the log-likelihood function to compare the top terms or entities between collections. Unfortunately, these methods rely on the two corpora being the same size and assume they are statistically independent, which will not be the case between the whole collection and the selected exemplars. To address this, we could apply a method developed by Lijffijt [198] that addresses these issues. These methods all require that we apply them to one term or entity at a time. We can also leverage the collection metrics created by Nwala et al. [245] and those from Chapter 4.1 to compare the exemplars (small collection) with their collection of origin. To better understand how to evaluate algorithms by applying the features from Chapter 4.1, we would need to select exemplars from more collections than the eight from the experiment detailed in Chapter 4.4.

What other types of stories exist? AlNoamany's dimensions of page and time were our starting dimensions, but others exist. *Fixed pages, sliding time* could include multiple specific seeds from the collection. A *fixed source, sliding time* would select exemplars from a given domain name. One could expand this to *fixed sources, sliding time* to cover specific media ecosystems. With the current polarization among US media, one could apply this new story type to tell a story from a particular political perspective. With exemplars restricted by query or topic modeling, one could tell a *fixed topic, sliding time* story. While *sliding page, sliding time* stories best serve Natasha and Ling, we want to emphasize that not just archivists but users should be able to tell stories in any dimension to address their needs. All of these dimensions have implications for how we select exemplars.

We have focused on selecting exemplars from HTML documents. Many web archive collections contain PDFs, which have their own document metadata. Work with PDFs has been the focus of summarizing scholarly corpora. What about other document formats? Not all mementos are documents. How do we select exemplar images, audio, or videos? Peng [263], Laganière et al. [190], Potapov et al. [271], Mei et al. [219], Yao et al. [336], and Kanehira et al. [173] all propose solutions to selecting exemplar clips from videos. Zhuang et al. [340], Gong and Liu [108], and Mundur et al. [229] provide solutions select exemplar striking images from videos.

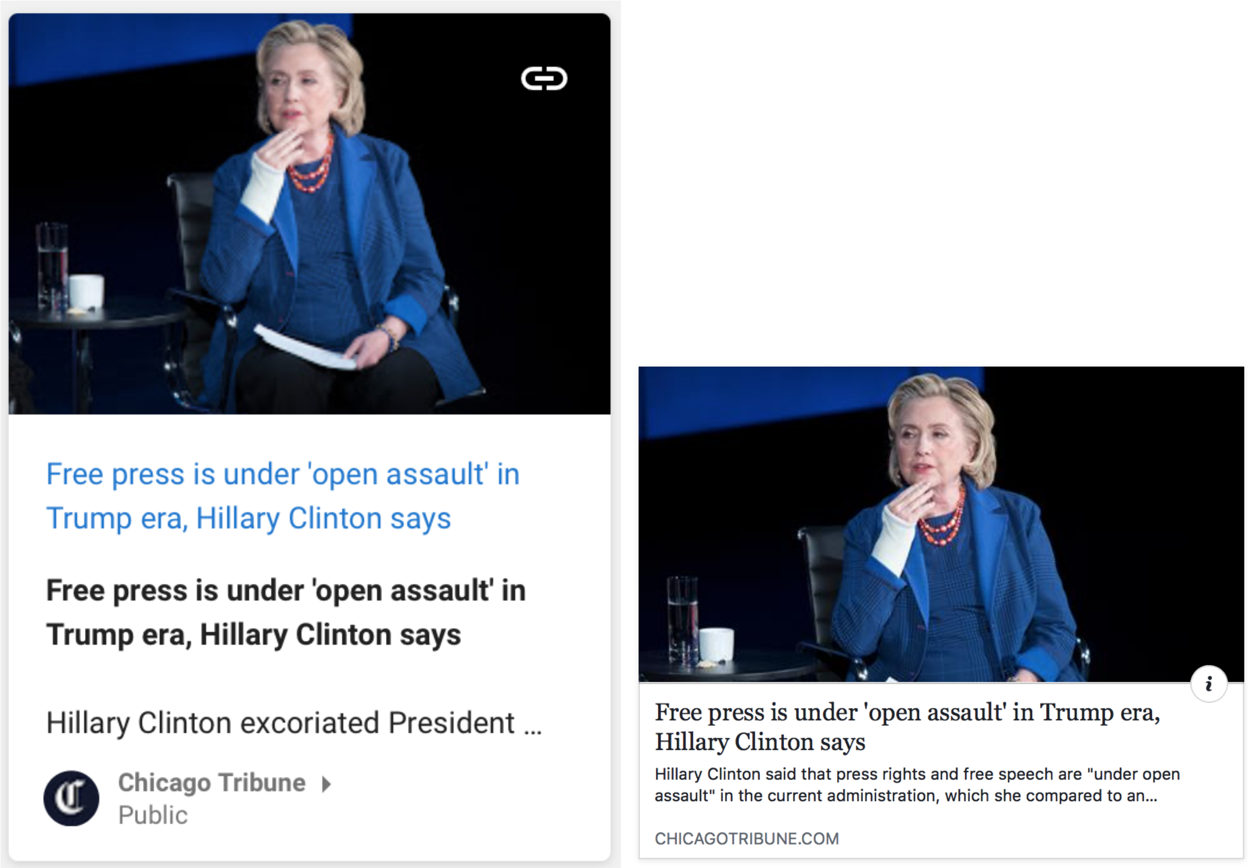
Our DSA algorithms from Chapter 4.3 assume that a web archive collection is structured like Archive-It: a collection of mementos, grouped by each seed's TimeMap, augmented with collection metadata. In Chapter 2.3 we mentioned collections from Confier, the Croatian Web Archive, Library of Congress, and Trove. In Chapter 6 we introduced examples that created stories from the web archive collections of Trove, maintained by the National Library of Australia

(NLA). Collections at NLA consist of mementos that archivists selected as being particularly meaningful. Algorithm steps like filtering duplicates and filtering off-topic mementos may not be needed because humans have already done this work. We created Hypercane (Chapter 4.5) with this in mind, realizing that one algorithm would not fit all collections, but perhaps pieces of algorithms could be shared between solutions. Croatian Web Archive (HAW) collections are formatted similarly to those at NLA. Perhaps some of these pieces could be shared by algorithms applied there as well. Our work can continue in that direction, finding algorithms that work best for these types of collections.

We can surface patterns from the documents in general web archives, like the Internet Archive. In Chapter 6 we discussed how the SHARI process helps those like Olayinka compare news between years. SHARI analyzes a portion of the live web and creates order from that sample before creating or discovering the mementos of that content in the Internet Archive or Archive.Today. What other techniques can we apply to general web archives? Kanhabua et al. [174] may offer a path for selecting exemplars. Their solution leverages existing live web search engines to find original resources. Kanhabu then discovers their mementos. From here, we can build a collection of mementos so we can then reduce it and summarize what is present in that part of a general web archive.

Our current DSA toolkit allows a user to specify a time period and original resources. With the Memento Protocol [312], Hypercane can locate the mementos corresponding to these resources. Hypercane also allows a user to crawl a web archive and find additional linked mementos. We can stop the crawl at a given depth and then score each memento by its path depth. By combining these concepts, a user could conceivably build a "newspaper" for any time period. Combine this with Hypercane's additional filtering abilities, and one could create fixed sources, sliding time stories across general web archives. We could also create fixed topic, sliding time stories. This work would allow historians to further extend their analysis of past news to address research questions, like *how does a society ingest outsiders?* We are currently working with the developers of StoryGraph to explore how this would work within the DSA toolkit.

Nwala et al. [245] analyzed social media micro-collections to find quality seeds to be feed into web archives. They also analyzed search engine result pages (SERPs) [246] for quality seeds. In both cases they were trying to automate the creation of a web archive collection. We could apply this to also surface collections from existing general web archives and web archive collections. Klein et al. [181] used data outside of web archives to generate event collections by analyzing changes in Wikipedia edit frequency to determine an event datetime. Klein et al. then scraped the Wikipedia references from the page edit at that datetime and crawled web archives with the



Free press is under 'open assault' in Trump era, Hillary Clinton says

Free press is under 'open assault' in Trump era, Hillary Clinton says

Hillary Clinton excoriated President ...

Chicago Tribune
Public

Free press is under 'open assault' in Trump era, Hillary Clinton says

Hillary Clinton said that press rights and free speech are "under open assault" in the current administration, which she compared to an...

CHICAGOTRIBUNE.COM

(a) Social card in portrait format from Google+. (Screenshot taken in 2018)

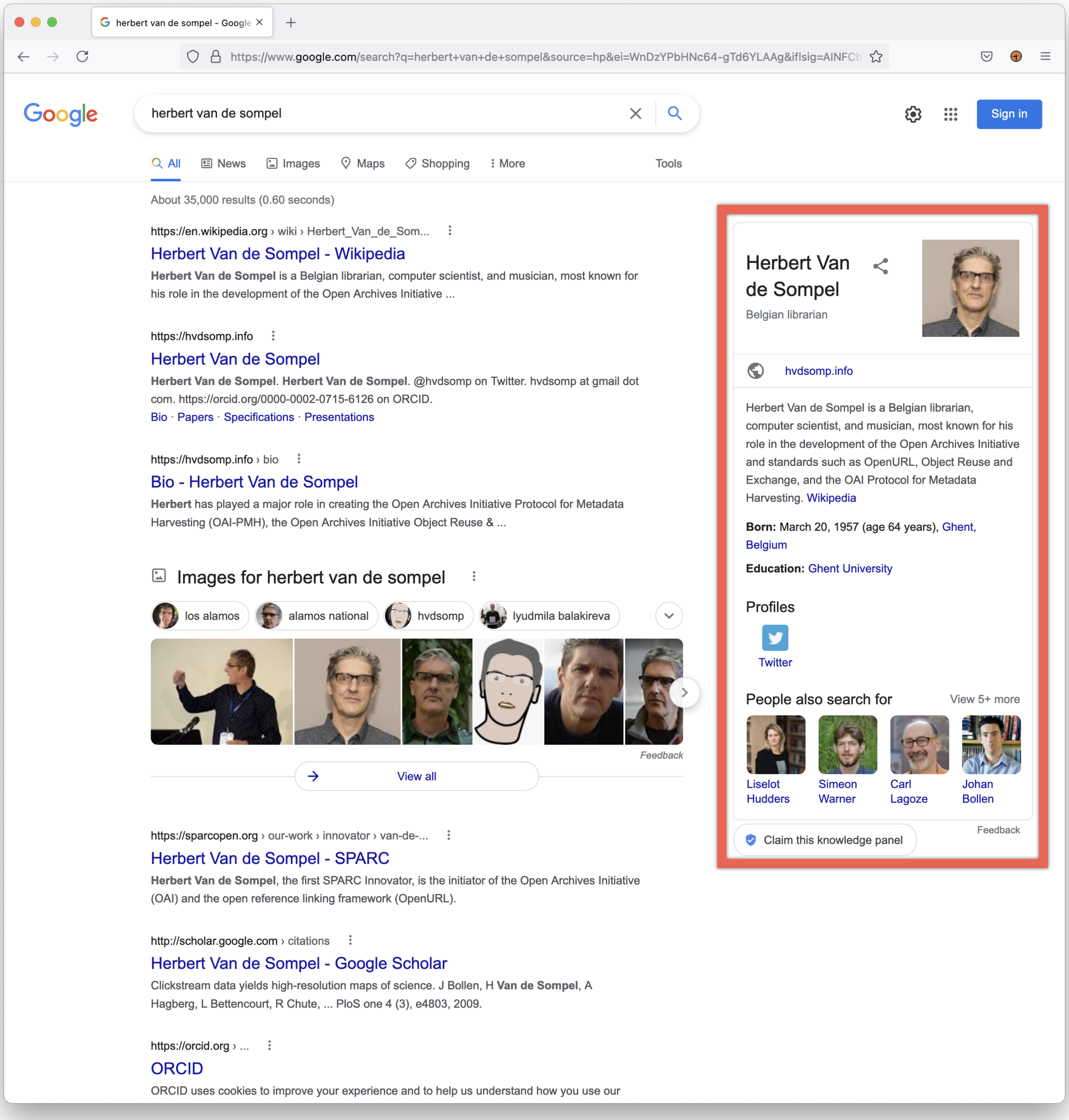
(b) Social card in landscape format from Facebook. (Screenshot taken in 2018)

Fig. 188. Do users learn more easily from social cards that are presented in portrait or landscape?

Memento Protocol to surface information about that event. Their methods could be combined with Kanehira et al. [173] to select exemplars from web archive collections and to generate story metadata from external sources.

7.1.2 EVALUATING VISUALIZATIONS FOR UNDERSTANDING

We could do more work analyzing the effectiveness of surrogates for collection understanding. We conducted a user study to determine which surrogate works best for understanding, but how do the specific fields of document metadata affect collection understanding? Do users understand best with just titles? Does the favicon or domain name in the social card influence how a user views the story's content and hence the underlying collection? What will eye tracking tell us about how users scan across different surrogate types? What parts of the social card do users spend more


herbert van de sompel - Google
https://www.google.com/search?q=herbert+van+de+sompel&source=hp&ei=WnDzYPbHNc64-gTd6YLAAg&iflsig=AINFCb
Google
herbert van de sompel
Sign in
All
News
Images
Maps
Shopping
More
Tools
los alamos
alamos national
hvdsomp
lyudmila balakireva
Feedback
View all
Feedback
Claim this knowledge panel

About 35,000 results (0.60 seconds)

https://en.wikipedia.org › wiki › Herbert_Van_de_Som...

Herbert Van de Sompel - Wikipedia

Herbert Van de Sompel is a Belgian librarian, computer scientist, and musician, most known for his role in the development of the Open Archives Initiative ...

https://hvdsomp.info

Herbert Van de Sompel

Herbert Van de Sompel. **Herbert Van de Sompel**. @hvdsomp on Twitter. hvdsomp at gmail dot com. https://orcid.org/0000-0002-0715-6126 on ORCID.

Bio · Papers · Specifications · Presentations

https://hvdsomp.info › bio

Bio - Herbert Van de Sompel

Herbert has played a major role in creating the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), the Open Archives Initiative Object Reuse & ...

Images for herbert van de sompel

https://sparcopen.org › our-work › innovator › van-de-...

Herbert Van de Sompel - SPARC

Herbert Van de Sompel, the first SPARC Innovator, is the initiator of the Open Archives Initiative (OAI) and the open reference linking framework (OpenURL).

http://scholar.google.com › citations

Herbert Van de Sompel - Google Scholar

Clickstream data yields high-resolution maps of science. J Bollen, H **Van de Sompel**, A Hagberg, L Bettencourt, R Chute, ... PloS one 4 (3), e4803, 2009.

https://orcid.org › ...

ORCID

ORCID uses cookies to improve your experience and to help us understand how you use our

Herbert Van de Sompel

Belgian librarian

hvdsomp.info

Herbert Van de Sompel is a Belgian librarian, computer scientist, and musician, most known for his role in the development of the Open Archives Initiative and standards such as OpenURL, Object Reuse and Exchange, and the OAI Protocol for Metadata Harvesting. Wikipedia

Born: March 20, 1957 (age 64 years), Ghent, Belgium

Education: Ghent University

Profiles

Twitter

People also search for

View 5+ more

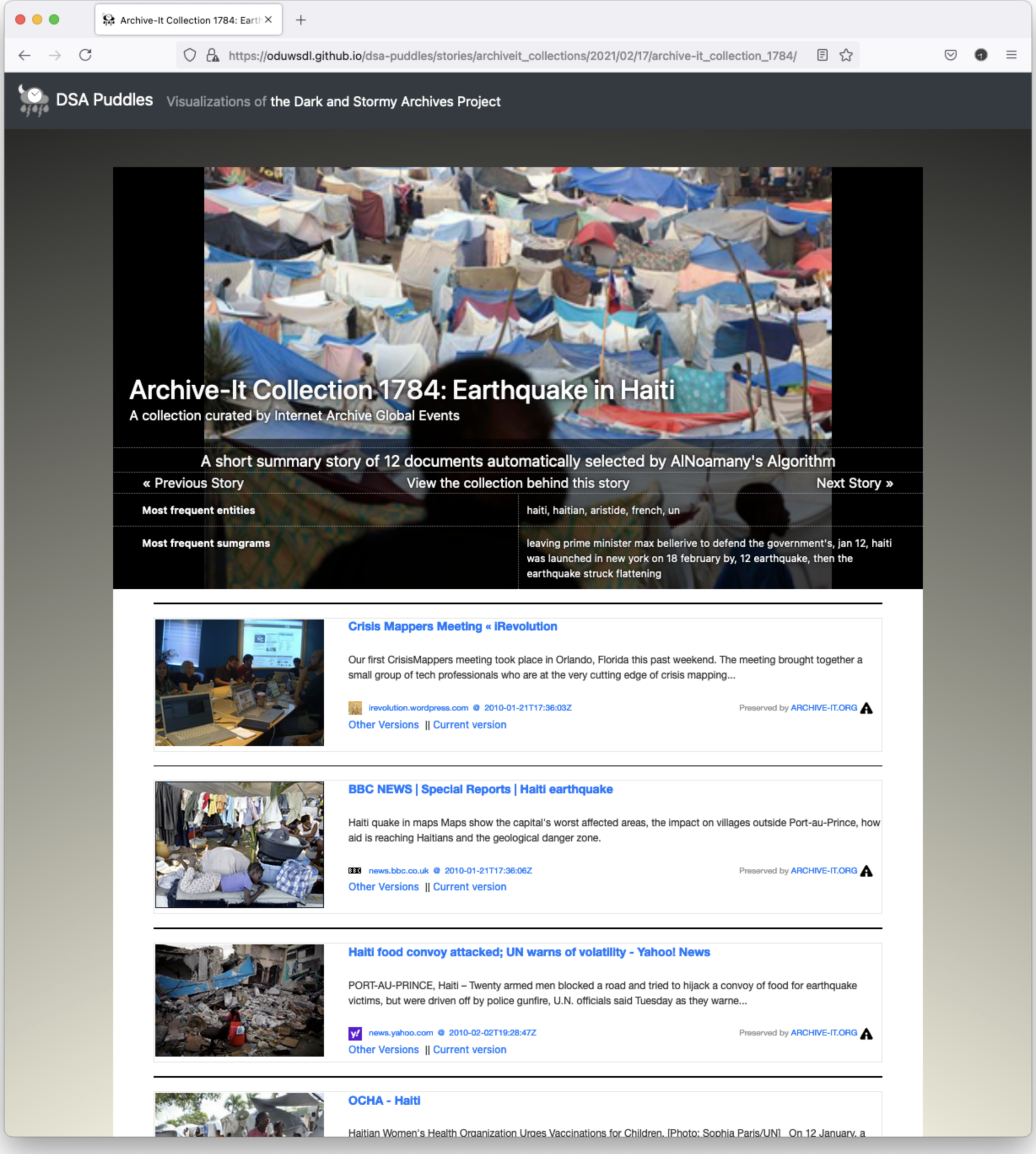
Liselot Hudders

Simeon Warner

Carl Lagoze

Johan Bollen

Fig. 189. Google provides entity cards (outlined in red), which they call *knowledge panels*, for certain named entities.


Archive-It Collection 1784: Eart
https://oduwsdl.github.io/dsa-puddles/stories/archiveit_collections/2021/02/17/archive-it_collection_1784/
DSA Puddles
Visualizations of the Dark and Stormy Archives Project

Archive-It Collection 1784: Earthquake in Haiti

A collection curated by Internet Archive Global Events

A short summary story of 12 documents automatically selected by AlNoamany's Algorithm

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

Most frequent entities	haiti, haitian, aristide, french, un
Most frequent sumgrams	leaving prime minister max bellerive to defend the government's, jan 12, haiti was launched in new york on 18 february by, 12 earthquake, then the earthquake struck flattening

Crisis Mappers Meeting « iRevolution

Our first CrisisMappers meeting took place in Orlando, Florida this past weekend. The meeting brought together a small group of tech professionals who are at the very cutting edge of crisis mapping...

irevolution.wordpress.com @ 2010-01-21T17:36:03Z   Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

BBC NEWS | Special Reports | Haiti earthquake

Haiti quake in maps Maps show the capital's worst affected areas, the impact on villages outside Port-au-Prince, how aid is reaching Haitians and the geological danger zone.

BBC news.bbc.co.uk @ 2010-01-21T17:36:06Z   Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

Haiti food convoy attacked; UN warns of volatility - Yahoo! News

PORT-AU-PRINCE, Haiti – Twenty armed men blocked a road and tried to hijack a convoy of food for earthquake victims, but were driven off by police gunfire, U.N. officials said Tuesday as they warne...

y! news.yahoo.com @ 2010-02-02T19:28:47Z   Preserved by ARCHIVE-IT.ORG

Other Versions || Current version

OCHA - Haiti

Haitian Women's Health Organization Urges Vaccinations for Children. [Photo: Sophia Paris/UN] On 12 January, a

Fig. 190. Hypercane automatically chooses the best image for a story, like this image of people living in tents from the Archive-It collection *Earthquake in Haiti*.

time examining? Even though users may perform well with one set of document metadata, they may prefer a different set. We could contrast users' surrogate preferences with their performance.

What presentation of surrogates works best? We have assumed that the Storify presentation of cards listed vertically works best because it is the search engines and social media paradigm. What about cards side-by-side? As shown in Figure 188, are portrait cards more effective than landscape ones? Also, does the screen size and orientation make a difference? For example, are some surrogates better for smartphones compared to desktop computers?

How well do surrogates match the content of their document? Existing research evaluates automatic summaries against human-generated summaries [199]. What if we have no human-generated summaries? How do we evaluate our surrogates? Grusky et al. [110] introduce concepts like density and coverage to describe how well an extractive summary summarizes a document. How well do these metrics correspond to human judgments? How well do they correspond with human understanding?

In Chapter 5.3, we evaluated different machine learning techniques to choose the same striking image as an article's author and we will replace MementoEmbed's existing image scoring function [156] with the classifier model from this work. How does a striking image in a social card help the reader's understanding? Would the reader have preferred a different image? Do multiple striking images improve the reader's performance? Should a system generate striking images based on a reader's past history or behavior so that the images are more meaningful to them?

Other types of cards exist, and we need to evaluate them for understanding. Bota et al. [39] explores the effectiveness of *entity cards* (Figure 189) on search behavior. Instead of analyzing a collection to produce for *exemplar mementos*, we should produce *exemplar entities*. Would a mix of entity cards with social cards work better for collection understanding? Perhaps entity types of cards work better for *Self-archiving* and *Subject-based* collections but not *Time Bounded - Spontaneous* collections? Designing a user study to determine this could be helpful. A fascinating study could be analyzing other types of exemplars (e.g., seeds, sources) and the best surrogates to represent them.

We select the overall story striking image, as seen in Figure 190, using the same method applied to individual striking images. Does this method work well for overall stories? We can evaluate this by applying the striking images chosen from the stories at Wakelet[1] as ground truth for a supervised learning study. For that matter, does this increased information scent truly help for understanding, or does it just make the experience more pleasant for users? As seen in Figure 190, we supply named entities and most frequent sumgrams [238] to enhance a user's understanding of

[1] https://wakelet.com/

the collection. Is this needed? Are the cards alone sufficient?

What other visualizations can we add to our stories? How do we visualize events that cover substantial periods of time, like the entire COVID-19 pandemic, Hurricane Katrina, or the Black Lives Matter movement? Perhaps we could automatically generate a timeline of incidents where each incident links to a specific card addressing that portion of the overall event. If a collection focuses on one of these events, but only supplies specific sources, then a visualization of the event will differ depending on the web archive collection from which they come.

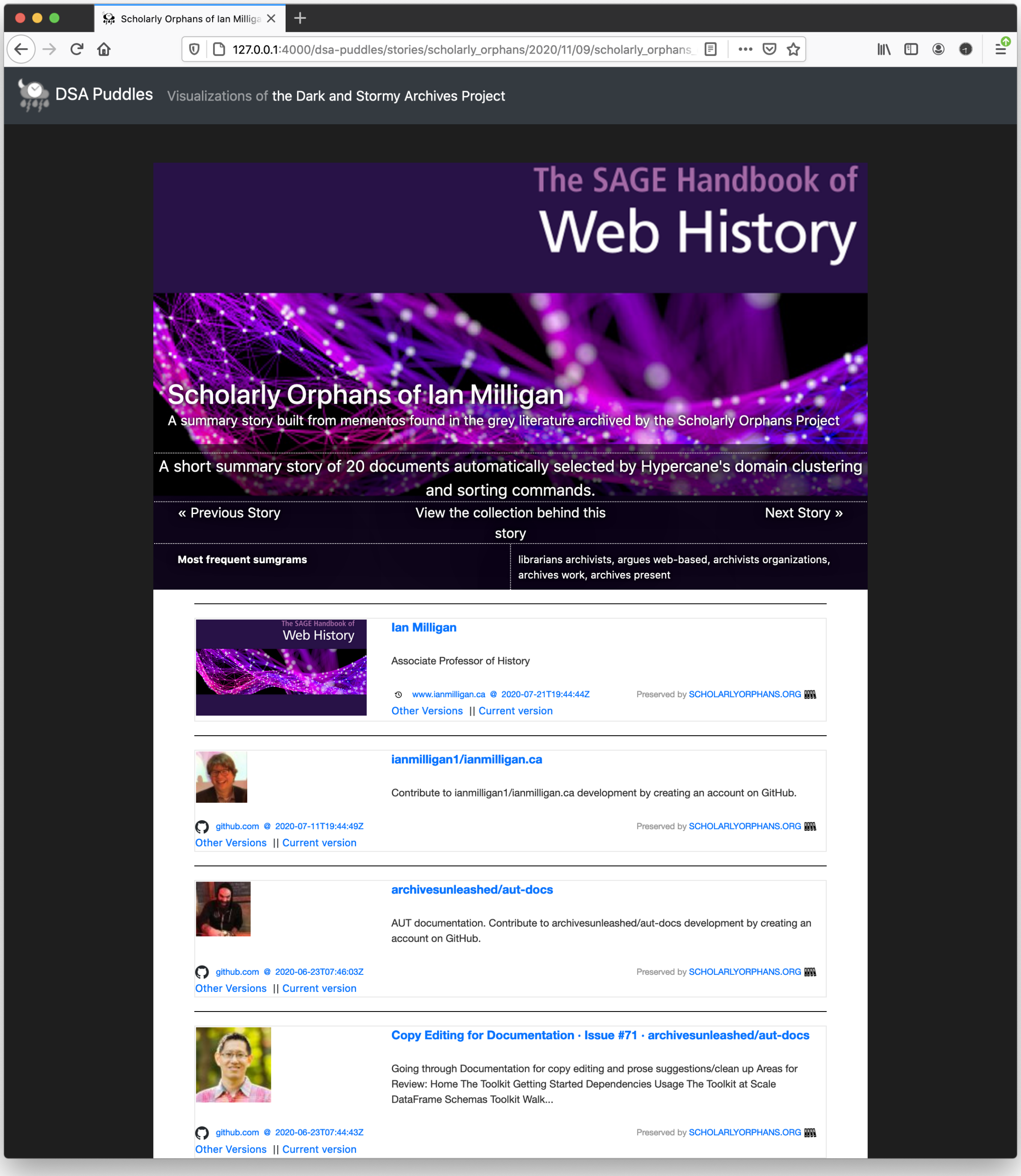
Finally, can we invert the concept? Maybe storytelling is a better interface for *creating* web archive collections. We could conduct a user study to evaluate different interfaces for creating collections. Seeing as social cards are an interface that users are already familiar with, they might be easier to use than the existing Archive-It interface. We would, however, need to address multiple issues with this paradigm, potentially including feedback from ethnographic studies [249] on how archivists construct web archive collections. For example, we would have to determine what metadata the card would display, how the card helps users configure crawling times, and how cards can help with batch processing.

7.2 BEYOND WEB ARCHIVES

We intentionally modeled the problem of storytelling with web archives generically. Our model for storytelling is about summarizing a collection of *things*. That summary takes the form of visualizations that summarize individual exemplars as surrogates and augments the resulting story with story metadata. However, this model need not only apply to mementos or even traditional documents.

Figure 191 demonstrates the concept of summarizing a scholar's grey literature. Here we executed Hypercane, Raintale, and MementoEmbed against a web archive collection from the Scholarly Orphans project that contains mementos from Ian Milligan's work outside of scholarly publications. We selected exemplars to summarize a scholar's output during a period of time, generated story and document metadata, and visualized and distributed it for others to see. It includes content from sites like GitHub, Milligan's home page, and WordPress. Stories like these are not only useful for archivists trying to feature the content of their collections. They can be used to highlight a scholar's accomplishments for the year, both to the public and institutional committees. The source of this story could be live web documents, but a web archive would further preserve the evidence of Milligan's contributions.

We can extend this idea further. Figure 192 shows how Google Scholar attempts to do something similar for a scholar's traditional publications. There is ordering by citation, there is story



Scholarly Orphans of Ian Milliga
127.0.0.1:4000/dsa-puddles/stories/scholarly_orphans/2020/11/09/scholarly_orphans_
DSA Puddles
Visualizations of the Dark and Stormy Archives Project

The SAGE Handbook of
Web History

Scholarly Orphans of Ian Milligan

A summary story built from mementos found in the grey literature archived by the Scholarly Orphans Project

A short summary story of 20 documents automatically selected by Hypercane's domain clustering and sorting commands.

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

Most frequent sumgrams	librarians archivists, argues web-based, archivists organizations, archives work, archives present

Ian Milligan

Associate Professor of History

www.ianmilligan.ca @ 2020-07-21T19:44:44Z Preserved by SCHOLARLYORPHANS.ORG

Other Versions || Current version

ianmilligan1/ianmilligan.ca

Contribute to ianmilligan1/ianmilligan.ca development by creating an account on GitHub.

github.com @ 2020-07-11T19:44:49Z Preserved by SCHOLARLYORPHANS.ORG

Other Versions || Current version

archivesunleashed/aut-docs

AUT documentation. Contribute to archivesunleashed/aut-docs development by creating an account on GitHub.

github.com @ 2020-06-23T07:46:03Z Preserved by SCHOLARLYORPHANS.ORG

Other Versions || Current version

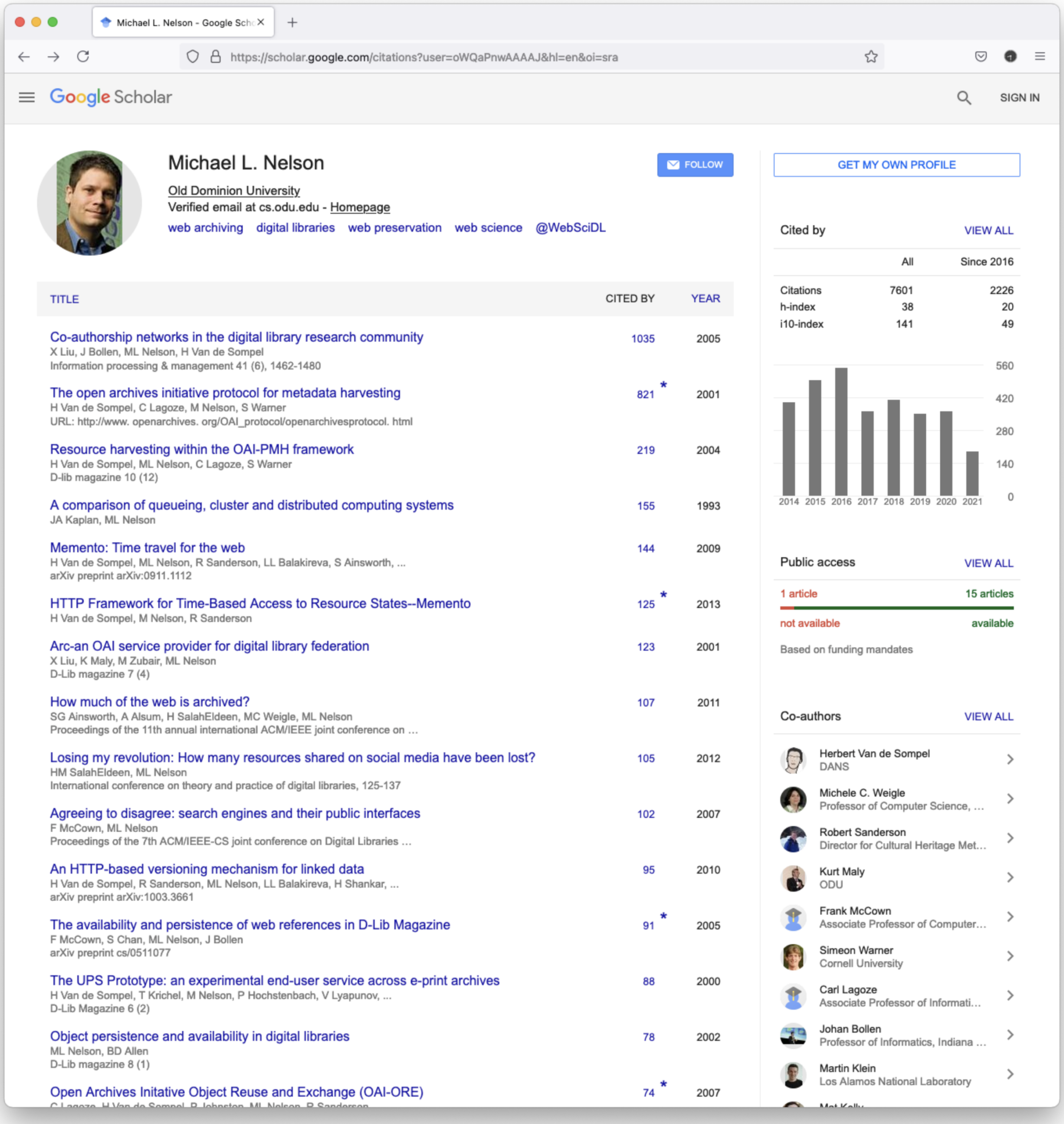
Copy Editing for Documentation · Issue #71 · archivesunleashed/aut-docs

Going through Documentation for copy editing and prose suggestions/clean up Areas for Review: Home The Toolkit Getting Started Dependencies Usage The Toolkit at Scale DataFrame Schemas Toolkit Walk...

github.com @ 2020-06-23T07:44:43Z Preserved by SCHOLARLYORPHANS.ORG

Other Versions || Current version

Fig. 191. We can apply our storytelling techniques to summarize other resources, like a scholar's grey literature.


Michael L. Nelson - Google Sch ×
+
https://scholar.google.com/citations?user=oWQaPnwAAAAJ&hl=en&oi=sra
Google Scholar
SIGN IN
FOLLOW
GET MY OWN PROFILE
VIEW ALL

Michael L. Nelson

Old Dominion University

Verified email at cs.odu.edu - Homepage

web archiving digital libraries web preservation web science @WebSciDL

TITLE	CITED BY	YEAR
Co-authorship networks in the digital library research community X Liu, J Bollen, ML Nelson, H Van de Sompel Information processing & management 41 (6), 1462-1480	1035	2005
The open archives initiative protocol for metadata harvesting H Van de Sompel, C Lagoze, M Nelson, S Warner URL: http://www. openarchives. org/OAI_protocol/openarchivesprotocol. html	821 *	2001
Resource harvesting within the OAI-PMH framework H Van de Sompel, ML Nelson, C Lagoze, S Warner D-lib magazine 10 (12)	219	2004
A comparison of queueing, cluster and distributed computing systems JA Kaplan, ML Nelson	155	1993
Memento: Time travel for the web H Van de Sompel, ML Nelson, R Sanderson, LL Balakireva, S Ainsworth, ... arXiv preprint arXiv:0911.1112	144	2009
HTTP Framework for Time-Based Access to Resource States--Memento H Van de Sompel, M Nelson, R Sanderson	125 *	2013
Arc-an OAI service provider for digital library federation X Liu, K Maly, M Zubair, ML Nelson D-Lib magazine 7 (4)	123	2001
How much of the web is archived? SG Ainsworth, A Alsum, H SalahEldeen, MC Weigle, ML Nelson Proceedings of the 11th annual international ACM/IEEE joint conference on ...	107	2011
Losing my revolution: How many resources shared on social media have been lost? HM SalahEldeen, ML Nelson International conference on theory and practice of digital libraries, 125-137	105	2012
Agreeing to disagree: search engines and their public interfaces F McCown, ML Nelson Proceedings of the 7th ACM/IEEE-CS joint conference on Digital Libraries ...	102	2007
An HTTP-based versioning mechanism for linked data H Van de Sompel, R Sanderson, ML Nelson, LL Balakireva, H Shankar, ... arXiv preprint arXiv:1003.3661	95	2010
The availability and persistence of web references in D-Lib Magazine F McCown, S Chan, ML Nelson, J Bollen arXiv preprint cs/0511077	91 *	2005
The UPS Prototype: an experimental end-user service across e-print archives H Van de Sompel, T Krichel, M Nelson, P Hochstenbach, V Lyapunov, ... D-Lib Magazine 6 (2)	88	2000
Object persistence and availability in digital libraries ML Nelson, BD Allen D-Lib magazine 8 (1)	78	2002
Open Archives Initative Object Reuse and Exchange (OAI-ORE) C Lagoze, H Van de Sompel, P Johnston, ML Nelson, R Sanderson	74 *	2007

Cited by

	All	Since 2016
Citations	7601	2226
h-index	38	20
i10-index	141	49

Public access VIEW ALL

1 article 15 articles

not available available

Based on funding mandates

Co-authors VIEW ALL

Herbert Van de Sompel	DANS
Michele C. Weigle	Professor of Computer Science, ...
Robert Sanderson	Director for Cultural Heritage Met...
Kurt Maly	ODU
Frank McCown	Associate Professor of Computer...
Simeon Warner	Cornell University
Carl Lagoze	Associate Professor of Informati...
Johan Bollen	Professor of Informatics, Indiana ...
Martin Klein	Los Alamos National Laboratory
Mat Kelly	

Fig. 192. Google Scholar lists a scholar’s publications.

and document metadata, and the whole was visualized and distributed. The key difference is that Figure 192 shows a complete list. No exemplars were selected. This visualization is a view into the entire collection. A reader can sort by title, year, and citation, but they cannot search within the work on this one scholar without searching the rest of Google Scholar. Instead, our storytelling processes could help a user promote the most impactful exemplars of their work. They could also apply our processes to convey the work most germane to a specific project so they could share it with a grant agency.

Ellen van Aken [311] noted that company employees want their corporate intranet search engines to work more like general web search engines (e.g., Google). Mukherjee and Mao [228] analyzed the needs of enterprise search users and determined how organizations need to manage intranet search engines more effectively. They discuss how corporate intranets often contain duplicate documents, which we can solve with our concept of filtering near-duplicates. Because most corporations do not hire a full-time librarian to organize their collection, their documents often have poor metadata, affecting their retrievability. Our concept of generating metadata can help solve this part of the problem. They suggest organizing documents visually into navigable structures (clustering) that will assist users in finding relevant information. Dumais et al. [74] performed a user study and established that these navigable structures help users find information better than a familiar search interface.

If we step back from our rigid storytelling concept, we see that some of our work may help solve this problem. Imagine a user exploring their corporate intranet through a series of stories clustered together via topics. A user starts their session with a query, and the system returns them a set of surrogates representing individual topics. To find their document, they can explore the documents within each topic or sub-topics or go back to higher-level topics. Pirolli and Card introduced this generic concept as *scatter/gather* [267]. It has been evaluated with positive results in recent user studies by Weimao et al. [176], Gong et al. [107], and Zhang et al. [339]. Our goal with storytelling is not to replace enterprise search engines. However, the results from our storytelling concepts can help users better understand topics on their way to finding the document they need, similar to a traditional library reference interview. If the corporate intranet search engine is already failing them, perhaps a journey through this scatter/gather interface would be more effective, especially considering DSA algorithms return documents with low retrievability (Chapter 4.4).

Of course, there are other ways our work can help with the corporate intranet. It could summarize an employee's output on the corporate intranet, possibly helping users fill out end-of-year assessments. Our stories would be an improvement over the flat list of "user contributions" offered

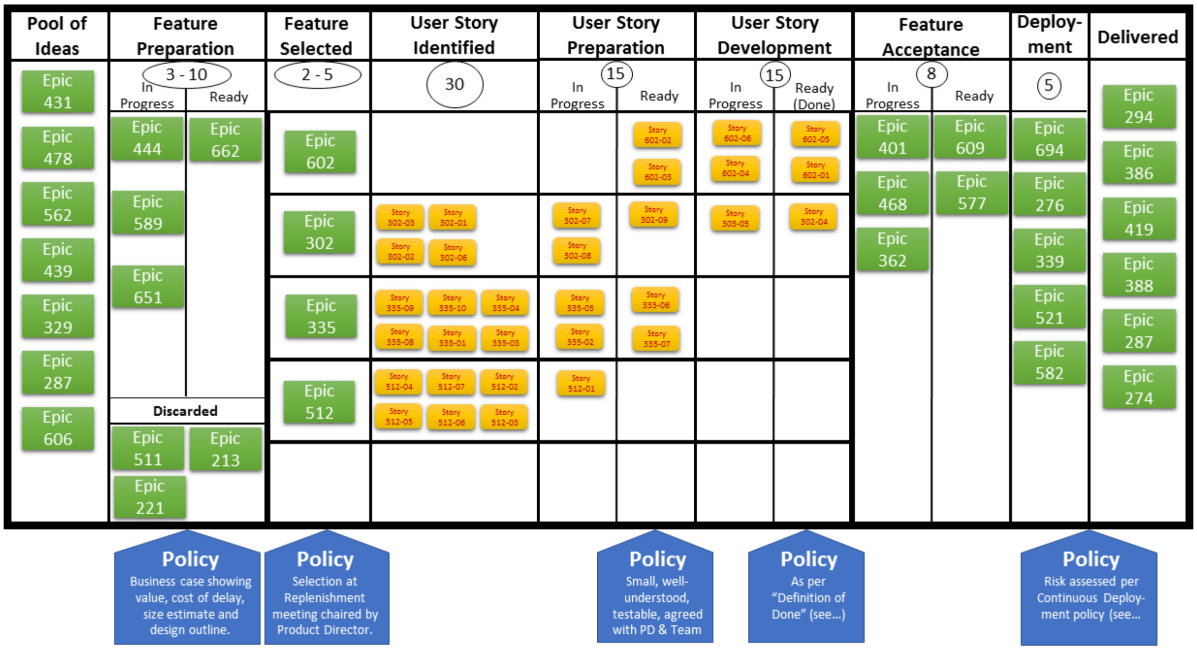


Fig. 193. Our storytelling model could summarize portions of a project's Kanban board. Example courtesy of Wikimedia Commons [321].

by corporate wikis. Consider extending this idea to projects, where a story summarizes the output of a particular team or a particular project phase, similar to a Kanban board (Figure 193).

7.3 CHAPTER SUMMARY

We covered potential future directions of our storytelling model. We first addressed where to take the work we presented in Chapters 4, 5, and 6. For selecting exemplars, we first need to detect off-topic mementos. We discussed several options for performing this off-topic detection by applying LDA, Spamsum, and the intra-TimeMap work started by Patel et al.

We also discussed additional structural features for describing and comparing collections, like mementos-per-seed and % of seeds in a given page category. We mentioned new directions for evaluating growth curves and how we might apply their shapes to discover events or improve our stories with information from external sources. We discussed additional algorithms for selecting exemplars and how to compare results. If we want to know if a set of exemplars is a good summary, we can compare probability distributions of the words in collections with the set of exemplars. We can also compare the structural features we have discussed and those developed by Nwala. We also mentioned that our focus has been on HTML documents and that mementos of other types of content, like video, exists in web archive collections. We need to determine if our social cards work for these or need a new surrogate type.

Other types of stories exist. *Fixed pages, sliding time* extends the singular fixed page, sliding time story to include multiple original resources. A *fixed source, sliding time* story would do the same for a single domain, and we could group domains in a *fixed sources, sliding time* story. We can demonstrate how a given publisher or political point of view covered an event with fixed sources. With topic modeling, we can create *fixed topic, sliding time* stories, providing stories that only cover the results of a specific query or cluster in topic modeling. These are just some of the additional dimensions with which one could experiment and grow our storytelling concept.

Our existing DSA algorithms may not be the best fit for selecting exemplars from web archives outside of Archive-It, like those from the National Library of Australia or the Croatian Web Archive. We noted that we could extend our techniques to explore general web archives and explore a topic by first building a collection from general web archives and then generating a summary. An example output of this process might be a "newspaper" for a given time period.

More work is needed to understand the effectiveness of surrogates for collection understanding. We construct surrogates from metadata fields. Perhaps only some of these fields are necessary for understanding. Someone could evaluate the ordering of surrogates and the arrangement of surrogates to see if it improves the user's performance and desired experience. We noted that perhaps different types of cards, like entity cards, might be a good augmentation for some stories. There is more work to be done in selecting striking images for stories and visualizing events that cover substantial periods of time.

The future of this model is not just limited to web archive collections. We can extend it to other areas, such as summarizing scholarly communications, improving corporate intranet discovery interfaces, or providing summaries of an individual employee's contributions to the company.

In our final chapter, we discuss how we got here. We revisit our research questions, and we provide an overview of our work.

CHAPTER 8

CONCLUSIONS

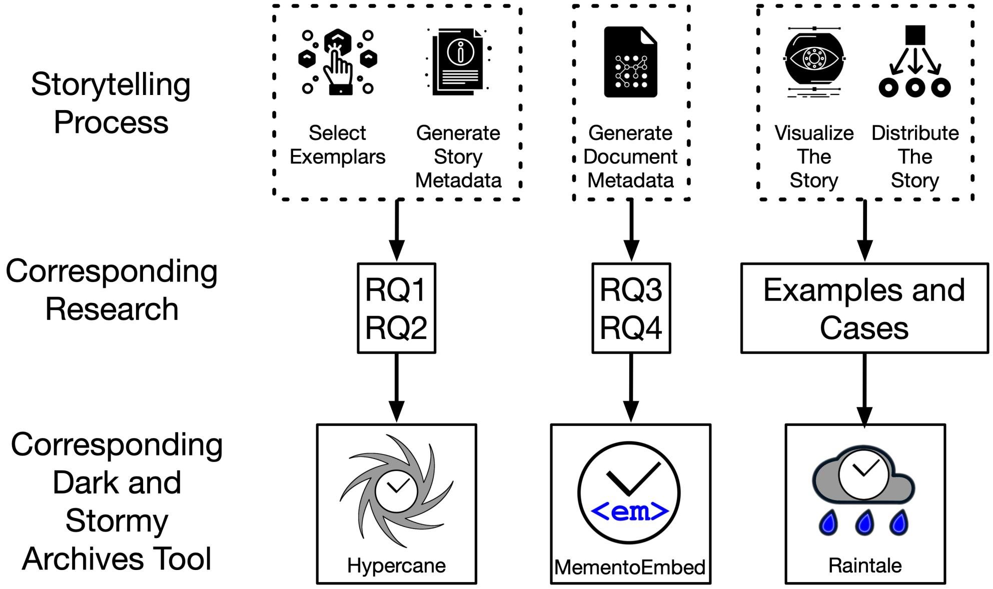


Fig. 194. The five processes for storytelling with a corpus mapped to the research in this dissertation.

This dissertation explores the application of social media storytelling to web archives for collection understanding. In Chapter 1, we introduced the needs of Natasha, Rustam, Olayinka, Elbert, and Ling, summarized in Table 1 on page 14. Natasha wanted to compare collections. Rustam wanted to explore how a single page changed its reporting as its authors learned new information about the Boston Marathon Bombing. Olayinka wanted to compare different news sources at a given date and time. Elbert wants to promote collections and help people like Natasha compare them. Ling inherited a collection and wants to understand it to make decisions about how to maintain it in the future.

Addressing their use cases through traditional means is challenging. Web archive collections consist of seeds which are the URIs of the original resources that archivists capture as mementos. For each seed, there are one or more mementos representing an original resource observed at different points in time. Some web archive collections contain hundreds of thousands of seeds (e.g., *Government of Canada Publications* has 339,166 seeds), and thus there is a multiplier effect concerning the number of mementos. There are more than 14,000 public Archive-It collections, many of which cover the same topic. Library science would dictate that experts apply metadata to answer questions about these collections, but few collections contain standardized metadata. Too many documents from too many collections with not enough metadata makes collection understanding an expensive proposition.

We turn to automation to develop visualizations that aid in collection understanding. Social media storytelling provides us with a visualization paradigm where documents are rendered as surrogates that contain document metadata. Because evaluating thousands of surrogates is no more helpful than evaluating thousands of mementos, we select exemplars so that a user only has to review the mementos that best describe the collection. For each exemplar, we generate document metadata and create a surrogate. These surrogates are bound together in a story augmented with story metadata and distributed so others can view it. We chose social media storytelling because it is an interface with which users are already familiar.

Our storytelling processes, as shown in Figure 194, are:

- **Select exemplars** to represent the collection.
- **Generate story metadata** to augment the story for better information scent.
- **Generate document metadata** so we can produce surrogates with good information scent.
- **Visualize the story** based on the metadata to produce surrogates and represent story metadata.
- **Distribute the story** so that users can view it and understand the collection.

We introduced four research questions with this model in mind. This chapter summarizes how we have addressed them before itemizing our contributions and finishing with some final words.

8.1 ANSWERS TO OUR RESEARCH QUESTIONS

Our research questions related to the different parts of our storytelling process. Research Question 1 centers on types of collections and structural features, critical concepts for selecting exemplars, and generating story metadata. Question 2 involves the development of algorithms for solving the problem of selecting exemplars. With Question 3, we seek to understand how to best generate document metadata by studying surrogates that visualize that metadata. Finally, with question 4, we want to develop surrogates even when the user has not provided sufficient document metadata. These questions span our storytelling processes and have helped us develop

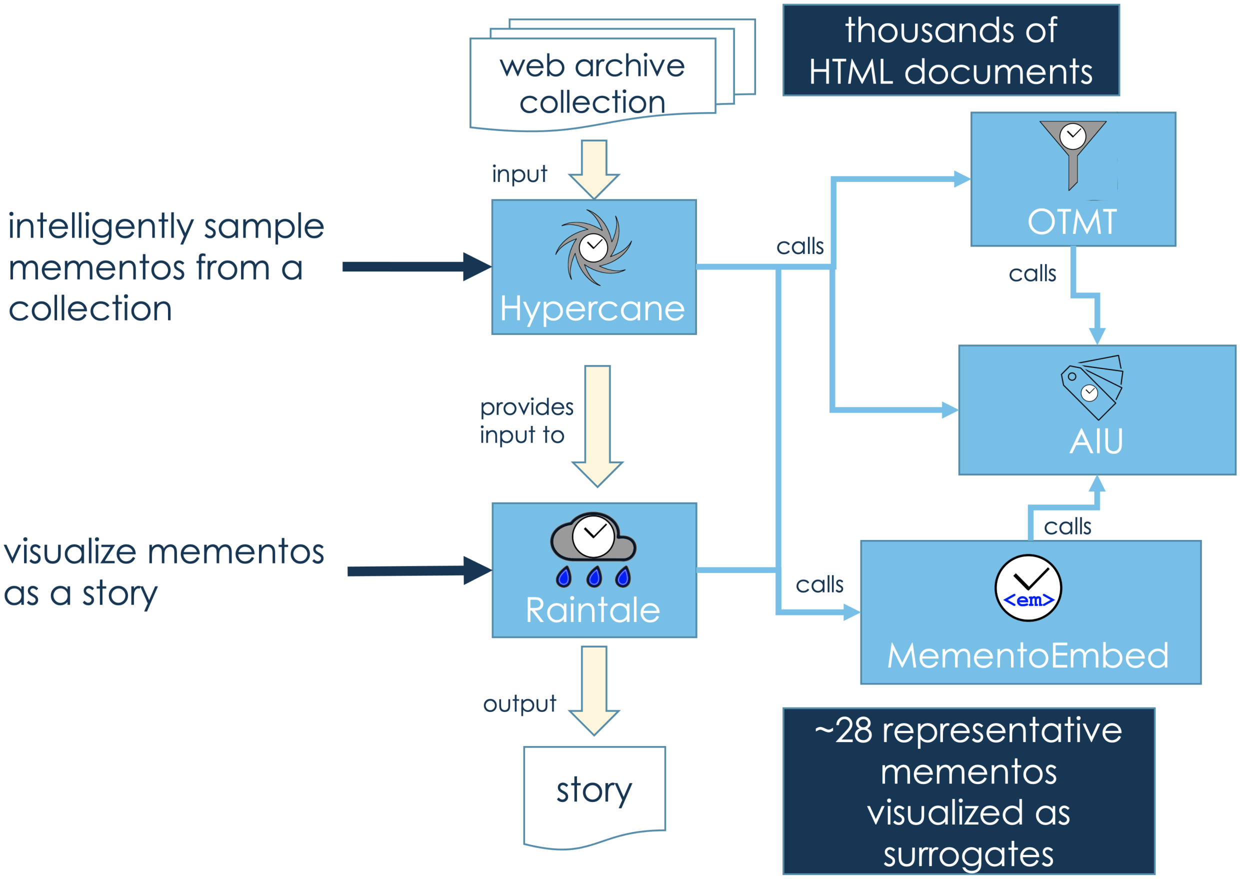

web archive collection
thousands of HTML documents
input
intelligently sample mementos from a collection
Hypercane
calls
OTMT
calls
AIU
provides input to
visualize mementos as a story
Raintale
calls
calls
<em>
MementoEmbed
output
story
~28 representative mementos visualized as surrogates

Fig. 195. The entire DSA architecture for storytelling.

tools – Hypercane, MementoEmbed, Raintale, and supporting libraries AIU and OTMT, shown in Figure 195 – that allow others to conduct storytelling with web archives.

8.1.1 RQ1: WHAT TYPES OF WEB ARCHIVE COLLECTIONS EXIST AND WHAT STRUCTURAL FEATURES DO THEY HAVE?

In Chapter 4.1 we explored the different types of web archive collections by manually reviewing the metadata of 3,382 Archive-It collections in 2017. We placed them into four different semantic categories:

- **Self-Archiving** collections consist of one or more domains belonging to the archiving organization or being archived as part of some initiative of which the collecting agency is part.
- **Subject-based Archiving** collections consist of seeds bound by a single topic.
- **Time Bounded - Expected** collections focus on an expected, planned event such as a sports competition or an election.
- **Time Bounded - Spontaneous** collections are created after a spontaneous event such as a disaster or the start of a movement.

We analyzed the different types of stories possible for each of these semantic categories. We also determined how many Archive-It collections from 2017 fit into each category. The Self-Archiving category dominated Archive-It with 54.1% of collections in 2017. Subject-based archiving made up 27.6% of collections. Time Bounded collections comprise even smaller amounts, with 14.1% in the category of Time Bounded - Expected and 4.2% in Time Bounded - Spontaneous.

We explored the structural features of the collections through seed features:

- **Seed URI domain diversity** quantifies the spread of the collection seeds across different sources.
- **Seed URI path depth diversity** indicates if the seed URIs consist solely of top-level pages or a mixture of top-level pages and more specific content.
- **Most frequent seed URI path depth** gives us an idea of whether most of the seeds represent top-level pages or pages deeper in a site.

- The **% Query string usage** helps us understand the behavior of seeds with respect to query strings, potentially indicating they come from a Self-Archiving collection because many organizations with a high percentage are government or finacial institutions that can be slow to adopt the technologies that make query strings obsolete.

With seed features, we can compare collections and compare our set of selected exemplars to their source collection.

We also explored what collection growth curves could tell us about collections. A collection growth curve, like the one shown in Figure 196, contains the lifespan of the collection on the x-axis and the cumulative percentage of URIs on the y-axis. The URI represented by the y-axis depends on which one of two curves the viewer is evaluating. The seed curve (shown in green) demonstrates the archivist's curatorial involvement with the collection. It represents when seeds were added and can indicate when the archivist lost or regained interest in altering the collection's sources. The seed memento curve (shown in red) demonstrates the crawling behavior of the collection. It represents when an archivist added mementos to the collection. For example, in Figure 196, the green line indicates that 30% of the seeds were added at the beginning of the collection's life, with the rest added more continuously. The seed mementos, shown by the red line, were added mostly throughout the collection's life, because this red line is close to the diagonal. The red line indicates that there are more crawls of some of the early seeds and that the crawling schedule was not changed much during the collection's life.

Growth curves have the following features:

- The **number of seeds** is a count of the collection's seeds.
- The **number of mementos** is a count of the collection's seed mementos.
- The **difference between the seed curve AUC and diagonal** gives us an idea of when an archivist added seeds to the collection.
- The **difference between the seed memento curve AUC and diagonal** informs us of the balanced size of temporal clusters in the collection.
- The **difference between the seed curve AUC and seed memento curve AUC** provides us with an understanding of the relationship between seed growth and seed memento growth.
- The **collection lifespan** is the difference between the first and last memento-datetimes of all mementos in the collection.

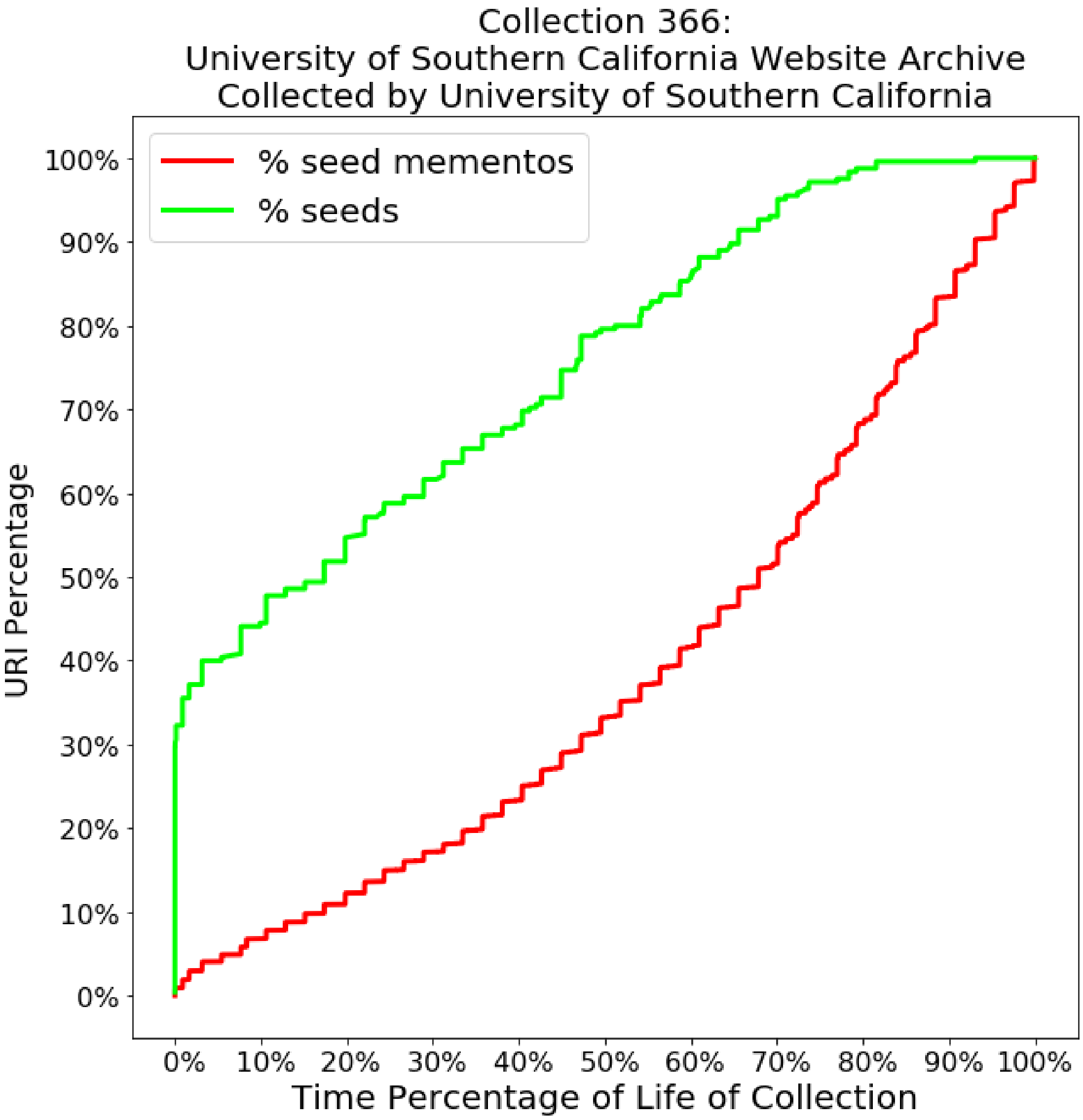


Fig. 196. An example of a collection growth curve.

Growth curves have implications for how we select exemplars. Temporal clustering is a necessary step for some types of stories. A clustering algorithm that does not consider the shape of the collection's growth curve may place mementos from the same time period in different clusters, leading to less representative exemplars.

In Chapter 4.5 we introduced Hypercane, which can produce reports of these seed features and growth curve features for a given collection. Thus, we have not only enumerated and quantified the types of web archive collections, but we also provided an implementation that calculates these structural features for further research.

8.1.2 RQ2: WHICH APPROACHES WORK BEST FOR SELECTING EXEMPLARS FROM WEB ARCHIVE COLLECTIONS?

The first step for selecting exemplars is to eliminate mementos of low information value, so further steps do not waste time evaluating them. Off-topic mementos, such as those from captured error pages, defaced websites, and pages whose topic has drifted, have low information value. Our solution from Chapter 4.2 builds on AlNomanay's research [10] that compares the first memento in a TimeMap to subsequent mementos to measure the degree to which subsequent mementos are off-topic. If subsequent mementos' degree exceeds some threshold, then we consider it off-topic. We evaluated multiple similarity measures to determine which worked best for identifying off-topic mementos and found that differences in word count (Figure 197) provides the best results.

Filtering for off-topic mementos is not the only step in selecting exemplars. In Chapter 4.3 we cover the primitives used to select exemplars:

- **Sample** refers to the action of creating a small sample by choosing k mementos from N such that $k << N$. This primitive can take the form of probabilistic sampling methods like random sampling, or *sample* can be a container for more complex intelligent sampling algorithms build from other primitives, like *filter* and *cluster*.

- **Filter** is a primitive that reduces the collection by some criteria, such as by removing off-topic or near-duplicate mementos.

- **Order** organizes the mementos by some criteria, such as sorting them by their memento-datetime.

- **Score** applies a function to our mementos so we can elevate some over others, such as how well a memento's content matches a query.

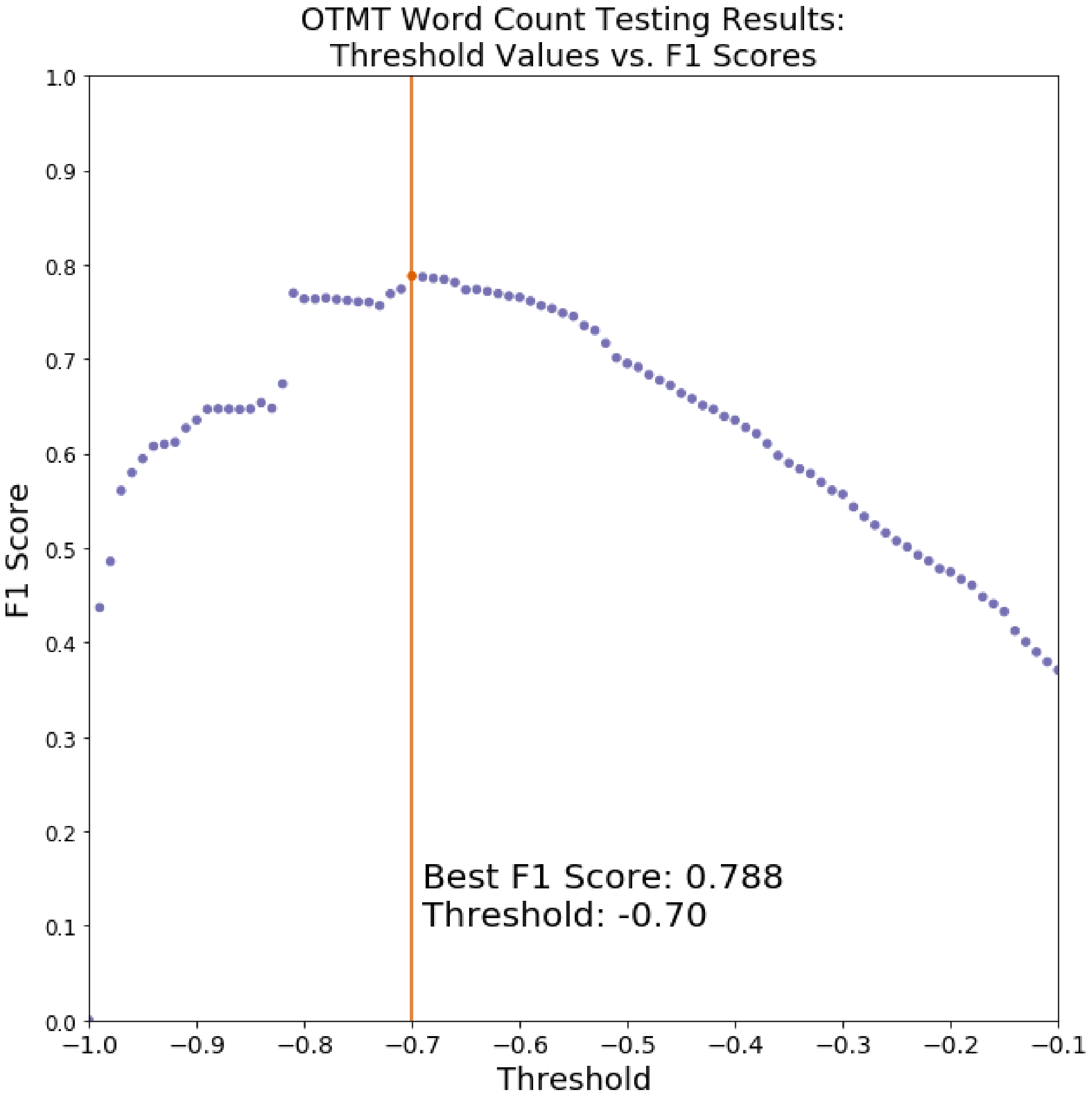
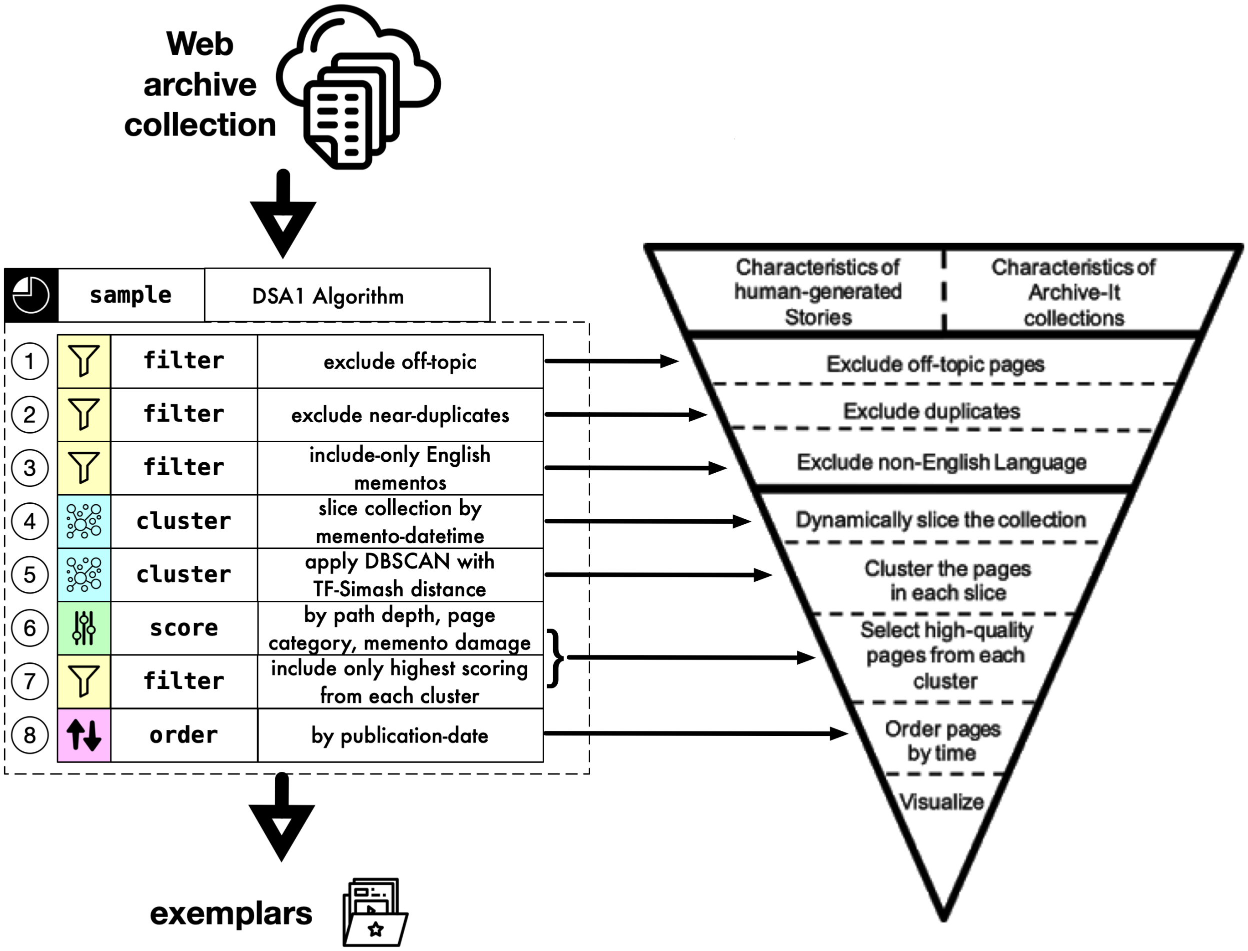


Fig. 197. Word Count was the similiarity metric that performed best for identifying off-topic mementos in a collection.


Web archive collection
exemplars
Characteristics of human-generated Stories
Characteristics of Archive-It collections
Exclude off-topic pages
Exclude duplicates
Exclude non-English Language
Dynamically slice the collection
Cluster the pages in each slice
Select high-quality pages from each cluster
Order pages by time
Visualize

		sample	DSA1 Algorithm
1		filter	exclude off-topic
2		filter	exclude near-duplicates
3		filter	include-only English mementos
4		cluster	slice collection by memento-datetime
5		cluster	apply DBSCAN with TF-Simash distance
6		score	by path depth, page category, memento damage
7		filter	include only highest scoring from each cluster
8		order	by publication-date

Fig. 198. We can combine our exemplar selection primitives to form AlNoamany's Algorithm, referred to in this dissertation as DSA1.

- **Cluster** groups mementos by some common feature and divides the collection by these groupings.

These primitives are derived from research like that performed by Sipos [286], Li [195], Zhang [338], Mihalcea [221], and AlNoamany [11]. We combine these primitives to reproduce AlNoamany's algorithm (which we also refer to as DSA1, shown in Figure 198) and use them to build three other algorithms.

Searching is the current manual method of discovering and selecting exemplars. A user like Natasha can employ a search engine to understand a collection. In Chapter 4.4 we create stories for eight collections with different structural features and of different semantic types. We then

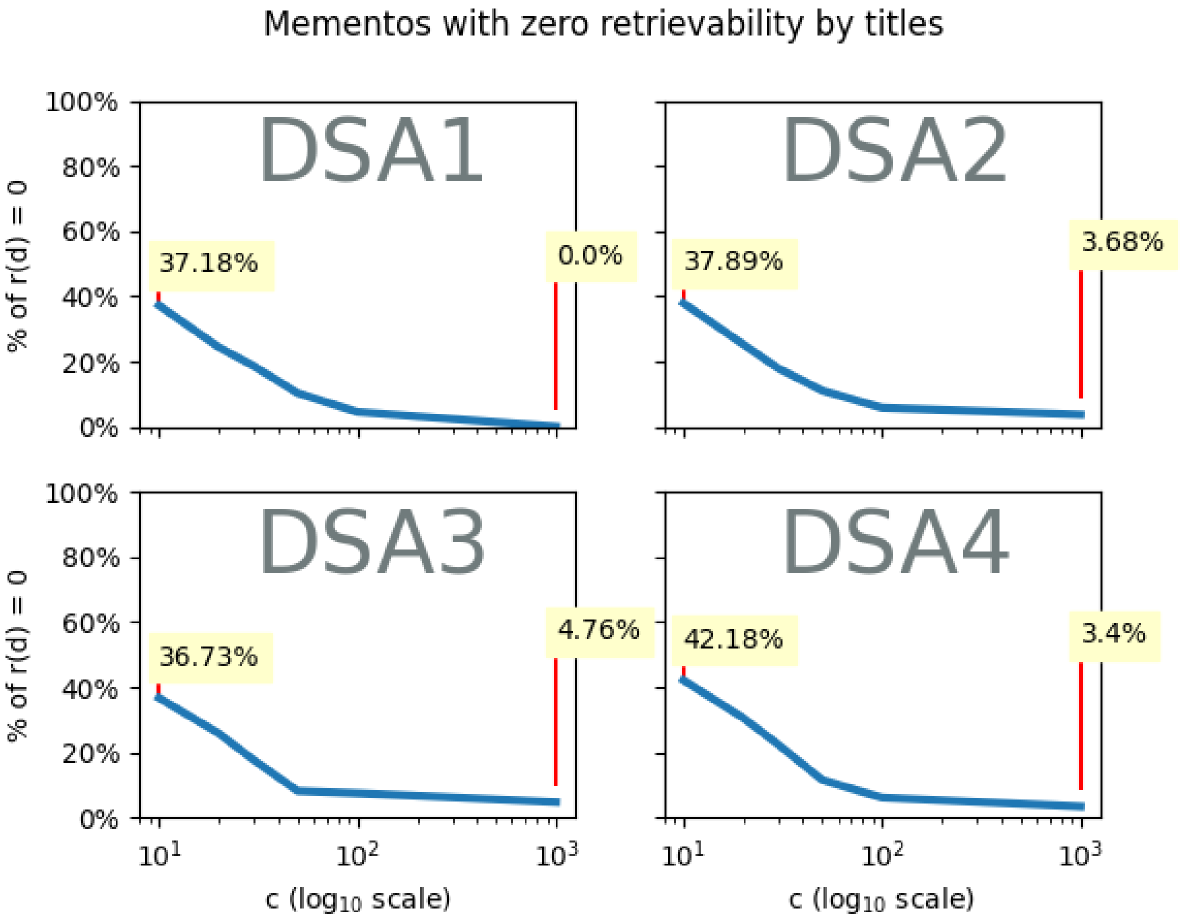


Fig. 199. Percentage of story mementos with zero retrievability when using titles as the search term.

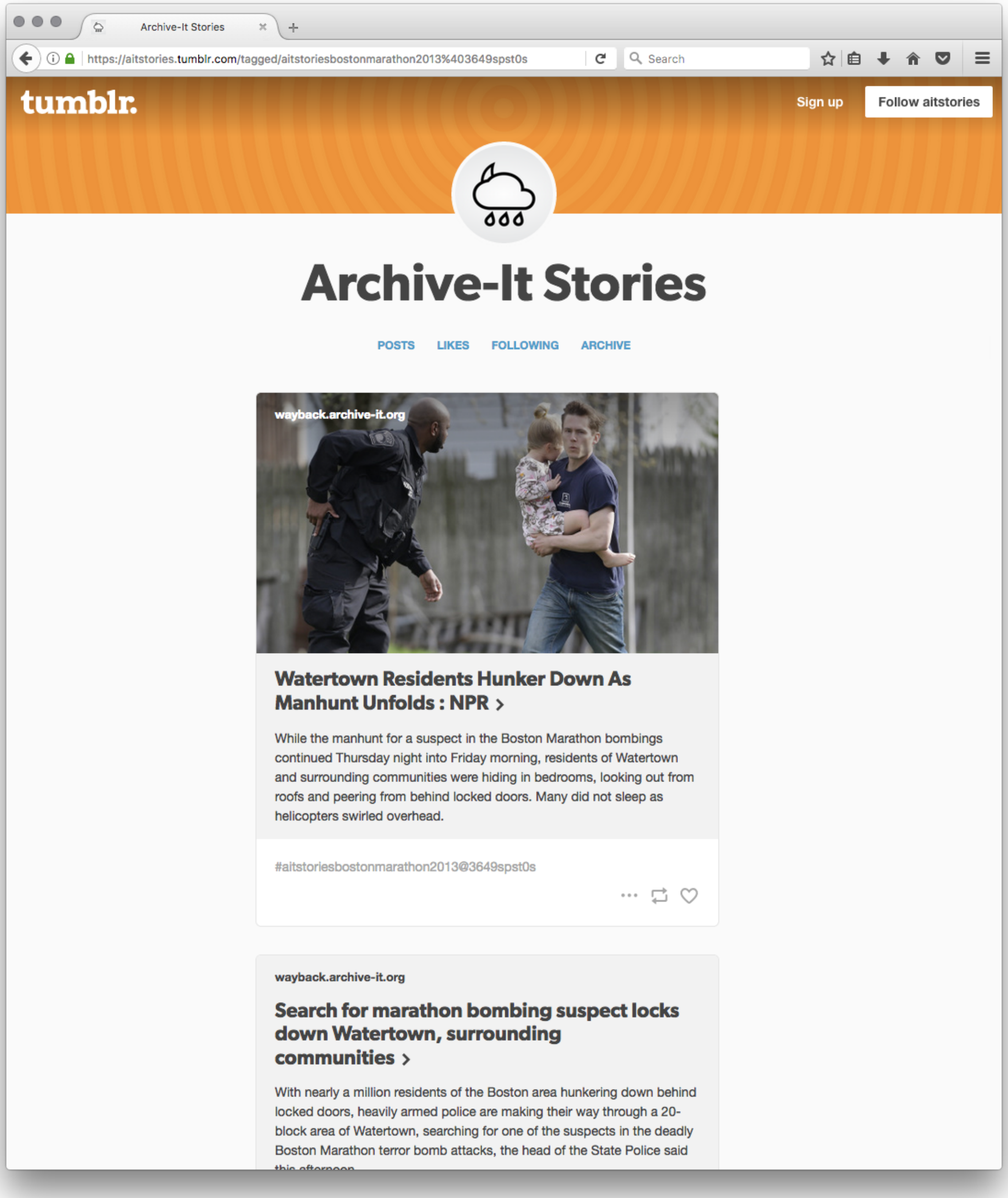
index them with SolrWayback, a standard web archive search engine. We also generate stories using each of our four DSA algorithms. We apply the concept of retrievability to measure how difficult the exemplars produced by our DSA algorithms are to retrieve using a search engine. Figure 199 visualizes how many exemplars are not retrieved at a given result count cutoff c when using a document title as the query, which Klein et al. showed to be very effective for retrieving a known document from a search engine [182]. Even with these queries, a user would need to examine at least one thousand results to find all of the exemplars surfaced by the DSA1 algorithm. For the other three DSA algorithms, some exemplars are still missing from search engine results. Thus, our methods surface more novel and quality exemplars than would otherwise be discovered conventionally with SolrWayback and existing query generation techniques.

We identified the best approaches for identifying off-topic mementos. We outlined a model of primitives to develop algorithms that surface exemplars from a collection. Finally, we show that algorithms derived from our model surface more novel exemplars from the collection than the current manual method of applying a search engine. The best method of identifying exemplars depends on the story that one wishes to tell. If everything in the collection is relevant, after filtering with methods such as off-topic detection, then our algorithms produce more novel results than search engines. Hypercane (Chapter 4.5) is our implementation of this work, and we used it in these experiments.

8.1.3 RQ3: WHAT SURROGATES WORK BEST FOR UNDERSTANDING GROUPS OF MEMENTOS?

Our initial thought was to reuse an existing platform like Storify because it was the de-facto standard. Adobe shut down Storify in 2018 [289], causing us to search for a replacement. We summarized the results of a review of 55 platforms in Chapter 5.1 (more details in Jones et al. [140, 156]) to uncover that none reliably produce sufficient surrogates for mementos. Some fail to produce surrogates at all, while others misattribute content to the web archive instead of its original source (Figure 200). From these results, no existing platform would work best for understanding groups of mementos.

In Chapter 5.2 we analyzed the default Archive-It surrogate consisting of a seed URI and first and last memento-datetimes and discovered that users could still glean information from the seed URI. We conducted a user study comparing user responses to the existing Archive-It surrogate, social cards, browser thumbnails, and combinations of cards and thumbnails. We discovered that users answered questions most accurately with social cards compared to the Archive-It surrogate at $p = 0.0569$. We also noted that social cards encouraged fewer users to interact with them



Archive-It Stories

Watertown Residents Hunker Down As Manhunt Unfolds : NPR ›

While the manhunt for a suspect in the Boston Marathon bombings continued Thursday night into Friday morning, residents of Watertown and surrounding communities were hiding in bedrooms, looking out from roofs and peering from behind locked doors. Many did not sleep as helicopters swirled overhead.

#aitstoriesbostonmarathon2013@3649spst0s

wayback.archive-it.org

Search for marathon bombing suspect locks down Watertown, surrounding communities ›

With nearly a million residents of the Boston area hunkering down behind locked doors, heavily armed police are making their way through a 20-block area of Watertown, searching for one of the suspects in the deadly Boston Marathon terror bomb attacks, the head of the State Police said this afternoon

Fig. 200. Tumblr creates cards for mementos, but misattributes them to the web archive rather than their original source.

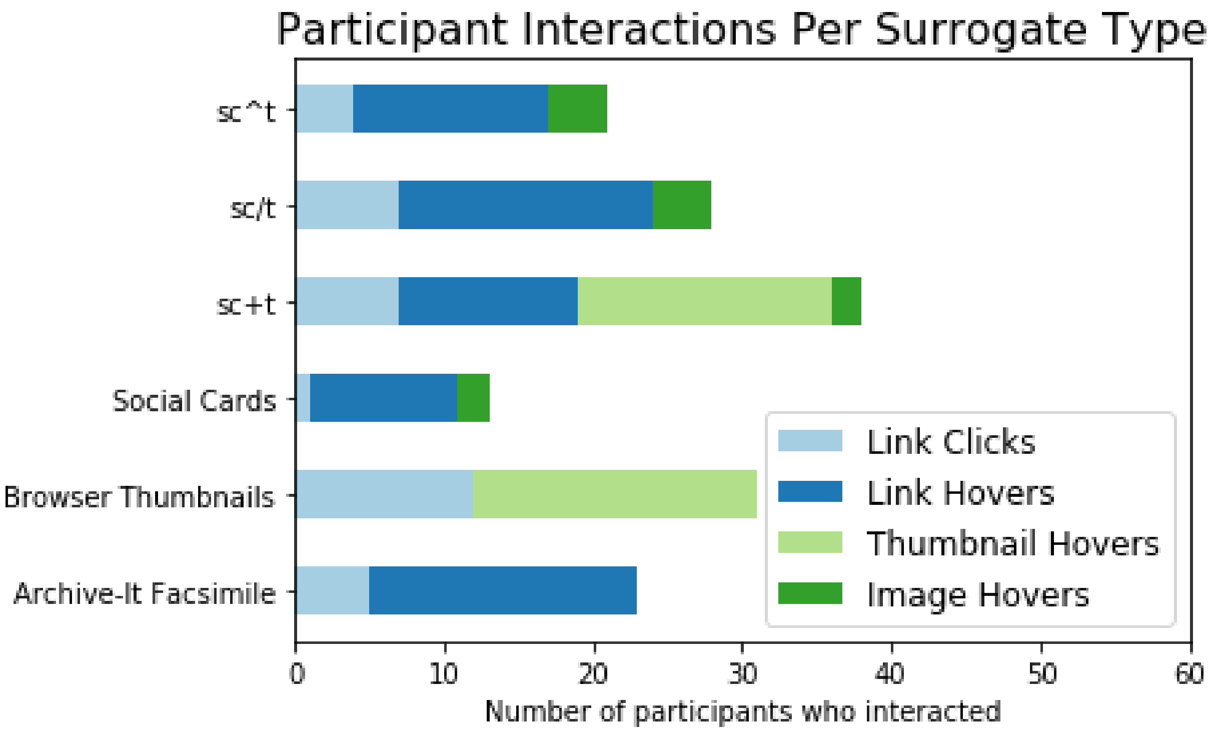


Fig. 201. The number of users interacting per surrogate type, broken down by image hovers, link hovers, and link clicks.

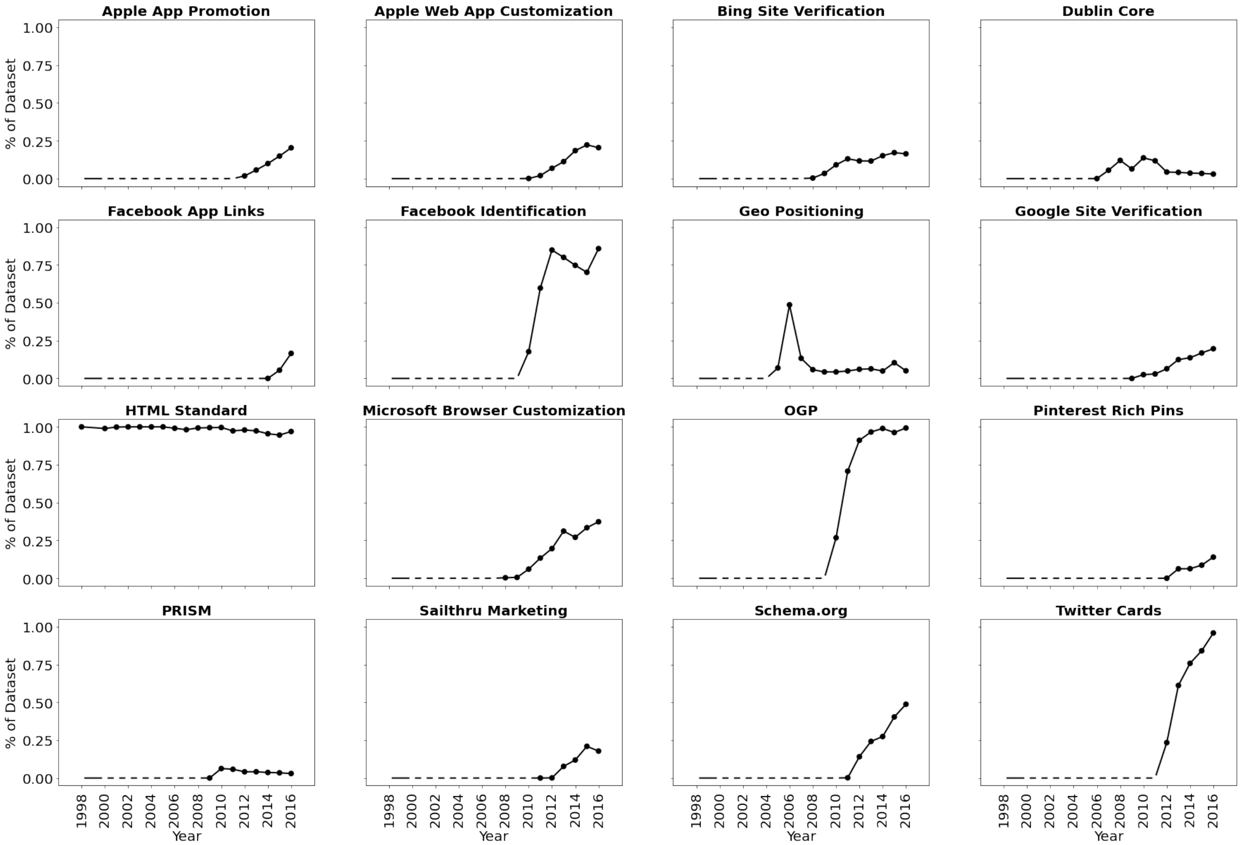


Fig. 202. By memento-datetime, the percentage of news articles each year that have adopted a given metadata category. Dashed lines indicate a y-axis value of 0. We see rapid growth in the adoption of OGP and Twitter Card metadata between 2010 and 2016.

(Figure 201). With social cards, users were able to correctly answer our questions without as much interaction. Thus, with these results, we concluded that social cards work best for understanding groups of mementos.

8.1.4 RQ4: WHAT METHODS THAT AUTOMATE THE CREATION OF SURROGATES PRODUCE RESULTS THAT BEST MATCH HUMANS' BEHAVIOR?

Social cards require that web page authors supply metadata so platforms can fill in the appropriate card fields for striking image, title, and description. In Chapter 5.3 we explore how to create social cards when this metadata is missing. This metadata may be absent due to a lack of interest on the part of the author. Many mementos, on the other hand, predate the existence of these metadata standards. This environment makes it challenging to produce good social cards for

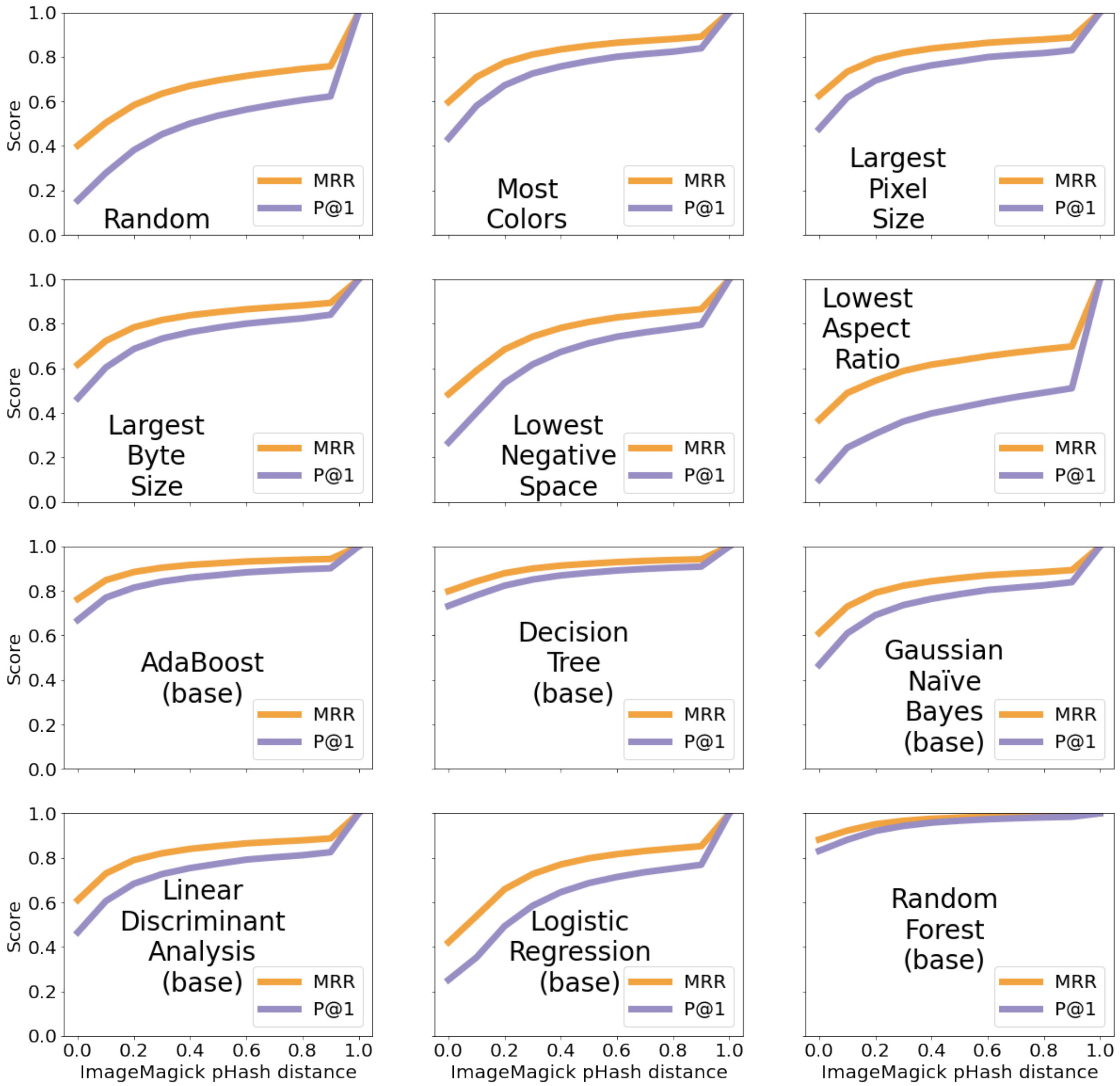


Fig. 203. This visualization demonstrates the *MRR* and *P*@1 results for different striking image prediction approaches as run against our sample. The best approach achieves the highest *MRR* and *P*@1 at the lowest pHash distance, making the ideal situation one where an approach's lines start higher into a graph's upper left corner. Items in parentheses are feature categories applied to the classifier, if applicable.

mementos reliably. We analyzed 296,162 mementos of news articles for which metadata fields were in use. As seen in Figure 202, once social card metadata fields (OGP and Twitter Cards) were standardized in 2010, adoption of these fields skyrocketed. For mementos captured before 2010, we can apply the HTML standard description field to create cards with an author-supplied description, but there is no similar field for striking images.

With this in mind, we analyzed how to best generate descriptions. As listed in Chapter 3.1, there are many algorithms, such as TextRank, for selecting sentences from the text. These algorithms can help us develop descriptions but require that we specify the length of the description. From analyzing the different description fields in our mementos, authors generate descriptions with mean sizes of 268 characters, 52 words, or two sentences.

Having adopted this solution for descriptions, we needed a way to select striking images. For our ground truth, we applied the striking images chosen by the authors of 37,522 news articles. For each article, we scored the images present in the article by applying different evaluation methods and chose the highest scoring image for that method. We then compared methods to each other in terms of precision at one ($P@1$) and mean reciprocal rank (MRR) to determine how accurate one combination of features and methods was to another. As shown in our results in Figure 203, we found that base features like image byte size, size in pixels, aspect ratio, and color count applied to the Random Forest classifier most often produced the same striking image as the author.

Social card surrogates consist of a title, description, and striking image, and different metadata fields define these units. Without metadata to guide our system, we must rely on automation to generate the metadata. By analyzing the existing metadata, we have determined how to automate the creation of surrogates such that their metadata best matches what web page authors would create. We realize this capability in MementoEmbed and its companion storytelling tool Raintale.

8.2 CONTRIBUTIONS

We have built on the work of AlNoamany to provide models and tools that enable future work with web archives. Now users can select exemplars, generate story and document metadata, and visualize and distribute their stories. We covered some future directions in Chapter 7 that focus on improvements to this process, including new structural features, additional ideas for off-topic detection, more story types, summaries for non-HTML file formats, and further evaluation. Our contributions support storytelling with web archiving and assist other research efforts with web archives and web science:

1. We established a model for storytelling with web archives consisting of five processes: select exemplars, generate story metadata, generate document metadata, visualize the story, and

distribute the story.

2. We evaluated Archive-It collections and developed a vocabulary for describing different types of collections which the web archiving community is applying to other work [1, 318]. We found that 54.1% of collections are organizations archiving their own content.

3. We introduced structural features for Archive-It collections and demonstrated how they could be applied to predict the semantic category of a collection [160].

4. We established that word count is an effective intra-TimeMap method of identifying off-topic mementos in a web archive collection [141, 164], outperforming more sophisticated and slower methods.

5. Based on existing work on selecting exemplars from documents and collections [14, 79, 83, 195, 206, 221, 338] we devised a set of primitives in Chapter 4.3 that could be combined to intelligently sample mementos from a web archive collection. These primitives are sample, filter, cluster, score, and order.

6. We established that four different algorithms produced from these primitives will select exemplars that are otherwise undiscoverable using a state-of-the-art web archive search engine and conventional query generation techniques.

7. We developed the tool Hypercane [144, 145, 146, 149, 157, 166] that can select exemplars and generate story metadata for storytelling. It applies the primitives from our model defined in Chapter 4.3.

8. Hypercane can also synthesize WARCs and generate queries to assist further research with web archives. We applied this capability in the final study of Chapter 4.4.

9. Our user study from Chapter 5.2 provides engineers support for choosing social cards over other surrogate types when designing a platform that requires surrogates [165]. Despite prior work evaluating users and surrogates for search engine result relevance, surrogates had never been evaluated for understanding.

10. Our evaluation of metadata in Chapter 5.3 sheds light on the reasons for metadata adoption. We have expanded this work into an additional paper [158] to provide more insight and show that the trend of high adoption for social card metadata continues to hold when we examine individual metadata fields even at the expense of older metadata standards.

11. We established methods for generating the metadata for social cards if a page's author fails to supply it [163]. Prior work did not evaluate as large a dataset and did not use the authors' input as ground truth. These results should assist some of the platforms we reviewed in Chapter 5.1 who do not seem to be able to produce social cards consistently.

12. We created the tool MementoEmbed [151, 156] to be an archive-aware document metadata and surrogate service. Future projects can apply the metadata generated by MementoEmbed for further document analysis.

13. We produced the tool Raintale [143, 156] that will allow others to visualize stories in a variety of formats and distribute them to popular social media platforms or even plain web servers.

Hypercane, MementoEmbed, and Raintale are currently being piloted with the National Library of Australia as part of a grant from the International Internet Preservation Consortium [234].

8.3 CLOSING THOUGHTS

AlNoamany's work [12] provided a foundation for us to build on, but it was tied to the web landscape of 2016. In five years, we have seen target platforms die (e.g., Storify), and new definitions of storytelling arise (e.g., Snapchat stories). This new environment required a rethinking of the whole process. Along the way, we opened up the processes of storytelling to new frontiers.

The questions we started with were analogous to our storytelling model. Which mementos are worthy of storytelling? How do we visualize them? How do we distribute the whole? How do we augment the story with additional content? How do we know if what we are doing is better than the existing state-of-the-art? We have addressed these questions and more.

As noted in Chapter 3.1 existing web archive exploration tools (e.g., ArchiveSpark) are designed for archivists who have access to their own WARCs. Instead of being constrained to these limitations, we applied the Memento Protocol [155, 312] to ensure that we and others could overcome them. With Hypercane, Raintale, and MementoEmbed, web archive users will be able to tell their own stories. Consider the next project where you need to apply documents from the past. You might need to decide between web archive collections. You may need to explore a particular aspect of a collection. With the tools and approaches we have presented, what story will you tell with web archives?

REFERENCES

[1] Samantha Abrams, Alexis Antracoli, Rachel Appel, Celia Caust-Ellenbogen, Sarah Denison, Sumitra Duncan, and Stefanie Ramsay. 2019. Sowing the Seeds for More Usable Web Archives: A Usability Study of Archive-It. *The American Archivist* 82, 2 (2019). https://doi.org/10.17723/aarc-82-02-19

[2] Myriam Abramson and David Aha. 2012. What's in a URL? Genre Classification from URLs. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/download/5252/5624

[3] Scott G. Ainsworth, Michael L. Nelson, and Herbert Van de Sompel. 2015. Only One Out of Five Archived Web Pages Existed as Presented. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. 257–266. https://doi.org/10.1145/2700171.2791044

[4] Hilal Al Maqbali, Falk Scholer, James Thom, and Mingfang Wu. 2010. Evaluating the Effectiveness of Visual Summaries for Web Search. In *Proceedings of the 15th Australasian Document Computing Symposium*. Melbourne, Australia, 1–8. http://www.cs.rmit.edu.au/adcs2010/proceedings/pdf/paper%2013.pdf

[5] Lulwah Alkwai. 2019. *Expanding the Usage of Web Archives by Recommending Archived Webpages Using Only the URI*. Ph.D. Dissertation. Old Dominion University, Norfolk, Virginia, USA. https://doi.org/10.25777/yk35-dd38

[6] Lulwah M. Alkwai, Michael L. Nelson, and Michele C. Weigle. 2020. Making Recommendations from Web Archives For "Lost" Web Pages. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. Virtual Event, China, 87–96. https://doi.org/10.1145/3383583.3398533

[7] Yasmin AlNoamany. 2016. *Using Web Archives to Enrich the Live Web Experience Through Storytelling*. Ph.D. Dissertation. Old Dominion University, Norfolk, Virginia, USA. https://doi.org/10.25777/zm0w-gp91

[8] Yasmin AlNoamany. 2016. Using Web Archives to Enrich the Live Web Experience Through Storytelling - Ph.D. defense presentation. https:

//www.slideshare.net/yasmina85/using-web-archives-to-enrich-the-live-web-experience-through-storytelling-65806319 Retrieved 25-May-2021.

[9] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. 2016. Characteristics of social media stories: What makes a good story? *International Journal on Digital Libraries* 17, 3 (2016), 239–256. https://doi.org/10.1007/s00799-016-0185-3

[10] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. 2016. Detecting off-topic pages within TimeMaps in Web archives. *International Journal on Digital Libraries* 17, 3 (2016), 203–221. https://doi.org/10.1007/s00799-016-0183-5

[11] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. 2017. Generating Stories From Archived Collections. In *Proceedings of the 2017 ACM on Web Science Conference*. Troy, New York, USA, 309–318. https://doi.org/10.1145/3091478.3091508

[12] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. 2017. *Stories From the Past Web*. Technical Report 1705.06218. Old Dominion University. http://arxiv.org/abs/1705.06218

[13] Yasmin AlNoamany, Michele C. Weigle, Michael L. Nelson, and Shawn Jones. 2017. Dataset: Human Stories Work for evaluating the DSA Framework. (2017). https://doi.org/10.6084/m9.figshare.5701054

[14] Yasmin AlNoamany, Michele C. Weigle, Michael L. Nelson, and Shawn Jones. 2017. Story Contents Uploaded to Storify, produced by the Dark and Stormy Archives Framework. (2017). https://doi.org/10.6084/m9.figshare.5701051.v1

[15] Omar Alonso, Vasileios Kandylas, Serge-Eric Tremblay, Jake M. Hofman, and Siddhartha Sen. 2017. What's Happening and What Happened: Searching the Social Web. In *Proceedings of the 2017 ACM Conference on Web Science*. Troy, New York, USA, 191–200. https://doi.org/10.1145/3091478.3091484

[16] Ahmed AlSum and Michael L. Nelson. 2014. Thumbnail Summarization Techniques for Web Archives. In *Advances in Information Retrieval*. Vol. 8416. Springer International Publishing, Cham, 299–310. https://doi.org/10.1007/978-3-319-06028-6_25

[17] American Library Association's ALCTS/LITA/RUSA Machine-Readable Bibliographic Information Committee and Network Development and MARC Standards Office Library of Congress. 1996. The MARC 21 Formats: Background and Principles. https://www.loc.gov/marc/96principl.html

[18] Lorin W Anderson, David R Krathwohl, Peter W Airasian, Kathleen A Cruikshank, Richard E Mayer, Paul R Pintrich, James Raths, and Merlin C Wittrock. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, abridged edition*. Longman, Harlow, England, UK.

[19] Ann Apps. 2005. Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata. http://www.dublincore.org/specifications/dublin-core/dc-citation-guidelines/

[20] Gaby Arancibia. 2014. How Storify can be a new platform for students. https://www.eschoolnews.com/2014/06/26/storify-platform-students-373/ Retrieved 25-May-2021.

[21] Archive-It. 2021. Archive-It software release notes. https://support.archive-it.org/hc/en-us/sections/201873356-Archive-It-software-release-notes

[22] Mohamed Aturban, Mat Kelly, Sawood Alam, John A. Berlin, Michael L. Nelson, and Michele C. Weigle. 2018. ArchiveNow: Simplified, Extensible, Multi-Archive Preservation. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. Fort Worth, Texas, USA, 321–322. https://doi.org/10.1145/3197026.3203880

[23] Mohamed Aturban, Michael L. Nelson, and Michele C. Weigle. 2017. *Difficulties of Timestamping Archived Web Pages* . Technical Report 1712.03140. Old Dominion University. https://arxiv.org/abs/1712.03140 https://arxiv.org/abs/1712.03140.

[24] Anne Aula, Rehan M. Khan, Zhiwei Guan, Paul Fontes, and Peter Hong. 2010. A comparison of visual and textual page previews in judging the helpfulness of web pages. In *Proceedings of the 19th international conference on World Wide Web*. Raleigh, North Carolina, USA, 51–60. https://doi.org/10.1145/1772690.1772697

[25] Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: An Evaluation Measure for Higher Order Information Access Tasks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (Napa Valley, California, USA). New York, NY, USA, 561–570. `https://doi.org/10.1145/1458082.1458157`

[26] Belinda Barnet. 2005. The Magical Place of Literary Memory: Xanadu. `http://tlweb.latrobe.edu.au/humanities/screeningthepast/firstrelease/fr_18/BBfr18a.html` Retrieved 14-July-2019.

[27] Belinda Barnet. 2008. The Technical Evolution of Vannevar Bush's Memex. *Digital Humanities Quarterly* 2, 1 (2008), 1–15. `http://www.digitalhumanities.org/dhq/vol/002/1/000015/000015.html`

[28] Belinda Barnet. 2010. Crafting the User-Centered Document Interface: The Hypertext Editing System (HES) and the File Retrieval and Editing System (FRESS). *Digital Humanities Quarterly* 4, 1 (2010), 1–9. `http://www.digitalhumanities.org/dhq/vol/4/1/000081/000081.html`

[29] Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the Similarity Function of TextRank for Automated Summarization. In *44 Jornadas Argentinas de Informática, Argentine Symposium on Artificial Intelligence (ASAI) 2015*. 65–72. `http://arxiv.org/abs/1602.03606`

[30] Regina Barzilay and Lillian Lee. 2004. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Boston, Massachusetts, USA, 113–120. `https://www.aclweb.org/anthology/N04-1015`

[31] T. Berners-Lee. 1989. Information Management: A Proposal. `https://www.w3.org/History/1989/proposal.html`

[32] Tim Berners-Lee. 1996. Universal Resource Identifiers – Axioms of Web Architecture. `https://www.w3.org/DesignIssues/Axioms.html#opaque` Retrieved 24-April-2018.

[33] Tim Berners-Lee, Tim Bray, Dan Connolly, Roy Fielding, Chris Lilley, David Orchard, Norman Walsh, and Stuart Williams. 2003. Architecture of the World Wide Web, Vol. 1. `https://www.w3.org/TR/webarch/` Retrieved 14-April-2019.

[34] T. Berners-Lee, R. Fielding, and L. Masinter. 2005. RFC 3986 - Uniform Resource Identifier (URI): Generic Syntax. https://tools.ietf.org/html/rfc3986

[35] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, California, USA. http://www.nltk.org/book/

[36] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 3/1/2003 (2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[37] Benjamin S. Bloom, David R. Krathwohl, and Bertram S. Masia. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* David McKay Company, Philadelphia, Pennsylvania, USA.

[38] Jay David Bolter and Michael Joyce. 1987. Hypertext and creative writing. In *Proceeding of the ACM conference on Hypertext*. Chapel Hill, North Carolina, USA, 41–50. https://doi.org/10.1145/317426.317431

[39] Horatiu Bota, Ke Zhou, and Joemon M. Jose. 2016. Playing Your Cards Right: The Effect of Entity Cards on Search Behaviour and Workload. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. Carrboro, North Carolina, USA, 131–140. https://doi.org/10.1145/2854946.2854967

[40] Karin Bredenberg, Regine Heberlein, Salvatore Vassallo, Wim van Dongen, Brian Tingle, Michael Rush, Ford Madox, and Ruth Tillman. 2019. EAD: Encoded Archival Description (EAD Official Site, Library of Congress). https://www.loc.gov/ead/

[41] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 5 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[42] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30, 1-7 (1998), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

[43] Martin Brinkmann. 2017. Adobe retires Storify but lets it live on as Storify 2 (sort of). https://www.ghacks.net/2017/12/13/adobe-retires-storify-but-lets-it-live-on-as-storify-2/ Retrieved 25-May-2021.

[44] Ilya Brown. 2021. Goodbye, Fleets. https://blog.twitter.com/en_us/topics/product/2021/goodbye-fleets

[45] Justin F. Brunelle, Mat Kelly, Hany SalahEldeen, Michele C. Weigle, and Michael L. Nelson. 2015. Not all mementos are created equal: measuring the impact of missing resources. *International Journal on Digital Libraries* 16, 3-4 (2015), 283–301. https://doi.org/10.1007/s00799-015-0150-6

[46] Justin F. Brunelle, Mat Kelly, Michele C. Weigle, and Michael L. Nelson. 2016. The impact of JavaScript on archivability. *International Journal on Digital Libraries* 17, 2 (2016), 95–117. https://doi.org/10.1007/s00799-015-0140-8

[47] Vannevar Bush. 1945. As We May Think: Information Literacy as a Discipline for the Information Age. *The Atlantic Monthly* 176, 1 (1945), 101–108. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

[48] Katriina Byström and Preben Hansen. 2005. Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology* 56, 10 (2005), 1050–1061. https://doi.org/10.1002/asi.20197

[49] Ricardo Campos, Arian Pasquali, Adam Jatowt, Vítor Mangaravite, and Alípio Mário Jorge. 2021. Automatic Generation of Timelines for Past-Web Events. In *The Past Web: Exploring Web Archives*. Springer International Publishing, Cham, 225–242. https://doi.org/10.1007/978-3-030-63291-5_18

[50] Robert Capra, Jaime Arguello, and Falk Scholer. 2013. Augmenting web search surrogates with images. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. San Francisco, California, USA, 399–408. https://doi.org/10.1145/2505515.2505714

[51] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Melbourne, Australia, 335–336. https://doi.org/10.1145/290941.291025

[52] Carlos Castillo. 2004. *Effective Web Crawling*. Ph.D. Dissertation. University of Chile, Santiago, Chile. http://chato.cl/papers/crawling_thesis/effective_web_crawling.pdf

[53] G. Changsheng, Z. Fuxi, and Y. Haitao. 2010. Sensitivity Analysis of Neural Network Parameters for Advertising Images Detection. In *2010 Second International Workshop on Education Technology and Computer Science*, Vol. 3. 76–79. https://doi.org/10.1109/ETCS.2010.380

[54] Moses S. Charikar. 2002. Similarity Estimation Techniques from Rounding Algorithms. In *Proceedings of the 34th ACM Symposium on Theory of Computing*. Montreal, Quebec, Canada, 380–388. https://doi.org/10.1145/509907.509965

[55] Mo Chen, Jian-Tao Sun, Hua-Jun Zeng, and Kwok-Yan Lam. 2005. A practical system of keyphrase extraction for web pages. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. Bremen, Germany, 277–278. https://doi.org/10.1145/1099554.1099625

[56] Ed H. Chi, Peter Pirolli, and James Pitkow. 2000. The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a Web site. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. The Hague, The Netherlands, 161–168. https://doi.org/10.1145/332040.332423

[57] Junghoo Cho, Hector Garcia-Molina, and Lawrence Page. 1998. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems* 30, 1 (1998), 161–172. https://doi.org/10.1016/S0169-7552(98)00108-1 Proceedings of the Seventh International World Wide Web Conference.

[58] François Chollet. 2017. *Deep Learning with Python*. Manning Publications Co.

[59] Alex Clark. 2015. Pillow (PIL Fork) Documentation. https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf Retrieved 4-February-2021.

[60] Wikimedia Commons. 2020. File:Copyright Card Catalog Files.jpg — Wikimedia Commons, the free media repository. https://commons.wikimedia.org/w/index.php?title=File:Copyright_Card_Catalog_Files.jpg&oldid=503166870 Retrieved 12-July-2021.

[61] Wikimedia Commons. 2021. File:WebCrawlerArchitecture.svg — Wikimedia Commons, the free media repository. https://commons.wikimedia.org/w/index.php?title=File:WebCrawlerArchitecture.svg&oldid=528110548 Retrieved 10-June-2021.

[62] Marcella Cornia, Stefano Pini, Lorenzo Baraldi, and Rita Cucchiara. 2018. Automatic Image Cropping and Selection Using Saliency: An Application to Historical Manuscripts. In *Digital Libraries and Multimedia Archives. IRCDL 2018*. 169–179. https://doi.org/10.1007/978-3-319-73165-0_17

[63] Miguel Costa, Daniel Gomes, and M. J. Silva. 2017. The evolution of web archiving. *International Journal on Digital Libraries* 18 (2017), 191–205. https://doi.org/10.1007/s00799-016-0171-9

[64] Miguel Costa and Mário J Silva. 2010. Understanding the Information Needs of Web Archive Users. In *Proceedings of the 10th International Web Archiving Workshop*. Vienna, Austria, 9–16. https://sobre.arquivo.pt/wp-content/uploads/understanding-the-information-needs-of-web-archive.pdf

[65] W. Bruce Croft, Donald Metzler, and Trevor Strohman. 2015. *Information Retrieval in Practice*. Pearson Education, Boston, Massachusetts, USA. https://ciir.cs.umass.edu/irbook/

[66] Tim Cuthbertson, Yuri Baburov, Sean Brant, Jerry Charumillind, Jan Weiß, Rick Harding, Nathan Breit, Mark Perdomo, Maxime Mouial, Drew Vogel, Zachary Denton, Andrey Popp, Gilles Dartiguelongue, Matthew Peters, Miguel Galves, Linas Valiukas, Dave Padovano, Marko Horvatić, and Martin Thurau. 2018. python-readability. https://github.com/buriy/python-readability Retrieved 5-April-2018.

[67] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Genoa, Italy, 449–454. http://www.lrec-conf.org/proceedings/lrec2006/pdf/440_pdf.pdf

[68] Ryan Deschamps, Samantha Fritz, Jimmy Lin, Ian Milligan, and Nick Ruest. 2019. The

Cost of a WARC: Analyzing Web Archives in the Cloud. In *2019 ACM/IEEE Joint Conference on Digital Libraries*. Urbana-Champaign, Illinois, USA, 261–264. https://doi.org/10.1109/JCDL.2019.00043

[69] Lee R. Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology* 26, 3 (1945), 297–302. https://doi.org/10.2307/1932409

[70] F. Dirfaux. 2000. Key frame selection to represent a video. In *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, Vol. 2. 275–278. https://doi.org/10.1109/ICIP.2000.899354

[71] Document Understanding Conference. 2014. DUC Past Data. https://duc.nist.gov/data.html

[72] Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland, 350–es. https://doi.org/10.3115/1220355.1220406

[73] Jackie Dooley and Kate Bowers. 2018. Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group. http://doi.org/10.25333/C3005C

[74] Susan Dumais, Edward Cutrell, and Hao Chen. 2001. Optimizing Search by Showing Results in Context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Seattle, Washington, USA, 277–284. https://doi.org/10.1145/365024.365116

[75] Susan T. Dumais. 2005. Latent semantic analysis. *Annual Review of Information Science and Technology* 38, 1 (2005), 188–230. https://doi.org/10.1002/aris.1440380105

[76] Geoff Duncan. 1998. Alas, HyperCard! *TidBITS* (1998). https://tidbits.com/1998/11/02/alas-hypercard/

[77] Susan Dziadosz and Raman Chandrasekar. 2002. Do Thumbnail Previews Help Users Make Better Relevance Decisions about Web Search Results?. In *Proceedings of the 25th*

Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland, 365–366. https://doi.org/10.1145/564376.564446

[78] Abigail Edge. 2014. How to: create stories using Storify. https://www.journalism.co.uk/skills/how-to-create-stories-using-storify/s7/a562894/ Retrieved 25-May-2021.

[79] H. P. Edmundson. 1969. New Methods in Automatic Extracting. *J. ACM* 16, 2 (1969), 264–285. https://doi.org/10.1145/321510.321519

[80] Thomas Egense, Jesper Lauridsen, Jørn Thøgersen, Toke Eskildsen, NielsGamborg, Asger Askov Blekinge, and Mat Kelly. 2021. SolrWayback. https://github.com/netarchivesuite/solrwayback Retrieved 25-May-2021.

[81] Samhaa R. El-Beltagy and Ahmed Rafea. 2009. KP-Miner: A keyphrase extraction system for English and Arabic documents. *Information Systems* 34, 1 (2009), 132–144. https://doi.org/10.1016/j.is.2008.05.002

[82] Douglas C. Engelbart, Richard W. Watson, and James C. Norton. 1973. The augmented knowledge workshop. In *Proceedings of the June 4-8, 1973, national computer conference and exposition*. New York, New York, USA, 9–21. https://doi.org/10.1145/1499586.1499593

[83] Gunes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research* 22 (2004), 457–479. https://doi.org/10.1613/jair.1523

[84] Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Portland, Oregon, USA, 226–231. https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

[85] Facebook. 2021. Facebook Reports First Quarter 2021 Results. https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-First-Quarter-2021-Results/default.aspx Retrieved 25-May-2021.

[86] Facebook. 2021. Facebook Stories: An Introduction for Content Creators. https://www.facebook.com/business/learn/lessons/facebook-stories-creators Retrieved 25-May-2021.

[87] Facebook. 2021. The Open Graph Protocol. http://ogp.me Retrieved 25-May-2021.

[88] Matthew Farrell, Edward McCain, Maria Praetzellis, Grace Thomas, and Paige Walker. 2018. Web Archiving in the United States: a 2017 Survey. https://ndsa.org/2018/12/12/announcing-publication-of-ndsa-s-2017-web-archiving-survey-report.html

[89] Steve Faulkner, Arron Eicholz, Travis Leithead, Alex Danilo, and Sangwhan Moon. 2017. HTML 5.2. https://www.w3.org/TR/html/ Retrieved 14-April-2019.

[90] R. Fielding, Y. Lafon, and J. Reschke. 2014. RFC 7233 - Hypertext Transfer Protocol (HTTP/1.1): Range Requests. https://tools.ietf.org/html/rfc7233

[91] R. Fielding, M. Nottingham, and J. Reschke. 2014. RFC 7234 - Hypertext Transfer Protocol (HTTP/1.1): Caching. https://tools.ietf.org/html/rfc7234

[92] R. Fielding and J. Reschke. 2014. RFC 7230 - Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing. https://tools.ietf.org/html/rfc7230

[93] R. Fielding and J. Reschke. 2014. RFC 7231 - Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content. https://tools.ietf.org/html/rfc7231

[94] R. Fielding and J. Reschke. 2014. RFC 7232 - Hypertext Transfer Protocol (HTTP/1.1): Conditional Requests. https://tools.ietf.org/html/rfc7232

[95] R. Fielding and J. Reschke. 2014. RFC 7235 - Hypertext Transfer Protocol (HTTP/1.1): Authentication. https://tools.ietf.org/html/rfc7235

[96] Kelly Fincham. 2011. 4 ways journalism educators are using Storify as a teaching tool. https://www.poynter.org/newsletters/2011/4-ways-journalism-educators-are-using-storify-as-a-teaching-tool/ Retrieved 25-May-2021.

[97] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of*

the 43rd Annual Meeting on Association for Computational Linguistics. Ann Arbor, Michigan, USA, 363–370. https://doi.org/10.3115/1219840.1219885

[98] Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"* (fourth ed.). Morgan Kaufmann, Burlington, Massachusetts, USA.

[99] Samantha Fritz and Ian Milligan. 2018. Analyze your Web Archives at Scale: The Archives Unleashed Cloud. https://archive-it.org/blog/post/analyze-your-web-archives-at-scale-the-archives-unleashed-cloud/ Retrieved 25-May-2021.

[100] Joe Garner. 1998. *We Interrupt This Broadcast*. Sourcebooks.

[101] L. Nancy Garrett, Karen E. Smith, and Norman Meyrowitz. 1986. Intermedia: issues, strategies, and tactics in the design of a hypermedia document system. In *Proceedings of the 1986 ACM conference on Computer-supported cooperative work*. Austin, Texas, USA, 163–174. https://doi.org/10.1145/637069.637090

[102] Bridget Gelms. 2014. Potential Uses for Teaching with Storify. https://www.insidehighered.com/blogs/gradhacker/potential-uses-teaching-storify Retrieved 25-May-2021.

[103] C. F. Goldfarb. 1991. Standards-HyTime: a standard for structured hypermedia interchange. *Computer* 24, 8 (1991), 81–84. https://doi.org/10.1109/2.84880

[104] Kenneth Goldsmith. 2013. *Seven American deaths and disasters*. PowerHouse Books.

[105] G. Golub and W. Kahan. 1965. Calculating the Singular Values and Pseudo-Inverse of a Matrix. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis* 2, 2 (1965), 205–224. https://doi.org/10.1137/0702016

[106] Daniel Gomes, João Miranda, and Miguel Costa. 2011. A Survey on Web Archiving Initiatives. In *Proceedings of Theory and Practice of Digital Libraries (TPDL)*. 408–420. https://doi.org/10.1007/978-3-642-24469-8_41

[107] Xuemei Gong, Weimao Ke, and Ritu Khare. 2012. Studying scatter/gather browsing for web search. *Proceedings of the American Society for Information Science and Technology* 49, 1 (2012), 1–4. https://doi.org/10.1002/meet.14504901328

[108] Yihong Gong and Xin Liu. 2000. Video summarization using singular value decomposition. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. (Cat. No.PR00662)*, Vol. 2. 174–180 vol.2. https://doi.org/10.1109/CVPR.2000.854772

[109] Google. 2019. Create good titles and snippets in Search Results. https://support.google.com/webmasters/answer/35624?hl=en Retrieved 14-July-2019.

[110] Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*. New Orleans, Louisiana, USA, 708–719. https://doi.org/10.18653/v1/N18-1065

[111] Rebecca Guenther. 2018. Metadata Application Profile for Description of Websites with Archived Versions Version 2. https://www.nyarc.org/sites/default/files/web-archiving-profile-version2.pdf

[112] Anthony Ha. 2013. Livefyre Acquires Storify, Says The Social Curation Service Will Still Operate As Standalone Product. https://techcrunch.com/2013/09/09/livefyre-acquires-storify/ Retrieved 25-May-2021.

[113] Frank Halasz and Mayer Schwartz. 1994. The Dexter hypertext reference model. *Commun. ACM* 37, 2 (1994), 30–39. https://doi.org/10.1145/175235.175237

[114] Austin Haugen. 2010. Abstract: The Open Graph Protocol Design Decisions. In *The Semantic Web – ISWC 2010*. 338–338. https://doi.org/10.1007/978-3-642-17749-1_25

[115] Takahiro Hayashi, Atsushi Ishikawa, and Rikio Onai. 2007. Landscape Image Retrieval with Query by Sketch and Icon. *Journal of Advanced Computational Intelligence and Intelligent Informatics* 11, 1 (2007), 61–70. https://doi.org/10.20965/jaciii.2007.p0061

[116] Brian Heater. 2017. Storify's standalone service is shutting down next year. https://techcrunch.com/2017/12/12/storifys-standalone-service-is-shutting-down-next-year/ Retrieved 25-May-2021.

[117] Cal Henderson. 2018. oEmbed. https://oembed.com Retrieved 24-April-2018.

[118] K. Holtman and A. Mutz. 1998. RFC 2295 - Transparent Content Negotiation in HTTP. https://tools.ietf.org/html/rfc2295

[119] Helge Holzmann, Vinay Goel, and Avishek Anand. 2016. ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. Newark, New Jersey, USA, 83–92. https://doi.org/10.1145/2910896.2910902

[120] Kai Hong and Ani Nenkova. 2014. Improving the Estimation of Word Importance for News Multi-Document Summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden, 712–721. http://www.aclweb.org/anthology/E14-1075

[121] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303 Retrieved 14-July-2021.

[122] Jianying Hu and Amit Bagga. 2004. Categorizing Images in Web Documents. *IEEE MultiMedia* 11, 1 (2004), 22–30. https://doi.org/10.1109/MMUL.2004.1261103

[123] IIPC. 2015. The CDX File Format (2015). https://iipc.github.io/warc-specifications/specifications/cdx-format/cdx-2015/ Retrieved 25-May-2021.

[124] Instagram. 2021. Instagram | Using Instagram Stories | Official Site. https://about.instagram.com/features/stories Retrieved 25-May-2021.

[125] Internet Archive. 2009. Internet Archive. https://web.archive.org/web/20090422100540/http://www.archive.org/index.php Retrieved 5-February-2021.

[126] Internet Archive. 2021. internetarchive/brozzler. https://github.com/internetarchive/brozzler [Online; accessed 14-July-2021].

[127] ISO/IEC/JTC1/SC34. 1997. ISO/IEC 10744:1997 : Information technology – Hypermedia/Time-based Structuring Language (HyTime). https://www.iso.org/standard/29303.html

[128] ISO/TC46/SC4. 2017. ISO28500:2017 Information and documentation – WARC file format. https://www.iso.org/standard/68004.html

[129] Paul Jaccard. 1912. The Distribution Of The Flora In The Alpine Zone. *New Phytologist* 11, 2 (1912), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

[130] Andy Jackson. 2015. Ten years of the UK web archive: what have we saved? https://anjackson.net/2015/04/27/what-have-we-saved-iipc-ga-2015/ Retrieved 3-April-2018.

[131] Andrew Jackson. 2019. Personal Communication.

[132] Andy Jackson, Toke Eskildsen, Thomas Egense, Gil Hoggarth, Leslie Bellony, Nick Ruest, Asger Askov Blekinge, William Palmer, and Mindaugas Vidmantas. 2021. ukwa/webarchive-discovery: WARC and ARC indexing and discovery tools. https://github.com/ukwa/webarchive-discovery Retrieved 25-May-2021.

[133] Adam Jatowt, Yukiko Kawai, and Katsumi Tanaka. 2008. Visualizing historical content of web pages. In *Proceedings of the 17th international conference on World Wide Web - WWW '08*. Beijing, China, 1221–1222. https://doi.org/10.1145/1367497.1367736

[134] Binxing Jiao, Linjun Yang, Jizheng Xu, and Feng Wu. 2010. Visual summarization of web pages. In *Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval*. Geneva, Switzerland, 499–506. https://doi.org/10.1145/1835449.1835533

[135] Thorsten Joachims. 1999. Transductive Inference for Text Classification Using Support Vector Machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, California, USA, 200–209. http://www1.cs.columbia.edu/~dplewis/candidacy/joachims99transductive.pdf

[136] Joint Steering Committee for Revision of AACR. 2005. *Anglo-American Cataloguing Rules*. American Library Association, Chicago, Illinois, USA.

[137] Karen Sparck Jones. 1972. A Statistical Interpretation Of Term Specificity And Its Application In Retrieval. *Journal of Documentation* 28, 1 (1972), 11–21. https://doi.org/10.1108/eb026526

[138] Shawn Jones. 2019. Off Topic Memento Toolkit (OTMT). https://github.com/oduwsdl/off-topic-memento-toolkit

[139] Shawn M. Jones. 2015. *Avoiding Spoilers on MediaWiki Fan Sites Using Memento*. Master's thesis. Old Dominion University, Norfolk, Virginia. http://doi.org/10.25777/d8hw-b984

[140] Shawn M. Jones. 2017. Where Can We Post Stories Summarizing Web Archive Collections? https://ws-dl.blogspot.com/2017/08/2017-08-11-where-can-we-post-stories.html Retrieved 3-May-2018.

[141] Shawn M. Jones. 2018. The Off-Topic Memento Toolkit. https://ws-dl.blogspot.com/2018/07/2018-07-02-off-topic-memento-toolkit.html Retrieved 12-July-2021.

[142] Shawn M. Jones. 2018. A Preview of MementoEmbed: Embeddable Surrogates for Archived Web Pages. https://ws-dl.blogspot.com/2018/08/2018-08-01-preview-of-mementoembed.html Retrieved 1-December-2018.

[143] Shawn M. Jones. 2019. Raintale – A Storytelling Tool For Web Archives. https://ws-dl.blogspot.com/2019/07/2019-07-11-raintale-storytelling-tool.html Retrieved 25-May-2021.

[144] Shawn M. Jones. 2020. Hypercane Hypercane Part 2: Synthesizing Output For Other Tools. https://ws-dl.blogspot.com/2020/06/2020-06-10-hypercane-part-2.html Retrieved 12-July-2021.

[145] Shawn M. Jones. 2020. Hypercane Part 1: Intelligent Sampling of Web Archive Collections. https://ws-dl.blogspot.com/2020/06/2020-06-03-hypercane-part-1-intelligent.html Retrieved 12-July-2021.

[146] Shawn M. Jones. 2020. Hypercane Part 3: Building Your Own Algorithms. https://ws-dl.blogspot.com/2020/06/2020-06-17-hypercane-part-3-building.html Retrieved 12-July-2021.

[147] Shawn M. Jones. 2020. Web API – MementoEmbed 0.2020.05.08.220018 documentation. https://mementoembed.readthedocs.io/en/latest/web_api.html Retrieved 25-May-2021.

[148] Shawn M. Jones. 2021. Dark and Stormy Archives. https://oduwsdl.github.io/dsa/ Retrieved 25-May-2021.

[149] Shawn M. Jones. 2021. Hypercane. https://oduwsdl.github.io/hypercane/ Retrieved 12-July-2021.

[150] Shawn M. Jones. 2021. Hypercane: Intelligent Sampling of Web Archive Collections — Hypercane 0.2021.06.10.005024 documentation. https://hypercane.readthedocs.io/en/latest/ Retrieved 25-May-2021.

[151] Shawn M. Jones. 2021. MementoEmbed: Archive-Aware Web Page Surrogates. https://mementoembed.readthedocs.io/en/latest/ Retrieved 25-May-2021.

[152] Shawn M. Jones and Yasmin AlNoamany. 2018. Off-Topic Gold Standard Dataset. https://github.com/oduwsdl/offtopic-goldstandard-data/tree/5139aca762e1ddac76da628436dbc48ae38807f2 Retrieved 25-May-2021.

[153] Shawn M. Jones, Himarsha Jayanetti, and Mat Kelly. [n.d.]. GitHub - oduwsdl/aiu: A library for interacting with web archive collections at Archive-It, Trove, Pandora, and more. https://github.com/oduwsdl/aiu Retrieved 25-May-2021.

[154] Shawn M. Jones, Martin Klein, and Herbert Van de Sompel. 2021. Robustifying Links To Combat Reference Rot. *Code4Lib Journal* 50, 2021-02-10 (2021). https://journal.code4lib.org/articles/15509

[155] Shawn M. Jones, Martin Klein, Herbert Van de Sompel, Michael L. Nelson, and Michele C. Weigle. 2021. Interoperability for Accessing Versions of Web Resources with the Memento Protocol . In *The Past Web: Exploring Web Archives*. Springer International Publishing , New York, 101 – 126 . https://doi.org/10.1007/978-3-030-63291-5_9

[156] Shawn M. Jones, Martin Klein, Michele C. Weigle, and Michael L. Nelson. 2020. *MementoEmbed and Raintale for Web Archive Storytelling*. Technical Report 2008.00137. Old Dominion University. https://arxiv.org/abs/2008.00137 Presented at 2020 Web Archiving and Digital Libraries Workshop.

[157] Shawn M. Jones, Valentina Neblitt-Jones, Michele C. Weigle, Martin Klein, and Michael L.

Nelson. 2021. Hypercane: Intelligent Sampling for Web Archive Collections. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2021*. [To be published in September 2021].

[158] Shawn M. Jones, Valentina Neblitt-Jones, Michele C. Weigle, Martin Klein, and Michael L. Nelson. 2021. It's All About The Cards: Sharing on Social Media Probably Encouraged HTML Metadata Growth. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2021*. [To be published in September 2021] preprint: `https://arxiv.org/abs/2104.04116`.

[159] Shawn M. Jones, Michael L. Nelson, and Herbert Van de Sompel. 2018. Avoiding spoilers: wiki time travel with Sheldon Cooper. *International Journal on Digital Libraries* 19, 1 (2018), 77–93. `https://doi.org/10.1007/s00799-016-0200-8`

[160] Shawn M Jones, Alexander Nwala, Michele C Weigle, and Michael L Nelson. 2018. The Many Shapes of Archive-It. In *Proceedings of the 15th International Conference on Digital Preservation*. Boston, Massachusetts, USA, 1–10. `https://doi.org/10.17605/OSF.IO/EV42P`

[161] Shawn M. Jones, Alexander C. Nwala, Martin Klein, Michele C. Weigle, and Michael L. Nelson. 2020. *SHARI – An Integration of Tools to Visualize the Story of the Day* . Technical Report arXiv:2008.00139. arXiv:2008.00139 [cs.DL] `https://arxiv.org/abs/2008.00139` Presented at 2020 Web Archiving and Digital Libraries Workshop.

[162] Shawn M. Jones, Herbert Van de Sompel, Harihar Shankar, Martin Klein, Richard Tobin, and Claire Grover. 2016. Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content. *PLOS ONE* 11, 12 (2016), e0167475. `https://doi.org/10.1371/journal.pone.0167475`

[163] Shawn M. Jones, Michele C. Weigle, Martin Klein, and Michael L. Nelson. 2021. Automatically Selecting Striking Images for Social Cards. In *Proceedings of the 13th ACM Web Science Conference*. 36–45. `https://doi.org/10.1145/3447535.3462505`

[164] Shawn M Jones, Michele C Weigle, and Michael L Nelson. 2018. The Off-Topic Memento Toolkit. In *Proceedings of the 15th International Conference on Digital Preservation*. Boston, Massachusetts, USA, 1–10. `https://doi.org/10.17605/OSF.IO/UBW87`

[165] Shawn M. Jones, Michele C. Weigle, and Michael L. Nelson. 2019. Social Cards Probably Provide For Better Understanding Of Web Archive Collections. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. Beijing, China, 2023–2032. https://doi.org/10.1145/3357384.3358039

[166] Shawn M. Jones, Michele C. Weigle, and Michael L. Nelson. 2021. Hypercane: Toolkit for Summarizing Large Collections of Archived Webpages. *SIGWEB Newsletter Autumn 2021* (2021). [to be published in 2021].

[167] Marcel Adam Just, G. Albyn Davis, and Patricia A. Carpenter. 1977. A Comparison of Aphasic and Normal Adults in a Sentence-Verification Task. *Cortex* 13, 4 (1977), 402–423. https://doi.org/10.1016/S0010-9452(77)80021-X

[168] Shaun Kaasten and Saul Greenberg. 2002. How People Recognize Previously Seen Web Pages from Titles, URLs and Thumbnails. In *Proceedings of Human Computer Interaction 2002 - People and Computers XVI - Memorable Yet Visible*. London, England, UK, 247–265. https://doi.org/10.1007/978-1-4471-0105-5_15

[169] Leander Kahney. 2002. HyperCard Forgotten, but Not Gone. *Wired* (2002). https://web.archive.org/web/20100206164413/http://www.wired.com/gadgets/mac/commentary/cultofmac/2002/08/54365

[170] P. Kalva, F. Enembreck, and A. Koerich. 2007. WEB Image Classification Based on the Fusion of Image and Text Classifiers. In *Ninth International Conference on Document Analysis and Recognition*, Vol. 1. 561–568. https://doi.org/10.1109/ICDAR.2007.4378772

[171] Ho Tin Kam. 1995. Random decision forest. In *Proceedings of the 3rd international conference on document analysis and recognition*, Vol. 1416. Montreal, Canada, August, 278282. https://doi.org/10.1109/ICDAR.1995.598994

[172] Min-Yen Kan and Hoang Oanh Nguyen Thi. 2005. Fast Webpage Classification Using URL Features. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. Bremen, Germany, 325–326. https://doi.org/10.1145/1099554.1099649

[173] Atsushi Kanehira, Luc Van Gool, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Viewpoint-Aware Video Summarization. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA. https://openaccess.thecvf.com/content_cvpr_2018/html/Kanehira_Viewpoint-Aware_Video_Summarization_CVPR_2018_paper.html

[174] Nattiya Kanhabua, Philipp Kemkes, Wolfgang Nejdl, Tu Ngoc Nguyen, Felipe Reis, and Nam Khanh Tran. 2016. How to Search the Internet Archive Without Indexing It. In *Proceedings of the 20th International Conference on Theory and Practice of Digital Libraries*. Hannover, Germany, 147–160. https://doi.org/10.1007/978-3-319-43997-6_12

[175] Epaminondas Kapetanios, Doina Tatar, and Christian Sacarea. 2014. *Natural Language Processing: Semantic Aspects*. Taylor & Francis, Boca Raton, Florida, USA. https://books.google.com/books?id=YXv6AQAAQBAJ

[176] Weimao Ke, Javed Mostafa, and Yong Liu. 2008. Toward responsive visualization services for scatter/gather browsing. *Proceedings of the American Society for Information Science and Technology* 45, 1 (2008), 1–10. https://doi.org/10.1002/meet.2008.1450450269

[177] Diane Kelly. 2007. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval* 3, 1-2 (2007), 1–224. https://doi.org/10.1561/1500000012

[178] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and Evaluation of Search Tasks for IIR Experiments using a Cognitive Complexity Framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. Northampton, Massachusetts, USA, 101–110. https://doi.org/10.1145/2808194.2809465

[179] Adam Kilgarriff. 1997. Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora. In *Association for Computational Linguistics Fifth Workshop on Very Large Corpora*. https://aclanthology.org/W97-0122/

[180] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies With Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 453–456. https://doi.org/10.1145/1357054.1357127

[181] Martin Klein, Lyudmila Balakireva, and Herbert Van de Sompel. 2018. Focused Crawl of Web Archives to Build Event Collections. In *Proceedings of the 10th ACM Conference*

on Web Science (Amsterdam, Netherlands). New York, NY, USA, 333–342. https://doi.org/10.1145/3201064.3201085

[182] Martin Klein, Jeffery Shipman, and Michael L. Nelson. 2010. Is This a Good Title?. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*. Toronto, Ontario, Canada, 3–12. https://doi.org/10.1145/1810617.1810621

[183] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE* 9, 12 (2014), e115253. https://doi.org/10.1371/journal.pone.0115253

[184] Eunyee Koh and Andruid Kerne. 2009. Deriving image-text document surrogates to optimize cognition. In *Proceedings of the 9th ACM Symposium on Document Engineering*. 84–93. https://doi.org/10.1145/1600193.1600212

[185] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. New York, New York, USA, 441–450. https://doi.org/10.1145/1718487.1718542

[186] Theodorich Kopetzky and Max Mühlhäuser. 1999. Visual preview for link traversal on the World Wide Web. *Computer Networks* 31, 11-16 (1999), 1525–1532. https://doi.org/10.1016/S1389-1286(99)00050-X

[187] Jesse Kornblum. 2006. Identifying almost identical files using context triggered piecewise hashing. *Digital Investigation* 3 (2006), 91–97. https://doi.org/10.1016/j.diin.2006.06.015 The Proceedings of the 6th Annual Digital Forensic Research Workshop.

[188] Kalin Kozhuharov. 2020. [6.9.11] CorruptImageProfile ‘xmp’ @ warning/profile.c/SetImageProfileInternal/1701 · Issue #110 · ImageMagick/ImageMagick6. https://github.com/ImageMagick/ImageMagick6/issues/110 Retrieved 5-February-2021.

[189] Neal Krawetz. 2011. Looks Like It - The Hacker Factor Blog. http://hackerfactor.com/blog/index.php%3F/archives/432-Looks-Like-It.html Retrieved 4-February-2021.

[190] Robert Laganière, Raphael Bacco, Arnaud Hocevar, Patrick Lambert, Grégory Païs, and Bogdan E. Ionescu. 2008. Video Summarization from Spatio-Temporal Features. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop* (Vancouver, British Columbia, Canada). New York, NY, USA, 144–148. https://doi.org/10.1145/1463563.1463590

[191] Cheryl Lederle. 2016. Your Students Can Archive the Internet — Apply Now. https://blogs.loc.gov/teachers/2016/07/your-students-can-archive-the-internet-apply-now/ Retrieved 3-May-2018.

[192] V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics – Doklady* 10, 8 (1966), 707–709.

[193] Noah Levitt, Alex Osborne, Andy Jackson, and Kristinn Sigurðsson. 2019. Home - internetarchive/heritrix3 Wiki. https://github.com/internetarchive/heritrix3/wiki Retrieved 14-April-2019.

[194] Peng Li, Jing Jiang, and Yinglin Wang. 2010. Generating templates of entity summaries with an entity-aspect model and pattern mining. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, 640–649. https://dl.acm.org/citation.cfm?id=1858747

[195] Peng Li, Yinglin Wang, Wei Gao, and Jing Jiang. 2011. Generating Aspect-oriented Multi-document Summarization with Event-aspect Model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK, 1137–1146. http://dl.acm.org/citation.cfm?id=2145432.2145553

[196] Ying Li, Zijian Zheng, and Honghua (Kathy) Dai. 2005. KDD CUP-2005 report: facing a great challenge. *ACM SIGKDD Explorations Newsletter* 7, 2 (2005), 91–99. https://doi.org/10.1145/1117454.1117466

[197] Zhiwei Li, Shuming Shi, and Lei Zhang. 2008. Improving Relevance Judgment of Web Search Results with Image Excerpts. In *Proceedings of the 17th International Conference on World Wide Web*. Beijing, China, 21–30. https://doi.org/10.1145/1367497.1367501

[198] Jefrey Lijffijt, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and

Heikki Mannila. 2014. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities* 31, 2 (2014), 374–397. https://doi.org/10.1093/llc/fqu064

[199] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Association for Computational Linguistics 2004 Workshop Text Summarization Branches Out*. Barcelona, Spain, 74–81. http://www.aclweb.org/anthology/W04-1013

[200] Jimmy Lin, Ian Milligan, Jeremy Wiebe, and Alice Zhou. 2017. Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. *Journal on Computing and Cultural Heritage* 10, 4 (2017), 1–30. https://doi.org/10.1145/3097570

[201] Mu Lin. 2015. Teaching "old" skills in a new way: How a journalism instructor integrates Storify in news writing classes. http://www.mulinblog.com/learning-old-skills-in-a-new-way-how-a-journalism-instructor-integrates-storify-in-news-writing-classes/ Retrieved 25-May-2021.

[202] Tzong-Jye Liu, Cheng-Nan Wu, Chia-Lin Lee, and Ching-Wen Chen. 2014. A self-adaptable image spam filtering system. *Journal of the Chinese Institute of Engineers* 37, 4 (2014), 517–528. https://doi.org/10.1080/02533839.2013.815005 https://doi.org/10.1080/02533839.2013.815005.

[203] Wei-Yin Loh. 2011. Classification and regression trees. *WIREs Data Mining and Knowledge Discovery* 1, 1 (2011), 14–23. https://doi.org/10.1002/widm.8

[204] Faidon Loumakis, Simone Stumpf, and David Grayson. 2011. This Image Smells Good: Effects of Image Information Scent in Search Engine Results Pages. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. Glasgow, Scotland, UK, 475–484. https://doi.org/10.1145/2063576.2063649

[205] Wenlong Lu, Luca Pagani, Liping Zhou, Xiaojun Liu, Jian Wang, Richard Leach, and Xiangqian (Jane) Jiang. 2019. Uncertainty-guided intelligent sampling strategy for high-efficiency surface measurement via free-knot B-spline regression modelling. *Precision Engineering* 56 (2019), 38–52. https://doi.org/10.1016/j.precisioneng.2018.09.002

[206] H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* 2, 2 (1958), 159–165. https://doi.org/10.1147/rd.22.0159

[207] Abigail Mabe, Dhruv Patel, Maheedhar Gunnam, Surbhi Shankar, Mat Kelly, Sawood Alam, Michael L. Nelson, and Michele C. Weigle. 2020. *Visualizing Webpage Changes Over Time*. Technical Report 2006.02487. Old Dominion University. arXiv:2006.02487 [cs.DL] https://arxiv.org/abs/2006.02487

[208] Takuya Maekawa, Takahiro Hara, and Shojiro Nishio. 2006. Image classification for mobile web browsing. In *Proceedings of the 15th international conference on World Wide Web*. Edinburgh, Scotland, UK, 43–52. https://doi.org/10.1145/1135777.1135789

[209] Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*. Banff, Alberta, Canada, 141–150. https://doi.org/10.1145/1242572.1242592

[210] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010

[211] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46. https://doi.org/10.1145/1121949.1121979

[212] Ashwin Maroli, Frank Taillandier, Matt Rogers, Alfred Xing, Nick Quaranto, Parker Moore, and Tom Preston-Werner. 2021. Jekyll • Simple, blog-aware, static sites. https://jekyllrb.com/ Retrieved 13-June-2021.

[213] Julien Masanes. 2006. *Web Archiving*. Springer International Publishing, New York City, New York, USA.

[214] Pranay Mathur, Aman Gill, and Aayush Yadav. 2017. Text Summarization in Python: Extractive vs. Abstractive techniques revisited. https://rare-technologies.com/text-summarization-in-python-extractive-vs-abstractive-techniques-revisited/ Retrieved 24-April-2018.

[215] H. Maurer. 1994. Hyper-G: Advancing the Ideas of the World-Wide-Web. http://kirste.userpage.fu-berlin.de/outerspace/doc/hyper-g-abs.html Retrieved 12-July-2021.

[216] David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. 2017. A Study of Snippet Length and Informativeness: Behaviour, Performance and User Experience. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Shinjuku, Tokyo, Japan, 135–144. https://doi.org/10.1145/3077136.3080824

[217] Marji McClure. 2006. Archive-It 2: Internet Archive Strives to Ensure Preservation and Accessibility. *EContent Magazine* (2006). https://web.archive.org/web/20121222160742/http://www.econtentmag.com/Articles/News/News-Feature/Archive-It-2-Internet-Archive-Strives-to-Ensure-Preservation-and-Accessibility-18132.htm

[218] Frank McCown, Sheffan Chan, Michael L. Nelson, and Johan Bollen. 2005. The Availability and Persistence of Web References in D-Lib Magazine. In *Proceedings of the 5th International Web Archiving Workshop*. Vienna, Austria. https://arxiv.org/abs/cs/0511077

[219] Shaohui Mei, Genliang Guan, Zhiyong Wang, Shuai Wan, Mingyi He, and David Dagan Feng. 2015. Video summarization via minimum sparse reconstruction. *Pattern Recognition* 48, 2 (2015), 522–533. https://doi.org/10.1016/j.patcog.2014.08.002

[220] D. Meyer. 1973. RFC 543 - Network Journal Submission and Delivery. https://tools.ietf.org/html/rfc543

[221] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain, 404–411. http://aclweb.org/anthology/W04-3252

[222] Ian Milligan. 2018. History is Always About Context (or why a history degree equips you to understand the context of a tweet). https://ianmilli.wordpress.com/2018/08/09/history-is-always-about-context-or-why-a-history-degree-equips-you-to-understand-the-context-of-a-tweet/

[223] MIT Libraries. 2018. MIT Libraries Institute Archives and Special Collections Web Archiving Metadata Application Profile. https://github.com/jfcarrano/

archives-docs/blob/master/mit_webarch_metaprofile.md Retrieved 25-May-2021.

[224] P. Mockapetris. 1983. RFC 882 - Domain Names - Concepts and Facilities. https://tools.ietf.org/html/rfc882

[225] Paul Mockapetris. 1983. RFC 883 - Domain Names - Implementation and Specification. https://tools.ietf.org/html/rfc883

[226] Paul Mockapetris. 1986. RFC 973 - Domain System Changes and Observations. https://tools.ietf.org/html/rfc973

[227] Julie B Morrison, Peter Pirolli, and Stuart K Card. 2001. A Taxonomic Analysis of What World Wide Web Activities Significantly Impact People's Decisions and Actions. In *Conference on Human Factors in Computing Systems 2001 Extended Abstracts on Human Factors in Computing Systems*. Seattle, Washington, USA, 163–164. https://doi.org/10.1145/634067.634167

[228] Rajat Mukherjee and Jianchang Mao. 2004. Enterprise Search: Tough Stuff: Why is It That Searching an Intranet is so Much Harder than Searching the Web? *Queue* 2, 2 (2004), 36–46. https://doi.org/10.1145/988392.988406

[229] Padmavathi Mundur, Yong Rao, and Yelena Yesha. 2006. Keyframe-based video summarization using Delaunay clustering. *International Journal on Digital Libraries* 6 (2006), 219–232. https://doi.org/10.1007/s00799-005-0129-9

[230] Eric Muntz. 2014. Social Cards, Segmentation Options, New Currencies, and More in v9.4. https://web.archive.org/web/20170204072827/https://blog.mailchimp.com/social-cards-segmentation-options-new-currencies-and-more-in-v9-4/

[231] Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. *Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond*. Technical Report 1602.06023. IBM. http://arxiv.org/abs/1602.06023

[232] Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. 2018. Entity-Aspect Linking: Providing Fine-Grained Semantics of Entities in Context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. Fort Worth, Texas, USA, 49–58. https://doi.org/10.1145/3197026.3197047

[233] National Institute of Standards and Technology. 2017. TAC 2010 Guided Summarization Task Guidelines. https://tac.nist.gov/2010/Summarization/Guided-Summ.2010.guidelines.html Retrieved 14-April-2019.

[234] Michael L. Nelson. 2021. Improving the Dark and Stormy Archives Framework by Summarizing the Collections of the National Library of Australia. https://netpreserve.org/projects/dark-and-stormy-archives/ Retrieved 25-May-2021.

[235] Michael L. Nelson and Michele C. Weigle. 2012. Why Care About the Past? https://www.slideshare.net/phonedude/why-careaboutthepast Retrieved 4-July-2021.

[236] Rodrigo Nogueira and Jimmy Lin. 2019. *From doc2query to docTTTTTquery*. Technical Report. University of Waterloo. https://cs.uwaterloo.ca/~jimmylin/publications/Nogueira_Lin_2019_docTTTTTquery.pdf

[237] Alexander Nwala. 2018. An exploration of URL diversity measures. https://ws-dl.blogspot.com/2018/05/2018-05-04-exploration-of-url-diversity.html Retrieved 4-November-2018.

[238] Alexander Nwala. 2019. Introducing sumgram, a tool for generating the most frequent conjoined ngrams. https://ws-dl.blogspot.com/2019/09/2019-09-09-introducing-sumgram-tool-for.html Retrieved 25-May-2021.

[239] Alexander Nwala. 2020. 365 dots in 2018 - top news stories of 2018. https://ws-dl.blogspot.com/2019/03/2019-03-05-365-dots-in-2018-top-news.html Retrieved 25-May-2021.

[240] Alexander Nwala. 2020. 365 dots in 2019 - top news stories of 2019. https://ws-dl.blogspot.com/2020/01/2020-01-04-365-dots-in-2019-top-news.html Retrieved 25-May-2021.

[241] Alexander Nwala. 2020. 365 Dots in 2019: Quantifying Attention of News Sources. In *Proceedings of the Computation + Journalism Symposium 2020*. Northeastern University, Boston, MA, USA. https://cpb-us-w2.wpmucdn.com/sites.northeastern.edu/dist/d/53/files/2020/02/CJ_2020_paper_31.pdf

[242] Alexander Nwala. 2021. 366 dots in 2020 - top news stories of 2020. https://ws-dl.blogspot.com/2021/01/2020-01-20-366-dots-in-2020-top-news.html Retrieved 25-May-2021.

[243] Alexander C. Nwala. 2015. What Did It Look Like? https://ws-dl.blogspot.com/2015/01/2015-02-05-what-did-it-look-like.html Retrieved 25-May-2021.

[244] Alexander C. Nwala. 2017. A survey of 5 boilerplate removal methods. http://ws-dl.blogspot.com/2017/03/2017-03-20-survey-of-5-boilerplate.html Retrieved 4-November-2018.

[245] Alexander C. Nwala, Michele C. Weigle, and Michael L. Nelson. 2018. Bootstrapping Web Archive Collections from Social Media. In *Proceedings of the 29th on Hypertext and Social Media*. Baltimore, Maryland, USA, 64–72. https://doi.org/10.1145/3209542.3209560

[246] Alexander C. Nwala, Michele C. Weigle, and Michael L. Nelson. 2018. Scraping SERPs for Archival Seeds: It Matters When You Start. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. Fort Worth, Texas, USA, 263–272. https://doi.org/10.1145/3197026.3197056

[247] Alexander C. Nwala, Michele C. Weigle, and Michael L. Nelson. 2020. *365 Dots in 2019: Quantifying Attention of News Sources*. Technical Report 2003.09989. Old Dominion University. https://arxiv.org/abs/2003.09989

[248] Information Sciences Institute University of Southern California. 1981. RFC 791 - Internet Protocol. https://tools.ietf.org/html/rfc791

[249] Jessica Ogden, Susan Halford, and Leslie Carr. 2017. Observing Web Archives: The Case for an Ethnographic Study of Web Archiving. In *Proceedings of the 2017 ACM on Web Science Conference (WebSci '17)*. Association for Computing Machinery, Troy, New York, USA, 299–308. https://doi.org/10.1145/3091478.3091506

[250] Tom O'Hara and Radim Řehůřek. 2015. making LSI reproducible across machines. https://groups.google.com/forum/#!topic/gensim/upiK51Hs_Pc Retrieved 2-February-2018.

[251] Maile Ohye. 2007. Google, duplicate content caused by URL parameters, and you. https://developers.google.com/search/blog/2007/09/google-duplicate-content-caused-by-url Retrieved 4-July-2021.

[252] David Orchard. 1995. Hyper-G. In *Proceedings of the Third International WWW Conference Demonstration and Paper Session*. Darmstadt, Germany. https://web.archive.org/web/19970614013822/http://www.pacificspirit.com/wwwConfNotes/hyperg.htm

[253] Tim Oren. 2004. A Eulogy for HyperCard. https://due-diligence.typepad.com/blog/2004/03/a_eulogy_for_hy.html Retrieved 14-July-2019.

[254] Our Hometown, Inc. 2018. Facebook & Twitter Social Media Cards. https://our-hometown.com/social-media-embedded-features/facebook-twitter-social-cards/ Retrieved 3-May-2018.

[255] Y. E. Ozkose, B. Celikkale, E. Erdem, and A. Erdem. 2019. Diverse Neural Photo Album Summarization. In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications*. 1–6. https://doi.org/10.1109/IPTA.2019.8936084

[256] Kalpesh Padia. 2012. *Visualizing digital collections at Archive-It.* Master's thesis. Old Dominion University, Norfolk, VA, USA. http://doi.org/10.25777/psw9-4x42

[257] Kalpesh Padia, Yasmin AlNoamany, and Michele C. Weigle. 2012. Visualizing digital collections at Archive-It. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. Washington, DC, USA, 15–18. https://doi.org/10.1145/2232817.2232821

[258] Seung-Taek Park, David M. Pennock, C. Lee Giles, and Robert Krovetz. 2002. Analysis of Lexical Signatures for Finding Lost or Related Documents. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Tampere, Finland). New York, NY, USA, 11–18. https://doi.org/10.1145/564376.564381

[259] Krutarth Patel, Cornelia Caragea, Mark E Phillips, and Nathaniel T Fox. 2020. Identifying Documents In-Scope of a Collection from Web Archives. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. 167–176. https://doi.org/10.1145/3383583.3398540

[260] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2/1/2011 (2011), 2825–2830. http://dl.acm.org/citation.cfm?id=1953048.2078195

[261] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2/1/2011 (2011), 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

[262] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70, May 2017 (2017), 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

[263] Yuxin Peng and Chong-Wah Ngo. 2006. Clip-based similarity measure for query-dependent clip retrieval and video summarization. *IEEE Transactions on Circuits and Systems for Video Technology* 16, 5 (2006), 612–627. https://doi.org/10.1109/TCSVT.2006.873157

[264] Matt Peters. 2017. Benchmarking Python Content Extraction Algorithms: Dragnet, Readability, Goose, and Eatiht - Moz. https://web.archive.org/web/20170518150333/https://moz.com/devblog/benchmarking-python-content-extraction-algorithms-dragnet-readability-goose-and-eatiht/

[265] Thomas A. Phelps and Robert Wilensky. 2004. Robust Hyperlinks: Cheap, Everywhere, Now. In *Digital Documents: Systems and Principles*. Berlin, Heidelberg, 28–43. https://doi.org/10.1007/978-3-540-39916-2_3

[266] Peter Pirolli and Stuart Card. 1999. Information Foraging. *Psychological Review* 106, 4 (1999), 643–675. https://doi.org/10.1037/0033-295X.106.4.643

[267] Peter Pirolli and Stuart K. Card. 1998. Information Foraging Models of Browsers for Very Large Document Spaces. In *Proceedings of the Working Conference on Advanced Visual Interfaces*. L'Aquila, Italy, 83—-93. https://doi.org/10.1145/948496.948509

[268] Daniel V. Pitti. 1999. Encoded Archival Description. *D-Lib Magazine* 5, 11 (1999). https://doi.org/10.1045/november99-pitti

[269] Jan Pomikalek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. Dissertation. Masaryk University, Brno, Czechia. https://is.muni.cz/th/45523/fi_d/phdthesis.pdf

[270] J. Postel and J. Reynolds. 1985. RFC 959 - File Transfer Protocol (FTP). https://tools.ietf.org/html/rfc959

[271] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. 2014. Category-Specific Video Summarization. In *Computer Vision – ECCV 2014*. Springer International Publishing, Cham, 540–555. https://doi.org/10.1007/978-3-319-10599-4_35

[272] Rojal Pradhan. 1997. Harmony – The Unix/X11 Client for Hyper-G. https://web.archive.org/web/19970414204427/http://vertex.cs.bsu.edu/hyperg.html

[273] Maria Praetzellis. 2016. Archive-It 4.9 Release Notes. https://support.archive-it.org/hc/en-us/articles/208110956-Archive-It-4-9-Release-Notes

[274] J. Ren, X. Shen, Z. Lin, and R. Měch. 2020. Best Frame Selection in a Short Video. In *2020 IEEE Winter Conference on Applications of Computer Vision*. Snowmass, CO, USA, 3201–3210. https://doi.org/10.1109/WACV45572.2020.9093615

[275] E. Rescorla. 2000. RFC 2818 - HTTP Over TLS. https://tools.ietf.org/html/rfc2818

[276] Rhizome. 2017. Our First Social Media President, Through Three Webrecorder Narratives: A Rhizome x Obama White House Collab. http://rhizome.org/editorial/2017/jan/05/the-first-social-media-president/

[277] Leonard Richardson. 2017. Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/ Retrieved 31-January-2018.

[278] Ian Ritchie. 2011. The day I turned down Tim Berners-Lee. https://www.ted.com/talks/ian_ritchie_the_day_i_turned_down_tim_berners_lee 14-July-2019.

[279] George Robertson, Mary Czerwinski, Kevin Larson, Daniel C. Robbins, David Thiel, and Maarten van Dantzich. 1998. Data mountain: using spatial memory for document management. In *Proceedings of the 11th annual ACM Symposium on User Interface Software and Technology*. San Francisco, California, USA, 153–162. https://doi.org/10.1145/288392.288596

[280] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC 1994)*. http://trec.nist.gov/pubs/trec3/papers/city.ps.gz

[281] Luiz Cláudio Santos Silva and Renelson Ribeiro Sampaio. 2014. Using Luhn's Automatic Abstract Method to Create Graphs of Words for Document Visualization. *Social Networking* 3, 2 (2014), 65–70. https://doi.org/10.4236/sn.2014.32008

[282] Christine Schmidt. 2017. Storify's demise shows nothing lasts forever (but the use of social media embeds in stories persists). https://www.niemanlab.org/2017/12/storifys-demise-shows-nothing-lasts-forever-but-the-use-of-social-media-embeds-in-stories-persists/ Retrieved 25-May-2021.

[283] Ian Sherr. 2016. Adobe buys Livefyre to turn your awful Internet comments into money. https://www.cnet.com/news/adobe-buys-livefyre-to-turn-your-awful-internet-comments-into-money/ Retrieved 25-May-2021.

[284] Nakatani Shuyo. 2014. Language Detection. https://github.com/shuyo/language-detection Retrieved 5-April-2018.

[285] similarweb. 2021. Tumblr.com Traffic Ranking & Marketing Analytics. https://www.similarweb.com/website/tumblr.com/#overview Retrieved 25-May-2021.

[286] Ruben Sipos, Adith Swaminathan, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Temporal corpus summarization using submodular word coverage. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. Maui, Hawaii, USA, 754–763. https://doi.org/10.1145/2396761.2396857

[287] Michael C. Sloan. 2010. Aristotle's *Nicomachean Ethics* as the Original *Locus* for the *Septem Circumstantiae*. *Classical Philology* 105, 3 (2010), 236–251. https://doi.org/10.1086/656196

[288] Snapchat. 2021. Snapchat - The fastest way to share a moment. https://www.snapchat.com/ Retrieved 25-May-2021.

[289] Storify. 2017. Storify End-of-Life. https://storify.com/faq-eol Retrieved 3-May-2018.

[290] Thorvald Julius Sørensen. 1948. A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons. *Kongelige Danske Videnskabernes Selskab* 5, 4 (1948), 1–34. http://www.royalacademy.dk/Publications/High/295_S%C3%B8rensen,%20Thorvald.pdf

[291] L. Tan, Y. Song, S. Liu, and L. Xie. 2012. ImageHive: Interactive Content-Aware Image Summarization. *IEEE Computer Graphics and Applications* 32, 1 (2012), 46–55. https://doi.org/10.1109/MCG.2011.89

[292] Zhenjun Tang, Yumin Dai, and Xianquan Zhang. 2012. Perceptual Hashing for Color Images Using Invariant Moments. *Applied Mathematics & Information Sciences* 6, 2S (2012), 8. http://www.naturalspublishing.com/files/published/54515x71g3omq1.pdf

[293] Jaime Teevan, Edward Cutrell, Danyel Fisher, Steven M. Drucker, Gonzalo Ramos, Paul André, and Chang Hu. 2009. Visual snippets: summarizing web pages for search and revisitation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Boston, Massachusetts, USA, 2023–2032. https://doi.org/10.1145/1518701.1519008

[294] Mallary Jean Tenore. 2010. How to use Storify as a reporting tool. https://www.poynter.org/reporting-editing/2010/how-to-use-storify-as-a-reporting-tool/ Retrieved 25-May-2021.

[295] Mallary Jean Tenore. 2011. 25 ways to use Facebook, Twitter & Storify to improve political coverage. https://www.poynter.org/reporting-editing/2011/25-ways-to-use-facebook-twitter-storify-to-improve-election-coverage/ Retrieved 25-May-2021.

[296] Mallary Jean Tenore. 2011. The 5 types of stories that make good Storifys. https://www.poynter.org/reporting-editing/2011/the-5-types-of-stories-that-make-good-storifys/ Retrieved 25-May-2021.

[297] The Archive-It Team. 2020. Archive-It and Archives Unleashed join forces to scale research use of web archives. https://archive-it.org/blog/post/archives-unleashed-partnership/ Retrieved 12-July-2021.

[298] The ImageMagick Development Team. 2021. ImageMagick. https://imagemagick.org Retrieved 5-February-2021.

[299] Masashi Toyoda and Masaru Kitsuregawa. 2012. The History of Web Archiving. *Proc. IEEE* 100, Special Centennial Issue (2012), 1441–1443.

[300] Myriam C. Traub, Thaer Samar, Jacco van Ossenbruggen, Jiyin He, Arjen de Vries, and Lynda Hardman. 2016. Querylog-Based Assessment of Retrievability Bias in a Large Newspaper Corpus. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. Newark, New Jersey, USA, 7–16. https://doi.org/10.1145/2910896.2910907

[301] Sebastian Tschiatschek, Rishabh Iyer, Haochen Wei, and Jeff Bilmes. 2014. Learning mixtures of submodular functions for image collection summarization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*. Montreal, Canada, 1413–1421.

[302] Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining (IJDWM)* 3, 3 (2007), 1–13. https://doi.org/10.4018/jdwm.2007070101

[303] Edward R. Tufte. 2001. *The visual display of quantitative information* (second ed.). Graphics Press, Cheshire, Connecticut, USA.

[304] Twitter. 2018. Optimize Tweets with Cards. https://developer.twitter.com/en/docs/tweets/optimize-with-cards/guides/getting-started Retrieved 3-November-2018.

[305] Twitter. 2021. About Twitter Fleets - what are they and how to delete. https://help.twitter.com/en/using-twitter/fleets Retrieved 25-May-2021.

[306] Twitter. 2021. Cards markup. https://developer.twitter.com/en/docs/twitter-for-websites/cards/overview/markup Retrieved 4-February-2021.

[307] Twitter. 2021. Why should your business use Twitter? https://business.twitter.com/en/blog/why-should-your-business-use-twitter.html Retrieved 25-May-2021.

[308] University of Texas at Austin School of Information. 2011. Introduction to EAD. https://tutorials.ischool.utexas.edu/index.php/Introduction_to_EAD Retrieved 25-May-2021.

[309] University of Virginia Library. 2018. University of Virginia Library Web Archiving Metadata Application Profile. https://docs.google.com/document/d/1M5kTUtUjob7YB7MpEd_Jl5lRsuZprxNZQZ6rELmJjeE/edit#heading=h.pz8x0vs3t9zo Retrieved 13-April-2019.

[310] Pertti Vakkari. 2005. Task-based information searching. *Annual Review of Information Science and Technology* 37, 1 (2005), 413–464. https://doi.org/10.1002/aris.1440370110

[311] Ellen van Aken. 2014. Why enterprise search is NOT like Google search. https://mydigitalworkplace.wordpress.com/2014/10/29/enterprise-vs-google-search/ Retrieved 4-July-2021.

[312] Herbert Van de Sompel, Michael Nelson, and Robert Sanderson. 2013. RFC 7089 - HTTP Framework for Time-Based Access to Resource States – Memento. https://tools.ietf.org/html/rfc7089

[313] Anne van Kesteren, Aryeh Gregor, Ms2ger, Alex Russell, and Robin Berjon. 2015. W3C DOM4. https://www.w3.org/TR/2015/REC-dom-20151119/ Retrieved 14-April-2019.

[314] C. J. Van Rijsbergen. 1979. *Information Retrieval* (second ed.). Butterworth-Heinemann, Oxford, England, UK.

[315] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond

SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management* 43, 6 (2007), 1606–1618. https://doi.org/10.1016/j.ipm.2007.01.023

[316] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[317] Fred Weinhaus. 2014. Tests Of Perceptual Hash (PHASH) Compare Metric. http://www.fmwconcepts.com/misc_tests/perceptual_hash_test_results_510/index.html Retrieved 4-February-2021.

[318] Gregory Wiedeman. 2019. Describing Web Archives: A Computer-Assisted Approach. *Journal of Contemporary Archival Studies* 6, 31 (2019). https://elischolar.library.yale.edu/jcas/vol6/iss1/31

[319] Wikibooks. 2018. HyperText Markup Language — Wikibooks, The Free Textbook Project. https://en.wikibooks.org/w/index.php?title=HyperText_Markup_Language&oldid=3453727

[320] Wikimedia Commons. 2015. File:HarvardCollegeLibrary CatalogCardSwi.jpg — Wikimedia Commons, the free media repository. https://commons.wikimedia.org/w/index.php?title=File:HarvardCollegeLibrary_CatalogCardSwi.jpg&oldid=144924593 Retrieved 12-July-2021.

[321] Wikimedia Commons. 2020. File:Sample Kanban Board.png — Wikimedia Commons, the free media repository. https://commons.wikimedia.org/w/index.php?title=File:Sample_Kanban_Board.png&oldid=445344614 Retrieved 4-July-2021.

[322] Wikipedia contributors. 2019. Marine mammal — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Marine_mammal&oldid=906443640 Retrieved 14-July-2021.

[323] Wikipedia contributors. 2021. 2016 Louisiana floods — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=2016_Louisiana_floods&oldid=1022074499 Retrieved 10-June-2021.

[324] Wikipedia contributors. 2021. Decision tree learning — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Decision_tree_learning&oldid=1032085462 Retrieved 7-July-2021.

[325] Wikipedia contributors. 2021. Encoded Archival Description — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Encoded_Archival_Description&oldid=998551556 Retrieved 18-June-2021.

[326] Wikipedia contributors. 2021. Small multiple — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Small_multiple&oldid=894091657 Retrieved 14-July-2021.

[327] Wikipedia contributors. 2021. Storytelling — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Storytelling&oldid=906157698 Retrieved 14-July-2021.

[328] Alex Wilhelm. 2012. Storify users posted 554,000 stories in 2012 which were viewed 367 million times. https://thenextweb.com/news/storify-users-in-2012-posted-554000-stories-which-were-viewed-367-million-times Retrieved 25-May-2021.

[329] Wired Staff. 2004. The Click Heard Round The World. *Wired* (2004). https://www.wired.com/2004/01/mouse/

[330] Gary Wolf. 1995. The Curse of Xanadu. *Wired* (1995). https://www.wired.com/1995/06/xanadu/

[331] Allison Woodruff, Andrew Faulring, Ruth Rosenholtz, Julie Morrsion, and Peter Pirolli. 2001. Using thumbnails to search the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Seattle, Washington, USA, 198–205. https://doi.org/10.1145/365024.365098

[332] Allison Woodruff, Ruth Rosenholtz, Julie B. Morrison, Andrew Faulring, and Peter Pirolli. 2002. A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for Web search tasks. *Journal of the American Society for Information Science and Technology* 53, 2 (2002), 172–185. https://doi.org/10.1002/asi.10029

[333] Shasha Xie and Yang Liu. 2008. Using corpus and knowledge-based similarity measure in Maximum Marginal Relevance for meeting summarization. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. Las Vegas, Nevada, USA, 4985–4988. https://doi.org/10.1109/ICASSP.2008.4518777

[334] Seungwon Yang, Kiran Chitturi, Gregory Wilson, Mohamed Magdy, and Edward A. Fox. 2012. A Study of Automation from Seed URL Generation to Focused Web Archive Development: The CTRnet Context. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*. Washington, DC, USA, 341–342. https://doi.org/10.1145/2232817.2232881

[335] Seungwon Yang, Andrea Kavanaugh, Nádia P. Kozievitch, Lin Tzy Li, Venkat Srinivasan, Steven D. Sheetz, Travis Whalen, Donald Shoemaker, Ricardo da S Torres, and Edward A. Fox. 2011. CTRnet DL for Disaster Information Services. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. Ottawa, Ontario, Canada, 437–438. https://doi.org/10.1145/1998076.1998173

[336] Ting Yao, Tao Mei, and Yong Rui. 2016. Highlight Detection With Pairwise Deep Ranking for First-Person Video Summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 982–990. https://openaccess.thecvf.com/content_cvpr_2016/html/Yao_Highlight_Detection_With_CVPR_2016_paper.html

[337] Christoph Zauner. 2010. *Implementation and Benchmarking of Perceptual Image Hash Functions*. Master's thesis. Upper Austria University of Applied Sciences. http://www.phash.org/docs/pubs/thesis_zauner.pdf

[338] Renxian Zhang, Wenjie Li, and Dehong Gao. 2012. Generating Coherent Summaries with Textual Aspects. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. Toronto, Ontario, Canada, 1727–1733. http://dl.acm.org/citation.cfm?id=2900929.2900973

[339] Yan Zhang, Ramona Broussard, Weimao Ke, and Xuemei Gong. 2014. Evaluation of a scatter/gather interface for supporting distinct health information search tasks. *Journal of the Association for Information Science and Technology* 65, 5 (2014), 1028–1041. https://doi.org/10.1002/asi.23011 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23011

[340] Yueting Zhuang, Yong Rui, T.S. Huang, and S. Mehrotra. 1998. Adaptive key frame extraction using unsupervised clustering. In *Proceedings 1998 International Conference on Image Processing. (Cat. No.98CB36269)*, Vol. 1. 866–870 vol.1. https://doi.org/10.1109/ICIP.1998.723655

[341] Radim Řehůřek. 2011. *Scalability of Semantic Analysis in Natural Language Processing*. Ph.D. Dissertation. Masaryk University, Brno, Czechia. https://radimrehurek.com/phd_rehurek.pdf

[342] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*. Valletta, Malta, 46–50. https://is.muni.cz/publication/884893/en

VITA

Shawn M. Jones
Department of Computer Science
Old Dominion University
Norfolk, VA 23529

e-mail: `jones.shawn.m@gmail.com`

Education

Doctor of Philosophy in Computer Science (2021)
Old Dominion University, Norfolk, Virginia, USA
Dissertation: *Improving Collection Understanding For Web Archives With Storytelling: Shining Light Into Dark and Stormy Archives*

Master of Science in Computer Science (2015)
Old Dominion University, Norfolk, Virginia, USA
Thesis: *Avoiding Spoilers on Mediawiki Fan Sites Using Memento*

Bachelor of Science in Computer Science (1999)
Old Dominion University, Norfolk, Virginia, USA

Employment

2015 - present – Graduate Research Assistant
Los Alamos National Laboratory Research Library

2013 - 2019 – Graduate Research Assistant
Old Dominion University Web Science and Digital Libraries Research Group

1998 - 2015 – Computer Scientist
Space and Naval Warfare Systems Center, Norfolk (SPAWAR)

Publications

An updated list of publications is available at `https://orcid.org/0000-0002-4372-870X`

Typeset using LaTeX.