

LA-UR-20-29413

Approved for public release; distribution is unlimited.

Title:	Summarizing web archives through storytelling with the Dark and Stormy Archives Project
Author(s):	Jones, Shawn Morgan
Intended for:	Coalition for Networked Information Fall 2020 Meeting, 2020-11-10/2020-11-13 (Washington, District Of Columbia, United States)
Issued:	2020-11-17

Disclaimer:

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Summarizing web archives through storytelling with the Dark and Stormy Archives Project

Shawn M. Jones

Los Alamos National Laboratory
Research Library Prototyping Team

Web Science and Digital Libraries Research Group
Old Dominion University

CNI Fall 2020 Meeting

Thanks to:



SIGIR
Special Interest Group
on Information Retrieval

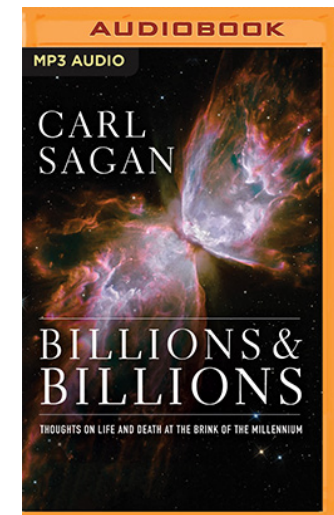




The abundance of data in web archives



@shawnmjones @StormyArchives



Web archives contain numerous documents



Brewster Kahle
@brewster_kahle

Wayback Machine now has 898,570,440,000 URL's -- serving millions of users as a foundation for what has been said on a constantly shifting WWW. (up 17 billion URLs in less than 1 month ago

[twitter.com/brewster_kahle...](https://twitter.com/brewster_kahle/status/1225167435399036939)) Go [@internetarchive](https://www.archive-it.org/) !



Brewster Kahle @brewster_kahle · Jan 10

Wayback Machine just grew to 881,352,519,000 web URL's. That is 881 Billion. For every one that becomes important in the news or in someones personal world, we crawl and store millions of others just-in-case. go [@internetarchive](https://www.archive-it.org/)

2:20 PM · Feb 5, 2020 · Twitter Web App

- https://twitter.com/brewster_kahle/status/1225167435399036939
- <https://www.webarchive.org.uk/en/ukwa/info/faq>
- <https://archive-it.org/collections/3572>

How big is the archive?

The UK Web Archive collects millions of websites each year and billions of individual assets (pages, images, videos, pdfs etc.). As of 2017 we have collected approximately 500TB of data and increasing this by over roughly 60 – 70 TB a year.



ARCHIVE-IT

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)

More ▼

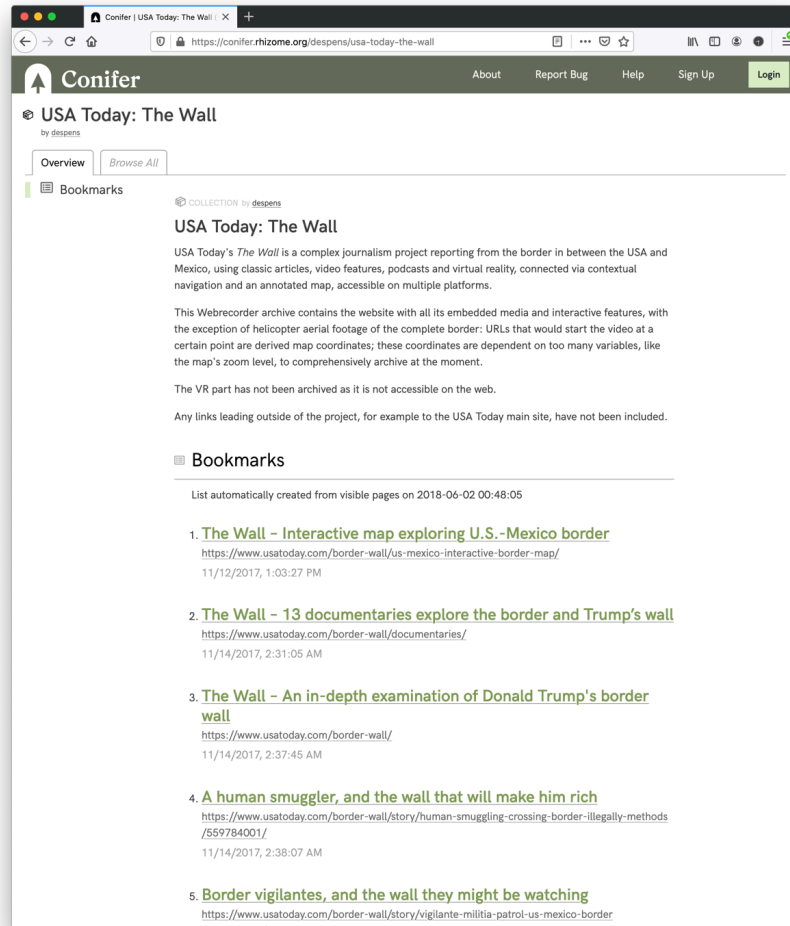
Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

339,166 Total Results



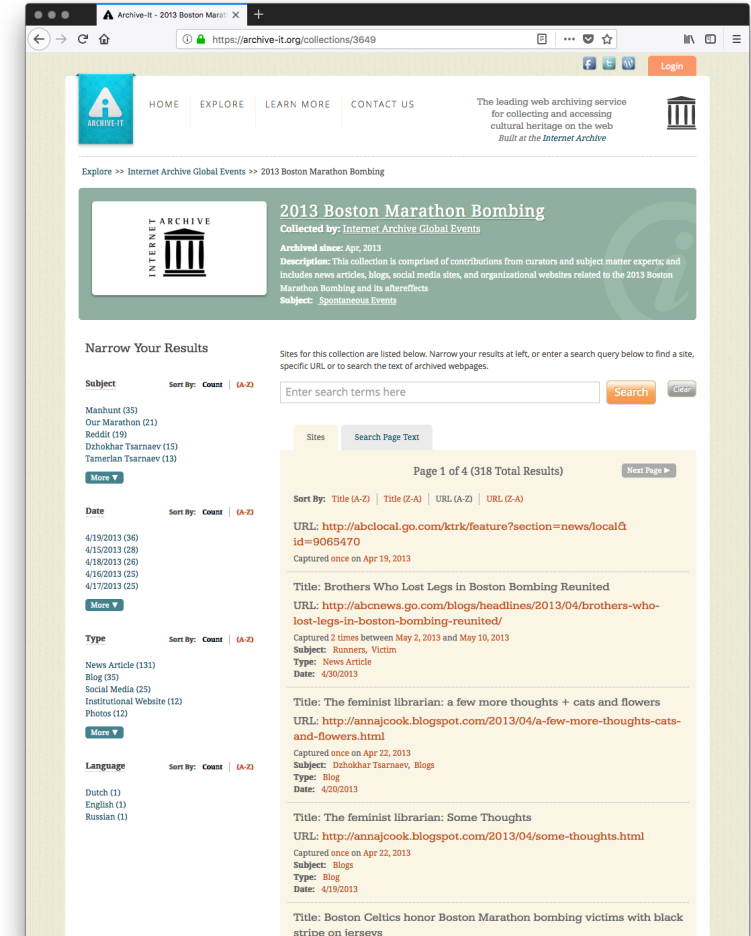
Mementos are the documents in web archive collections



Web archive collections consist of **seeds** and crawls of those seeds.

The **mementos** from these crawls are versions of pages from the time of the crawl.

The mementos are the documents in our collections.



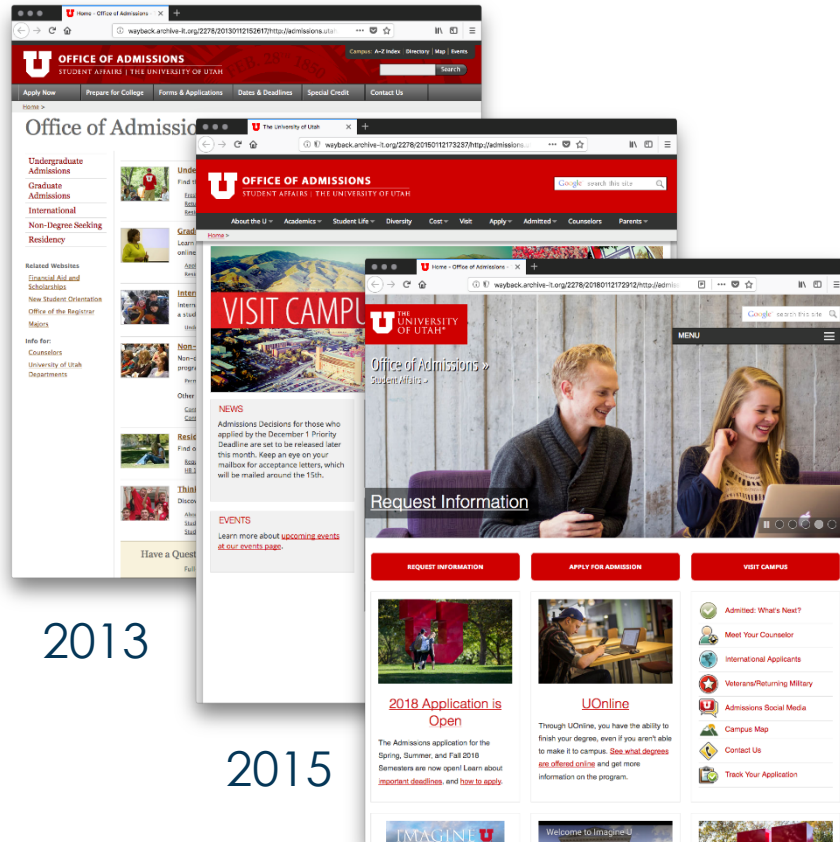
Collections with Conifer/Webrecorder

Collections with Archive-It



@shawnmjones @StormyArchives

Web archives have many versions (mementos) of the same page and *this is a feature...*



2013

2015

2018

University of Utah Office of Admissions
from the University of Utah Web Archive Collection



2/12/2015

3/5/2015

4/1/2015

Tumblr Black Lives Matter Blog
from the #blacklivesmatter Collection

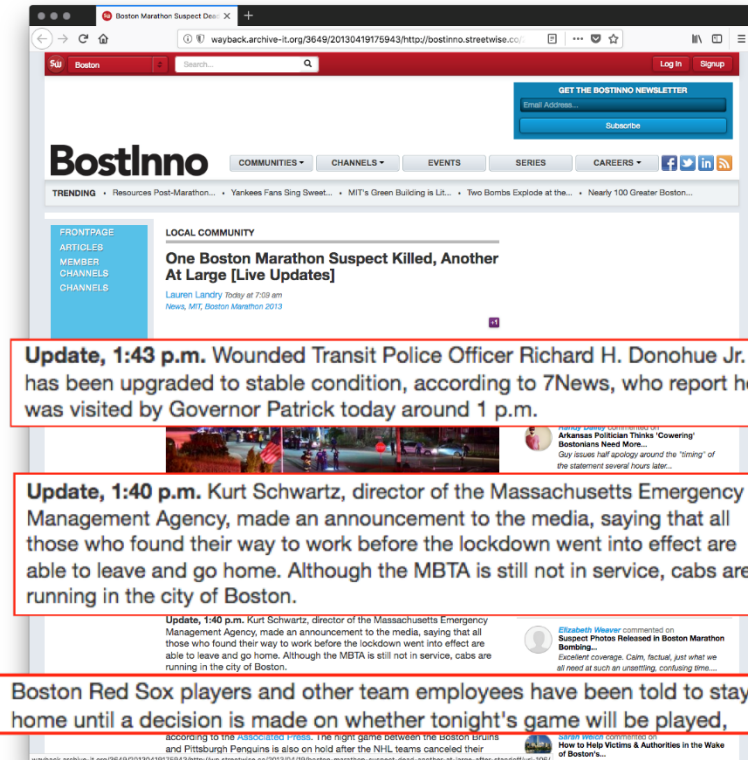


@shawnmjones @StormyArchives

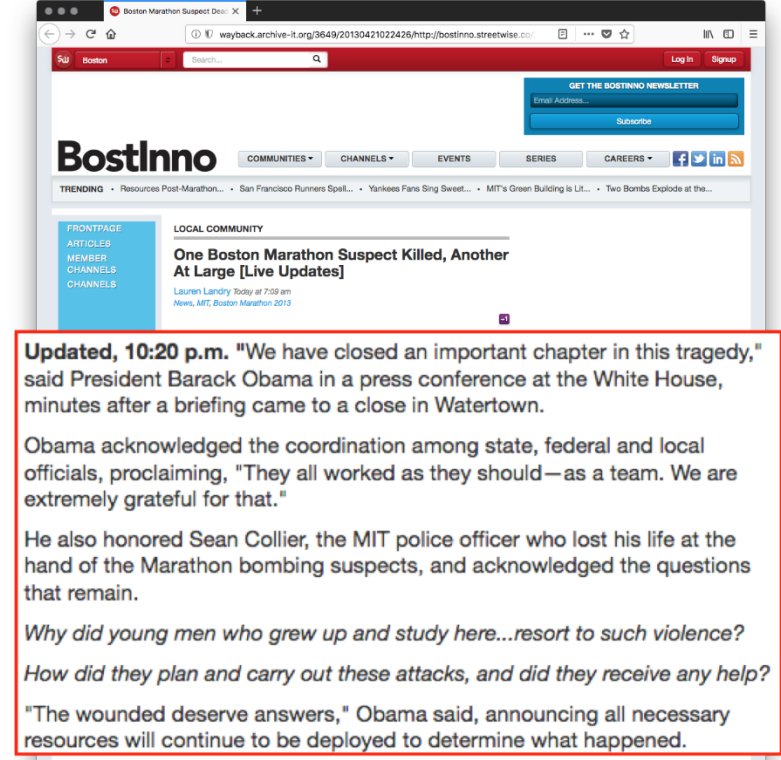
Different versions (mementos) allow us to see an unfolding news story



**Memento from
April 19, 2013 17:12**
Searching for suspects,
City on lockdown



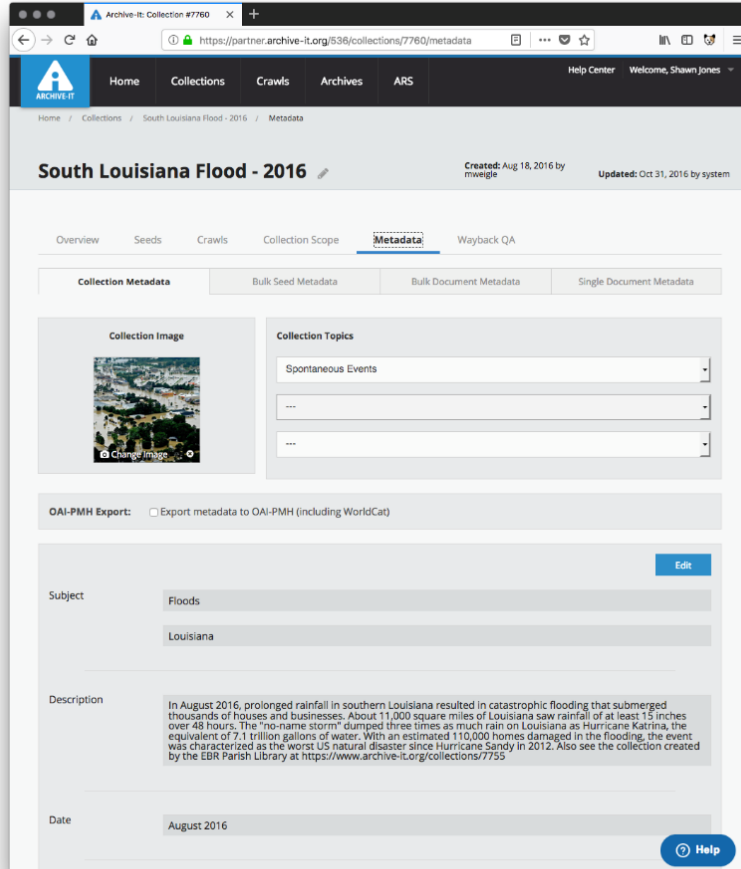
**Memento from
April 19, 2013 17:59**
Officer Donohue in hospital,
Lockdown loosened,
Will the Red Sox game be cancelled?



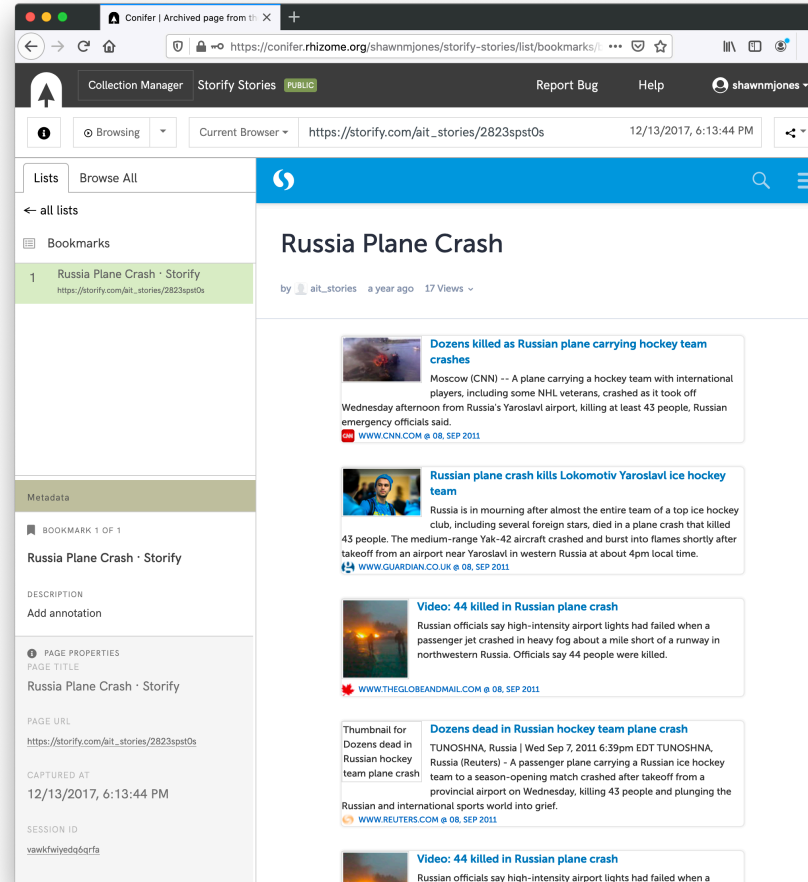
**Memento from
April 24, 2013 2:24**
Suspect Found,
Officer Collier lost life,
Obama speaks



Tools exist that easily allow users to add content to web archives...



Archive-It



Conifer/Webrecorder



Internet Archive's
Save Page Now



Archive.Today's
On-demand
Web Form



@shawnmjones @StormyArchives

... and these collections are used by other researchers



These collections live a life after their curator has stopped adding to them.

The collection curator is not the only user of the collection!

Consider Archive-It as an example...



There are many collections
on the same topic:
31 Archive-It collections
match the search query
“human rights”

Which one is best for
my needs?

Human Rights Documentation Initiative
Collected by: University of Texas at Austin Libraries, Human Rights Documentation Initiative
Archived since: Apr, 2009
Description: The University of Texas Libraries' Human Rights Documentation Initiative features fragile websites containing human rights documentation and related content from human rights organizations and advocates across the globe. Browse collection by topic, geographic area, language or do a keyword search at http://www.lib.utexas.edu/hrdi/hr_archive.
Subject: Society & Culture, Politics & Elections, Blogs & Social Media

Narrow Your Results

Group Sort By: Count | (A-Z)

- Association africaine de Défense des droits de l'Homme (ASADHO) (2)
- Cambodian Center for Human Rights (CCHR) (2)
- Change for Equality (2)
- Chinese Human Rights Defenders (2)
- East Turkistan Information Center (2)

More ▾

Subject Sort By: Count | (A-Z)

- advocacy, activism, and responses to persecution-activists (54)
- civil and political rights-women's rights (48)
- armed conflict and persecution-torture (34)

Sites Search Page Text

Page 1 of 3 (267 Total Results)

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

Title: Adaleh Center for Human Rights Studies
URL: <http://adaleh-center.org/new/>
Description: In Arabic and English. ADALEH Center for Human Rights is a non-governmental, non-profit organization founded in September 2003, with a mission of enforcing human rights in Jordan and Arab world. Projects include youth capacitation, training medical staff in torture detection, and capacitating lawyers in the field of human rights. Site contains news and publications.

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

Human Rights
Collected by: University of Manitoba
Archived since: Apr, 2014

Human Rights
Collected by: Duke University
Archived since: Nov, 2011
Description: *This collection is under development. Content will become available as it is crawled by Archive-It.* The Human Rights Archive, part of the David M. Rubenstein Rare Book & Manuscript Library, collects, preserves, and provides access to materials generated by organizations and individuals working to advance human rights. This Archive-It collection preserves web pages from the Archive's collections.

How are they different
from each other?



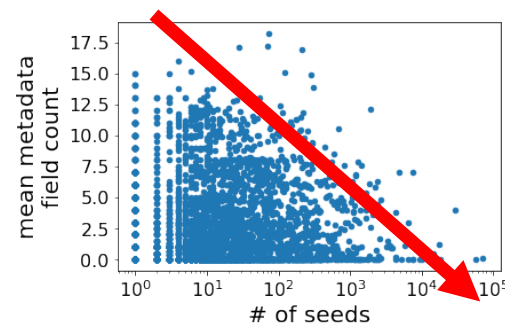
Metadata is there, but does not always help

Because metadata is optional it is not always present.

Metadata on Archive-It collections:

- many different curators
- different organizations
- different content standards
- different rules of interpretation
- **it is inconsistently applied**

This means that a user cannot reliably compare metadata fields to understand the differences between collections.



Archive-It - 2013 BART Strike

San Francisco Public Library

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.

Subject: Government, Government, Cities, Government, Counties, 2013 BART strike

Creator: Bay Area Rapid Transit

Publisher: Bay Area Rapid Transit

Source: Internet Archive

Format: Web page and PDF

Language: English

Collector: San Francisco Public Library

Narrow Your Results

Creator: San Francisco Bay Area Rapid Transit District (9)

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

9 seeds
with metadata

Archive-It - 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

Specific URL, or to search the text of archived webpages.

(132,599 Total Results)

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

URL: http://publications.gc.ca/collections/collection_2010/statcan/91-008-X/91-008-X2010007-fra.pdf

Captured 2 times between Apr 12, 2013 and Apr 18, 2013

URL: http://publications.gc.ca/collections/collection_2010/statcan/97-001-X/97-001-X2010003-fra.pdf

Captured 2 times between Apr 16, 2013 and Apr 18, 2013

URL: http://publications.gc.ca/collections/collection_2011/statcan/91-008-X/91-008-X2010015-fra.pdf

Captured once on Apr 23, 2013

URL: http://publications.gc.ca/collections/collection_2011/statcan/97-001-X/97-001-X2010004-fra.pdf

Captured once on Apr 23, 2013

Title: Geology, eastern Prince of Wales Island and adjacent smaller islands, Nunavut

URL: <http://publications.gc.ca/collection/R/GSC-CGC/M42E/974/2016A.pdf>

Description: Accompanies Bulletin 574, "Geology of eastern Prince of Wales Island and adjacent smaller islands, Nunavut" (see Cat. number M42/574B/575).

132,599 seeds
no metadata

Paradox:
More seeds = more effort
More seeds = greater user need for metadata



@shawnmjones @StormyArchives

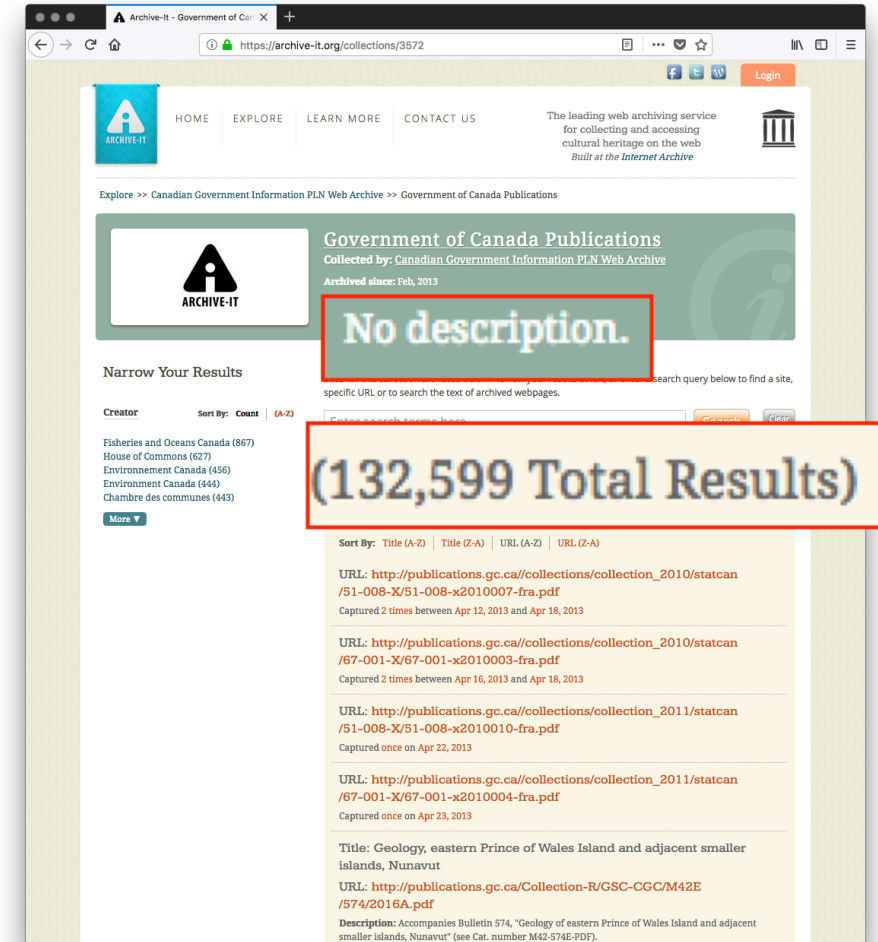
Jones, S.M., Weigle, M.C., & Nelson, M.L. (2019). Social Cards Probably Provide For Better Understanding Of Web Archive Collections. In ACM CIKM '19.
<https://doi.org/10.1145/3357384.3358039>

Reviewing mementos manually is costly

Some collections have 1000s of seeds

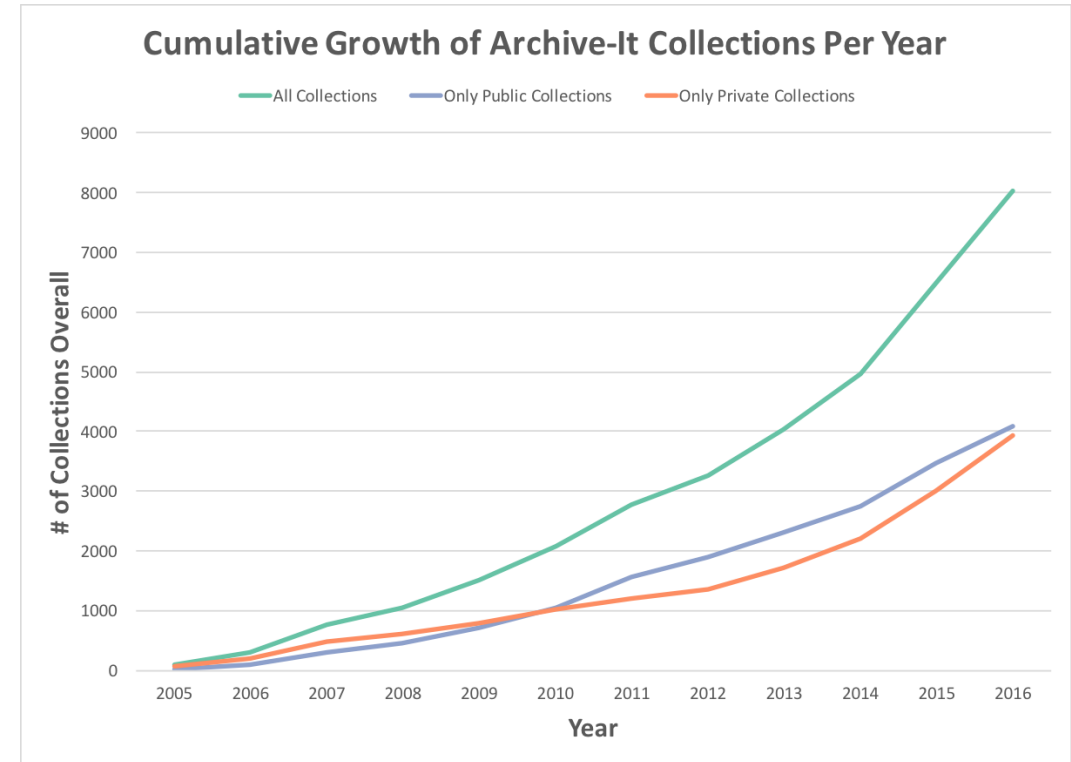
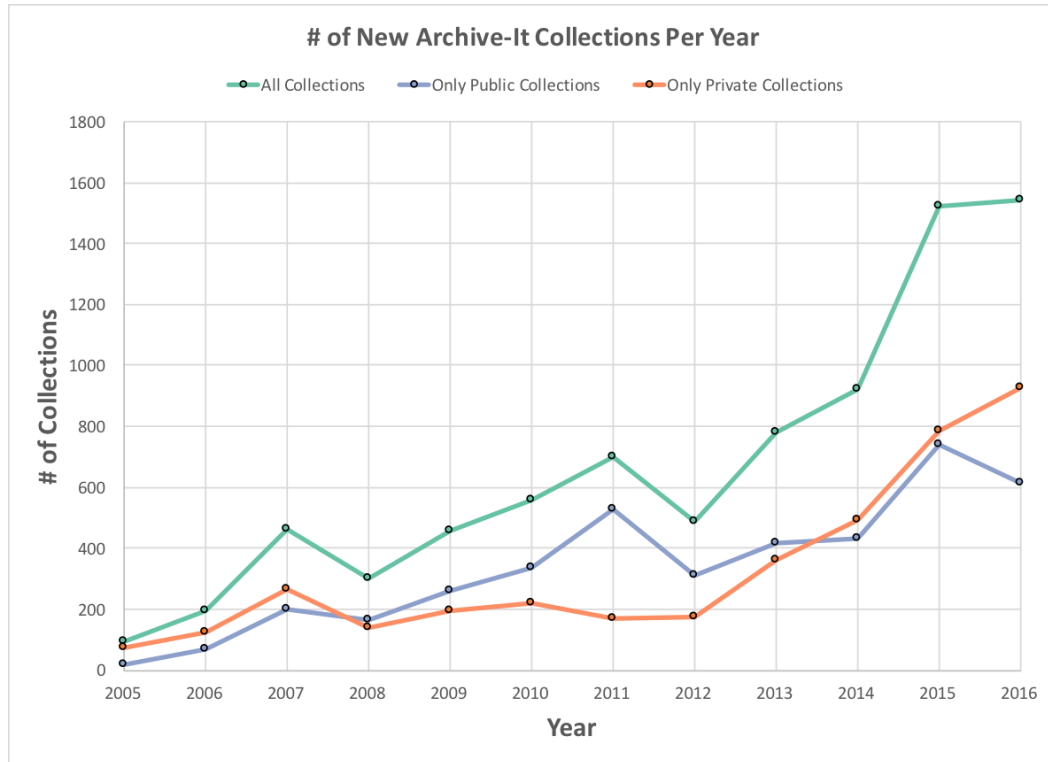
Each seed can have many mementos

In some cases, this can require reviewing 100,000+ documents to understand the collection



This collection has 132,599 seeds, many with multiple mementos

In addition to more mementos in web archives, more Archive-It collections are added every year

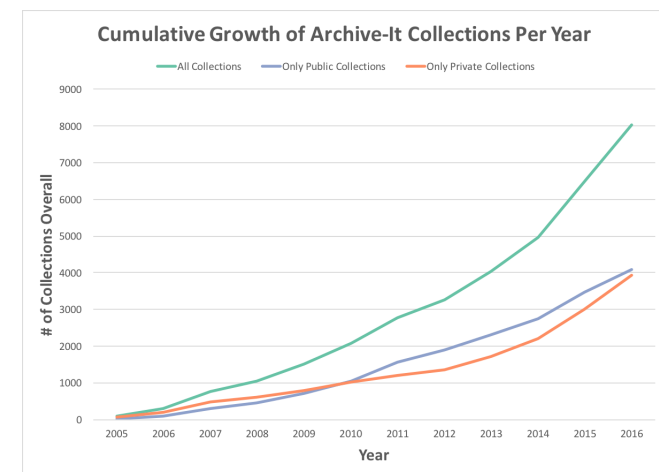
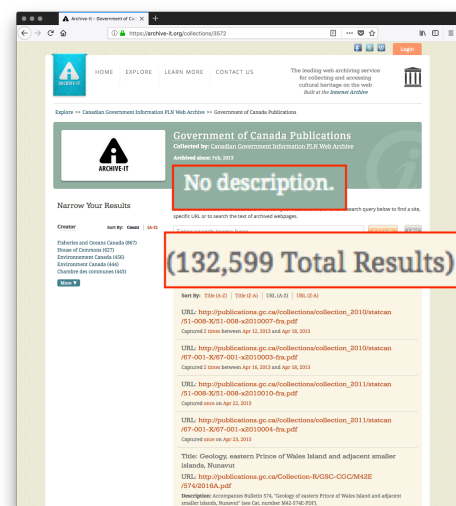
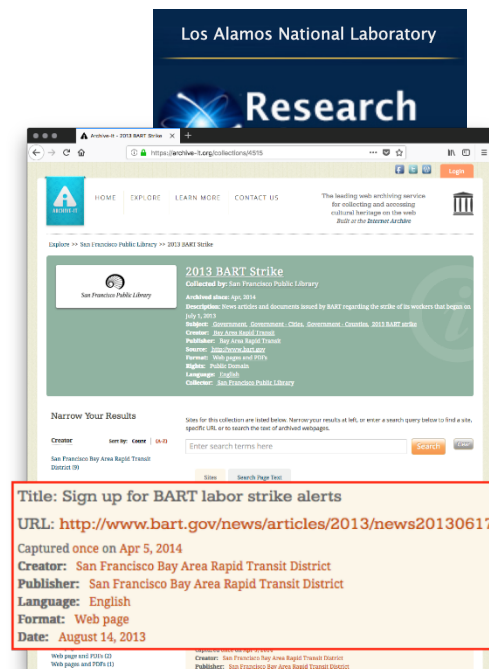
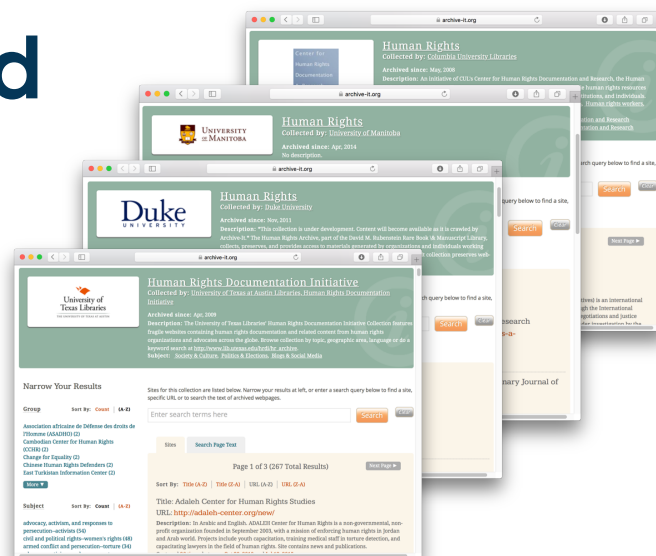


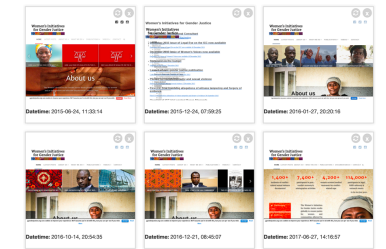
More than 15,000 collections exist



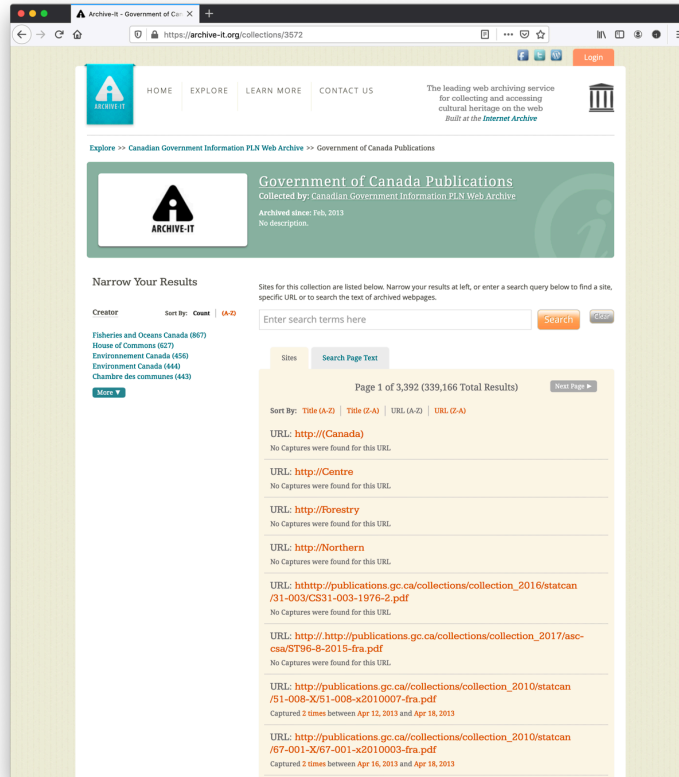
The problem, summarized

- There are multiple collections about the “same concept.”
- The metadata for each collection is non-existent, or inconsistently applied.
- Many collections have 1000s of seeds with multiple mementos.
- There are more than 15,000 collections.
- **Human review of these mementos for collection understanding is prohibitively expensive.**

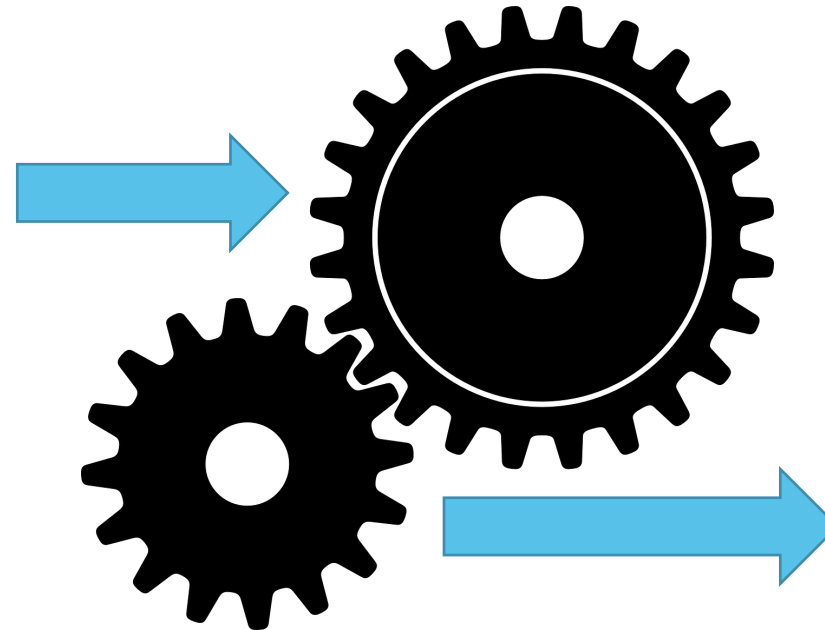




Ideally, a user would be able to glance at a visualization and gain understanding of the collection



Web Archive Collection



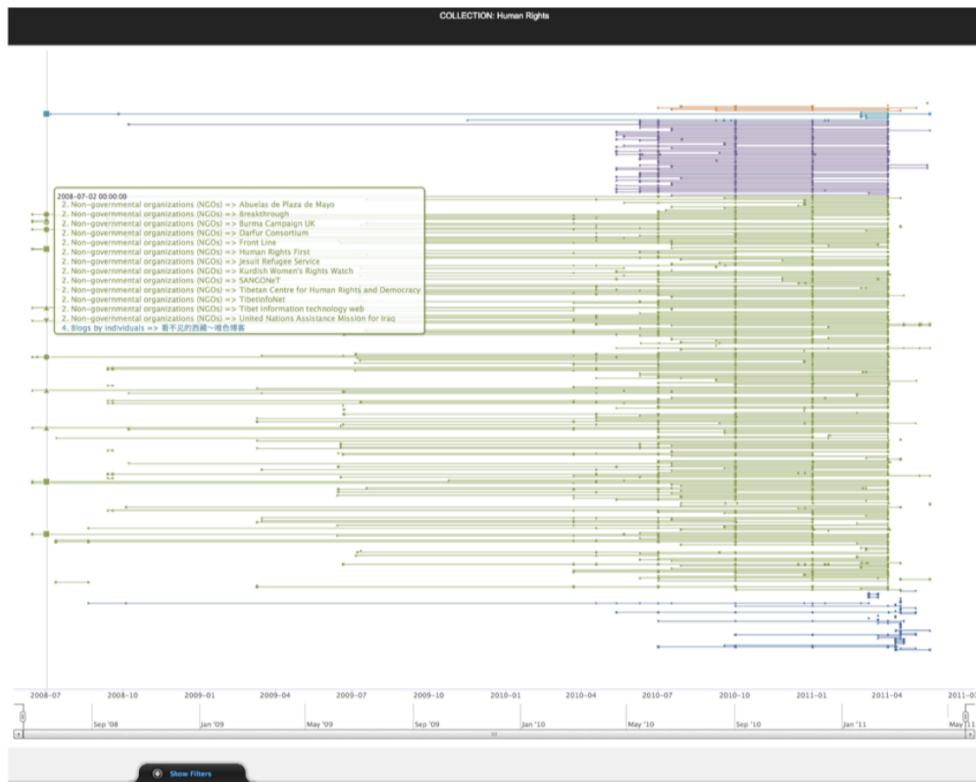
Automated
Solution



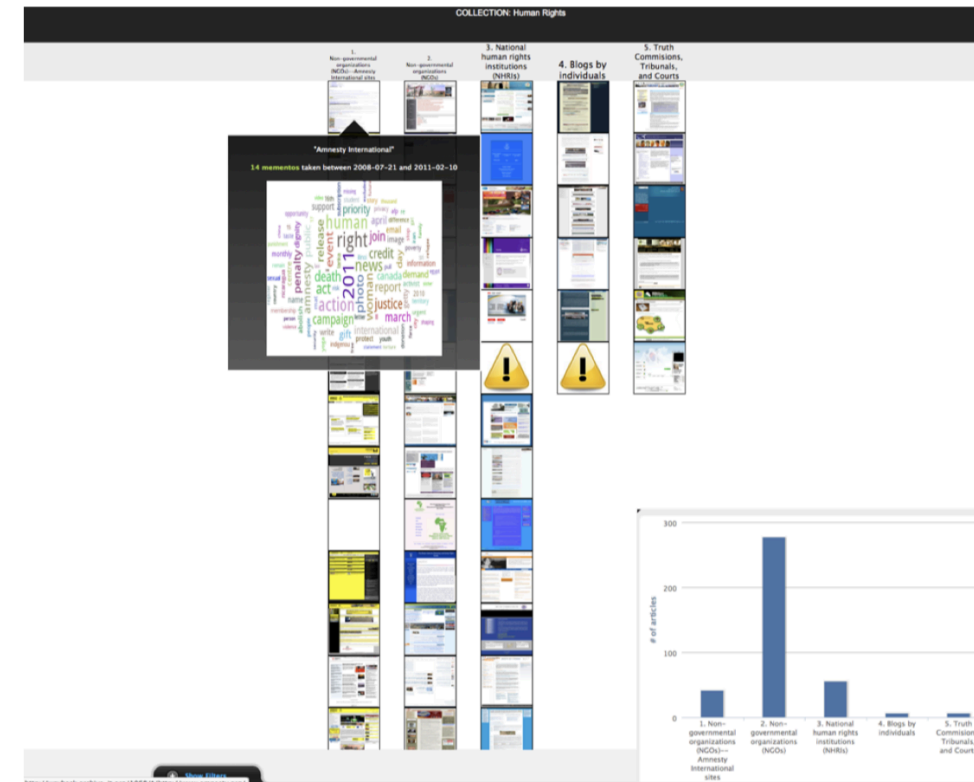
Visualization that
conveys understanding
at a glance



But techniques that apply to every page are hard to scale up and difficult to understand...



Timelines provide an overview, but do not provide understanding at a glance.



Thumbnails work well for small collections, but do not scale.



@shawnmjones @StormyArchives

K. Padia, Y. AlNoamany, and M. C. Weigle, (2012) "Visualizing digital collections at Archive-It," In JCDL '12. <https://doi.org/10.1145/2232817.2232821>

What if we could apply a visualization technique that users already know how to process?

Users would require no training and could instead get information immediately.

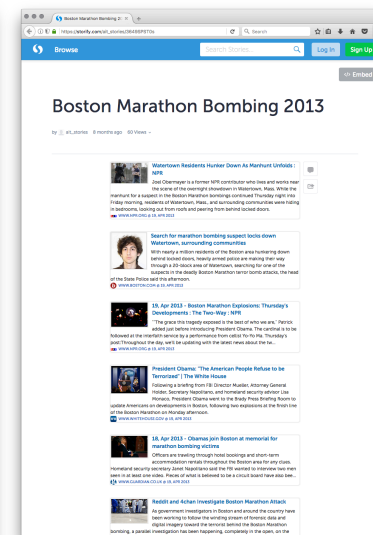
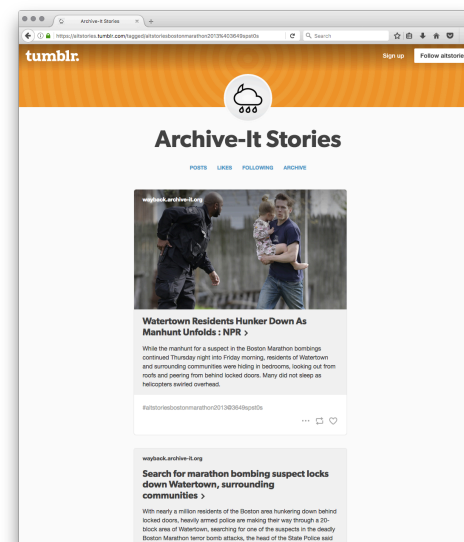
Users could quickly decide which collection meets their needs!



Social media storytelling



@shawnmjones @StormyArchives



Web surrogates provide a visual summary of a web resource drawn from the content of the resource

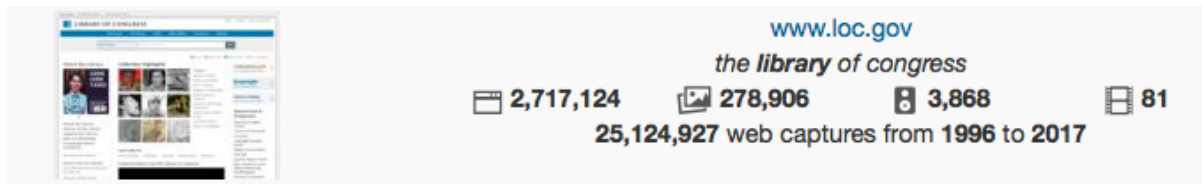
Home | Michele C. Weigle

www.cs.odu.edu/~mweigle

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](#)

Text snippet (example from Bing)



Text + Thumbnail (example from Internet Archive)

Instances



Instances archived after 01-jul-2008 and before 31-jan-2010



Browser Thumbnail (example from UK Web Archive)



Social Card (example from Facebook)



@shawnmjones @StormyArchives

S. M. Jones. "Let's Get Visual and Examine Web Page Surrogates."
<https://ws-dl.blogspot.com/2018/04/2018-04-24-lets-get-visual-and-examine.html>, 2018.

Social cards are summaries of web resources

Long URL: <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>

VS.

The same URL represented as a social card:



- URLs can be difficult for people to comprehend
- Social cards provide a title, small text snippet, and striking image from the web page behind a URL

And we use these cards all the time!



Facebook

2017-08-11: Where Can We Post Stories Summarizing Web ...
<https://ws-dl.blogspot.com/2017/08/2017-08-11-where-can-we-post-stories.html> ▾
 Aug 11, 2017 · In Yasmin AlNoamany's Dark and Stormy Archives (DSA), she uses social cards in Storify stories to summarize mementos from Archive-It collections. These stories take the form of 28 high ...
 Author: Shawn M. Jones

Bing



Twitter



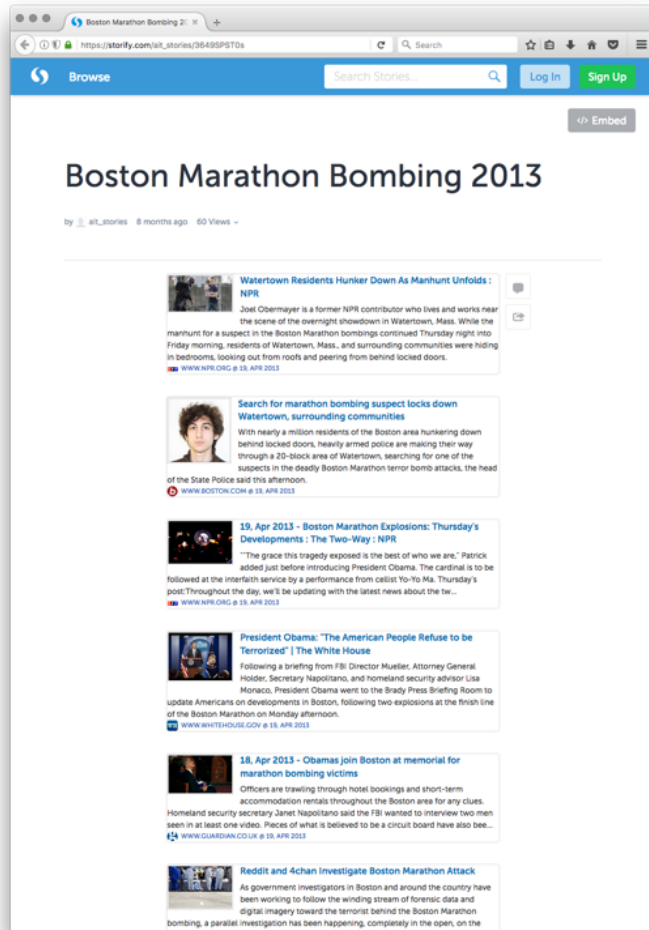
Tumblr

ws-dl.blogspot.com › 2017/08 › 2017-08-11-where-ca... ▾
Where Can We Post Stories Summarizing Web Archive ...
 Aug 11, 2017 — In effect, social cards are visualizations of web objects, piercing the veil ... Earlier attempts at visualizing Archive-It collections by Padia and ...

Google



Social media storytelling uses groups of social cards to provide a “summary of summaries”

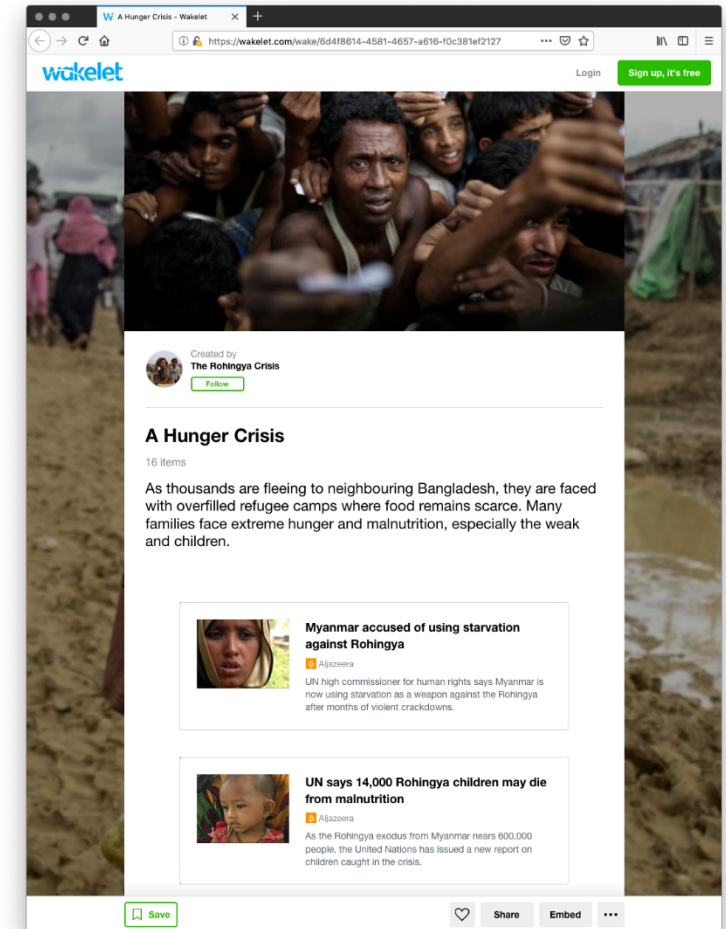


Each social card summarizes a web resource.

Each story groups the social cards, summarizing the topic.

Social cards contain the same information in the same place on each card, allowing for easy comparison.

We want to use this technique to summarize web archive collections **because users are already familiar with this visualization paradigm.**



6 resources are shown in this Storify story

2 resources are shown in this Wakelet story



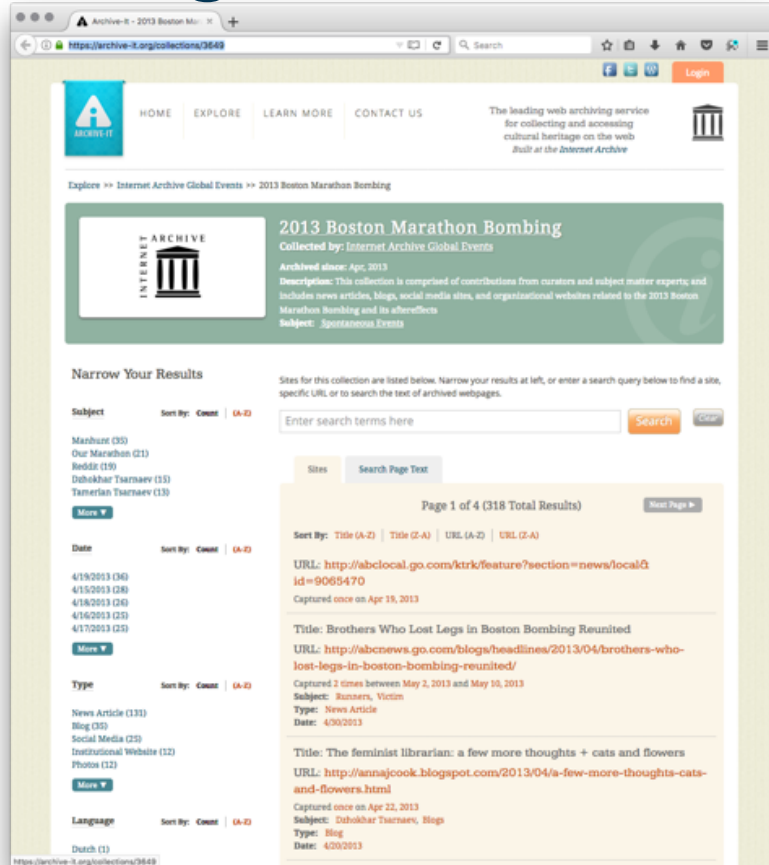
@shawnmjones @StormyArchives

Social media storytelling uses an interface that users already understand how to process!

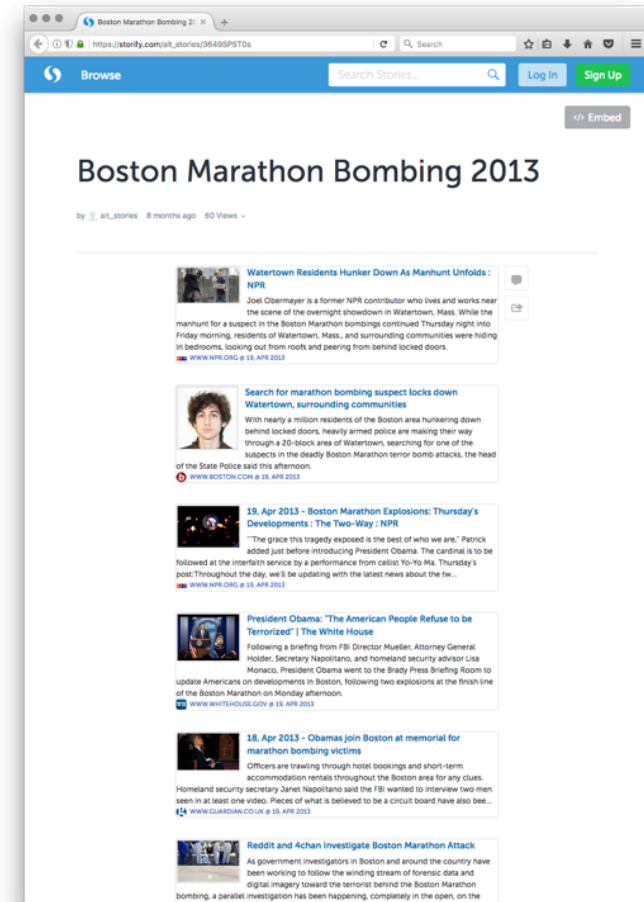
Users would require no training and could instead get information immediately.

Users could quickly decide which collection meets their needs!

Our solution: applying social media storytelling to web archives



From this:
318 seeds with
2421 mementos



To something like this:
a **social media story**
of ~28 cards



Storytelling With Web Archives Has Multiple Use Cases...

1. Promotion of the collection

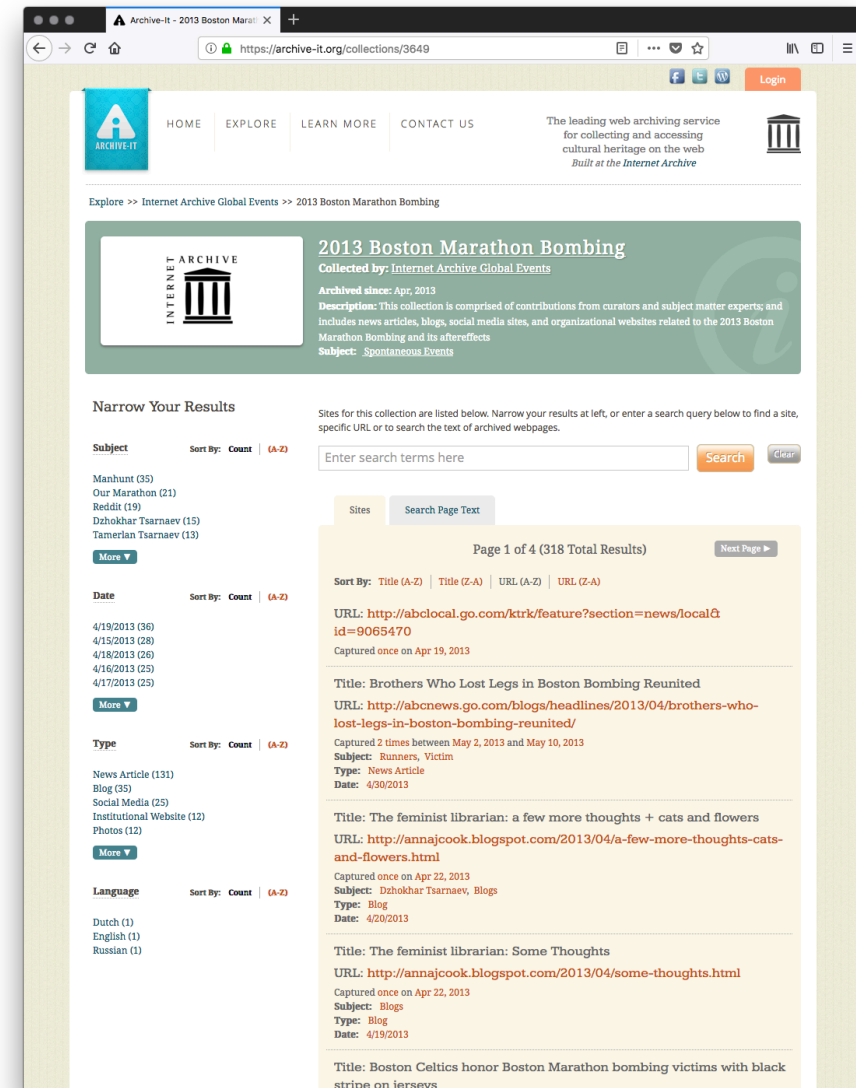
- Storytelling allows a curator to promote a collection, making others aware of it

2. Exploring aspects of the collection

- Users can explore a collection and expose specific sides of a news story or focus on specific people or places

3. Summarization

- Web archive collections are too large for manual review – we need a summary to understand what they contain



Some services equate “stories” with ephemera, which is not what we want



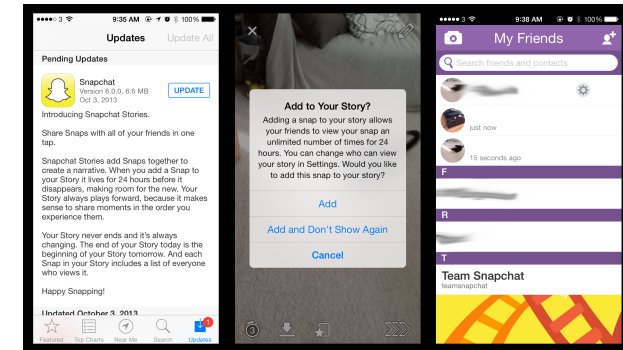
Facebook stories

Image ref:
<https://techcrunch.com/2018/04/05/facebook-stories-default/>



Instagram stories

Image ref:
<https://buffer.com/library/instagram-stories>



Snapchat stories

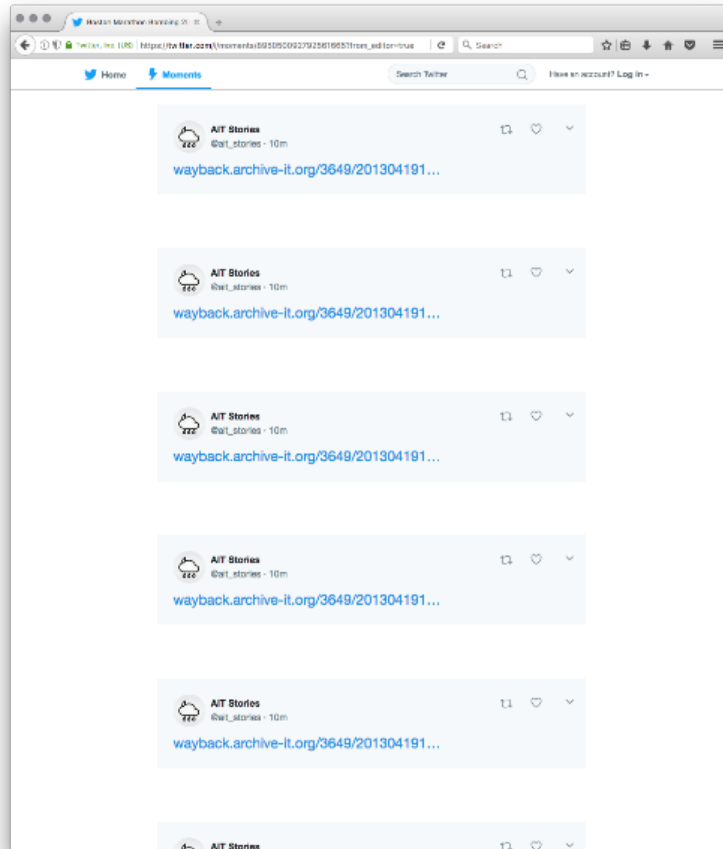
Image ref:
<https://techcrunch.com/2013/10/03/snapchat-gets-its-own-timeline-with-snapchat-stories-24-hour-photo-video-ales/>

These platforms delete the user's stories 24 hours after they are posted.

This form of social media storytelling is the opposite of what we are looking for.
We want the stories to be artifacts themselves.



Existing social platforms do not work well for archived resources (mementos)



twitter



@shawnmjones @StormyArchives



facebook

Jones, S.M., Klein, M., Weigle, M.C., & Nelson, M.L. (2020). MementoEmbed and Raintale for Web Archive Storytelling. In Web Archiving and Digital Libraries Workshop (WADL 2020). arXiv, abs/2008.00137. <https://arxiv.org/abs/2008.00137>

Existing card services create a confusing experience for pages from web archives (mementos)

CNN ARCHIVE-IT



Egypt - This Just In - CNN.com Blogs

To alleviate U.S. pressure on the Egyptian army, Saudi Arabia would consider matching the estimated \$1.5 billion the United States provides Egypt in military aid each year if Washington decides to cut its aid to the troubled North African nation, an Arab diplomat says.

[Read the article on wayback.archive-it.org >](https://archive-it.org/2011/02/25/egypt-this-just-in-cnn-com-blogs)

powered by embedly

embed.ly social card

This is confusing.

Who published these resources?
Archive-It?
CNN?

It is a matter of trust, because different sources are attributed.



@shawnmjones @StormyArchives

Jones, S.M., Klein, M., Weigle, M.C., & Nelson, M.L. (2020). MementoEmbed and Raintale for Web Archive Storytelling. In Web Archiving and Digital Libraries Workshop (WADL 2020). arXiv, abs/2008.00137. <https://arxiv.org/abs/2008.00137>

Current social media storytelling solutions don't work with archival content



Our stories should not be temporary.

Our summaries should not present users with confusing attribution.

Web archive collections
are vast and hard to
understand

Understanding them is
difficult with existing
visualizations

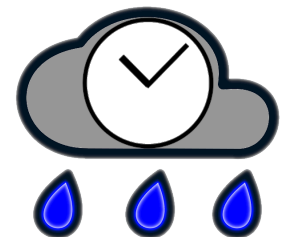
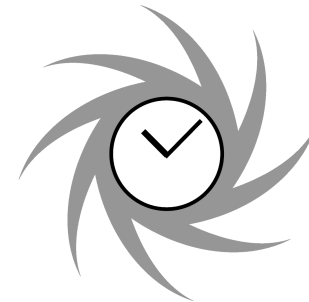
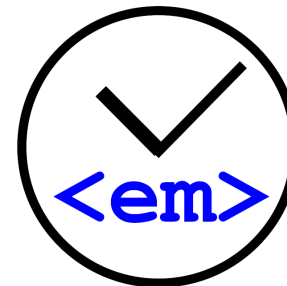
Existing social media
storytelling does a
poor job with archived
pages (mementos)



Solving the problem with the Dark and Stormy Archives Project



@shawnmjones @StormyArchives



Telling a story with web archives has 3 basic steps

1. Select the pages
for the story

2. Gather information
to summarize
each memento

3. Summarize
all mementos together
and publish the story

?



?



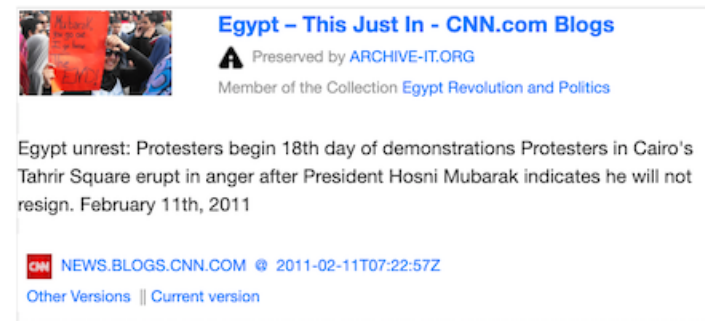
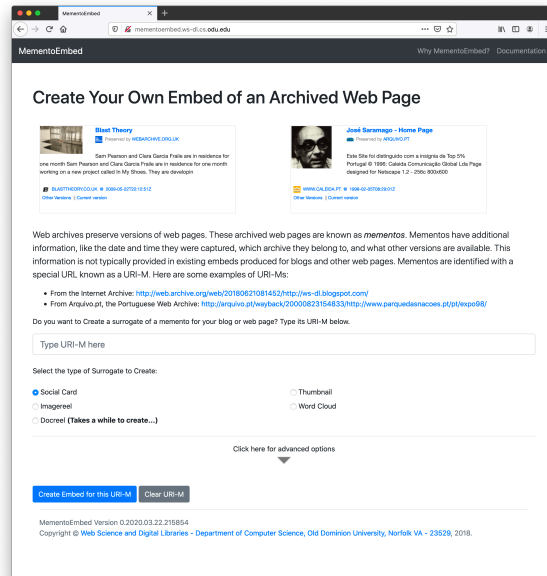
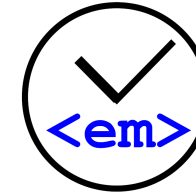
?



Neither social media services nor card services were reliable for web archive storytelling, so we created MementoEmbed...



We can summarize *individual* pages with MementoEmbed in different ways



Cards

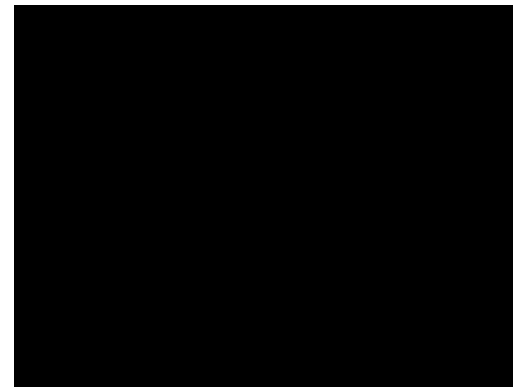


Browser Thumbnails

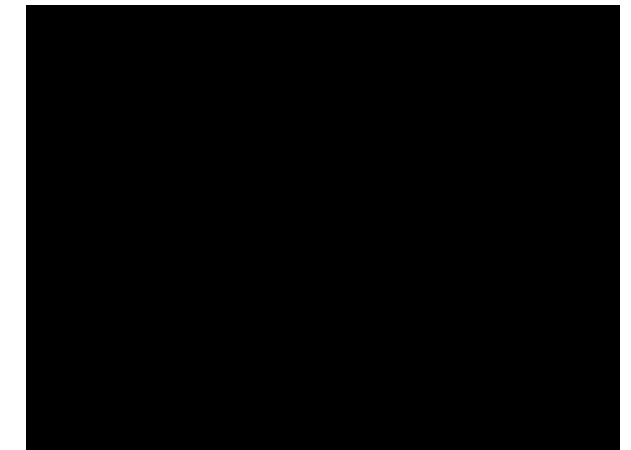


Word Clouds

All of these summaries represent the same page about the Egyptian Revolution.



Imagereels



Docreels



@shawnmjones @StormyArchives

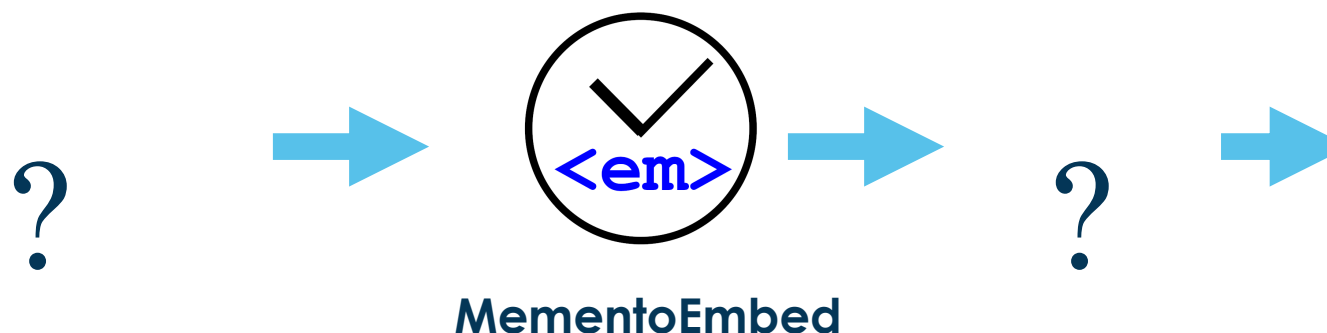
Jones, S.M., Klein, M., Weigle, M.C., & Nelson, M.L. (2020). MementoEmbed and Raintale for Web Archive Storytelling. In Web Archiving and Digital Libraries Workshop (WADL 2020). arXiv, abs/2008.00137. <https://arxiv.org/abs/2008.00137>

Now we can produce cards, but how do we create stories?

1. Select the pages
for the story

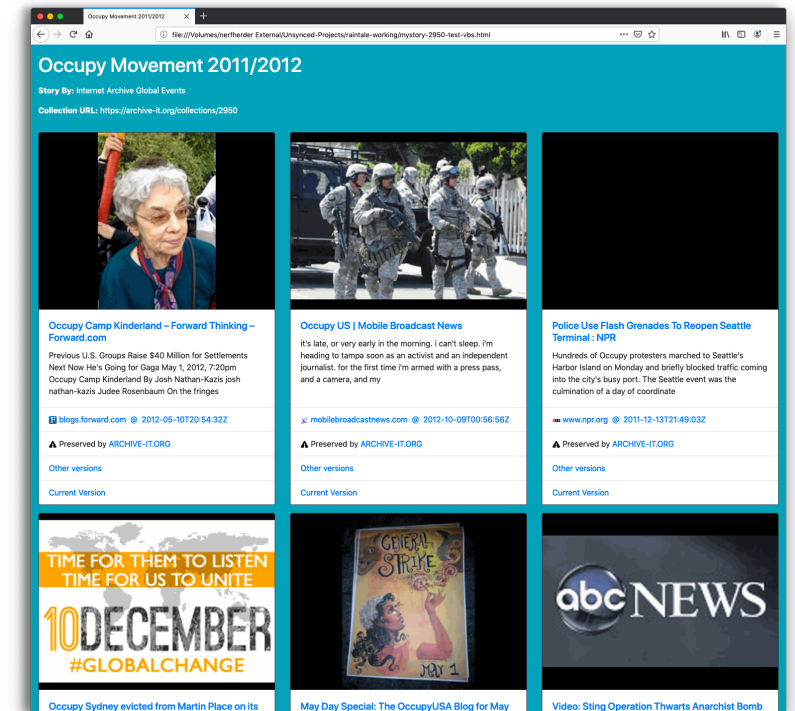
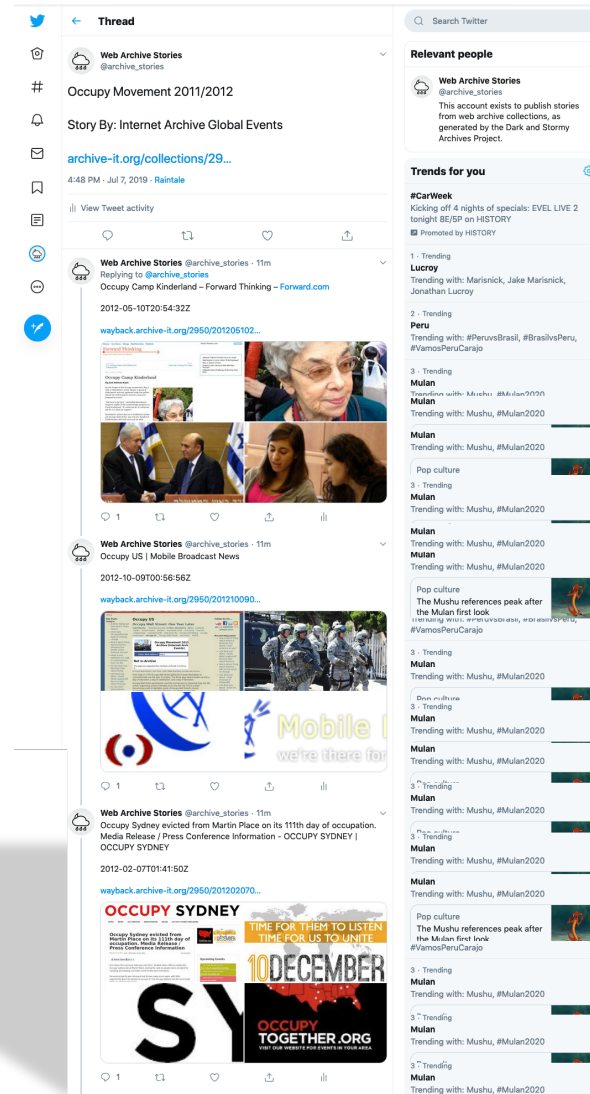
2. Gather information
to summarize
each memento

3. Summarize
all mementos together
and publish the story



MementoEmbed just produces information for individual pages, so we created Raintale to tell stories...

- Raintale generates social media stories for mementos
- Raintale publishes to file formats, like HTML, or social media services, like Twitter
- Raintale also supports templates for custom output



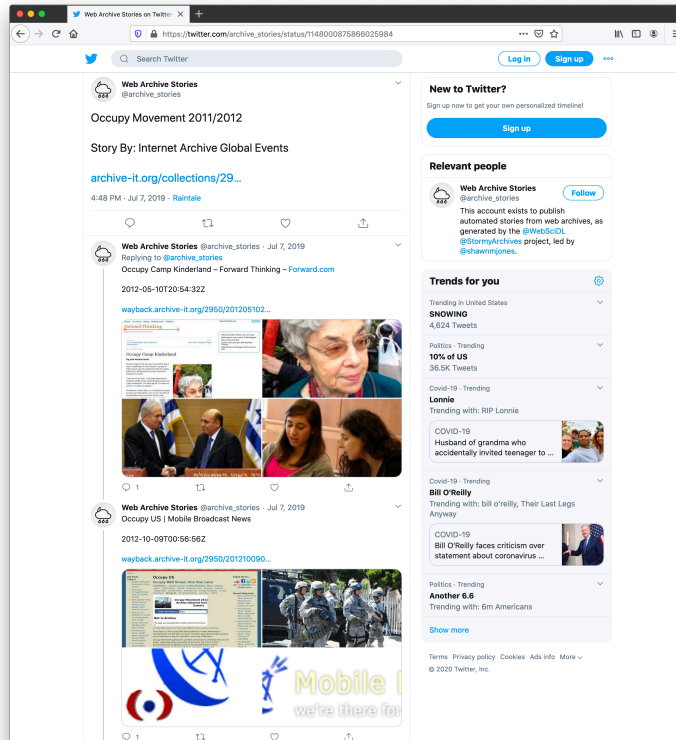
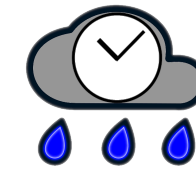
<https://oduwsdl.github.io/raintale/>



@shawnmjones @StormyArchives

Jones, S.M., Klein, M., Weigle, M.C., & Nelson, M.L. (2020). MementoEmbed and Raintale for Web Archive Storytelling. In Web Archiving and Digital Libraries Workshop (WADL 2020). arXiv, abs/2008.00137. <https://arxiv.org/abs/2008.00137>

We can publish *many* Mementos with **Raintale** in a variety of formats



Twitter Threads
with screenshots,
images, titles, dates, links



GitHub Pages
with cards and much more



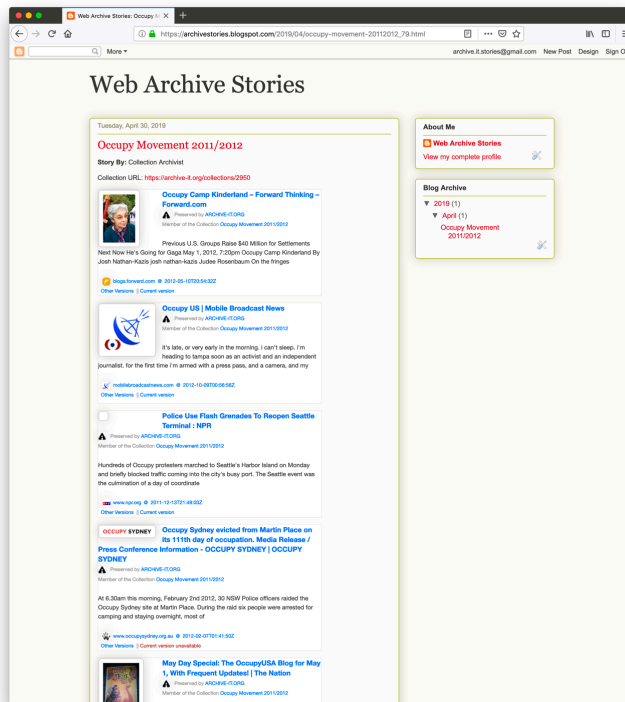
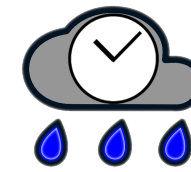
Videos
with top ranked images and sentences
to be uploaded to YouTube, Twitter, etc.
(experimental)



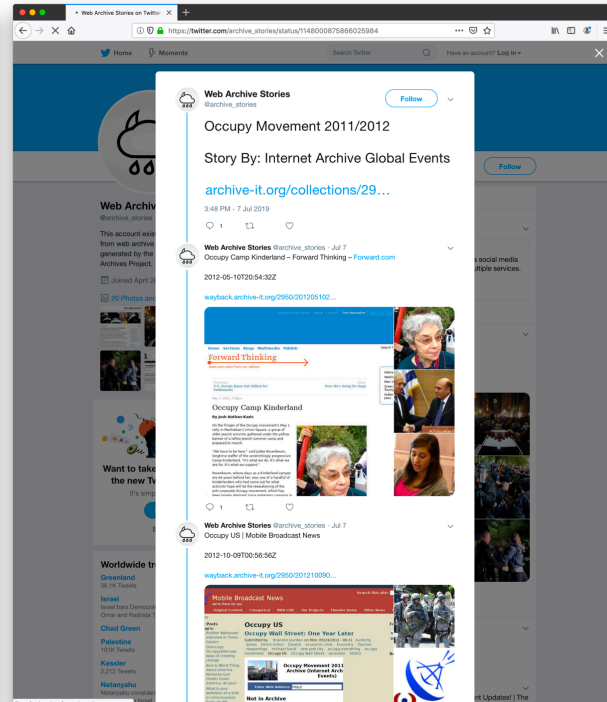
@shawnmjones @StormyArchives

Jones, S.M., Klein, M., Weigle, M.C., & Nelson, M.L. (2020). MementoEmbed and Raintale for Web Archive Storytelling. In Web Archiving and Digital Libraries Workshop (WADL 2020). arXiv, abs/2008.00137. <https://arxiv.org/abs/2008.00137>

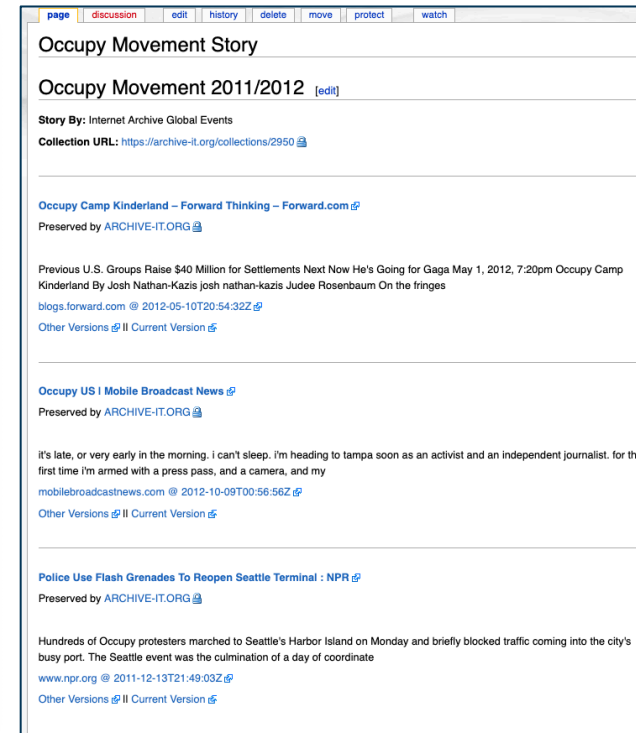
Raintale uses templates to let you customize the look and destination media of your story



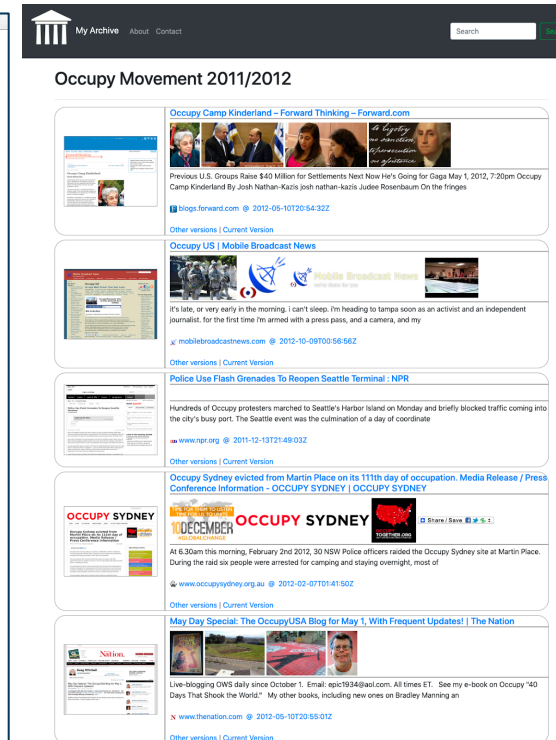
MementoEmbed
cards in Blogger



Customized Twitter
Threads



MediaWiki
Pages



Brand your own stories
for your institution



@shawnmjones @StormyArchives

Jones, S.M., Klein, M., Weigle, M.C., & Nelson, M.L. (2020). MementoEmbed and Raintale for Web Archive Storytelling. In Web Archiving and Digital Libraries Workshop (WADL 2020). arXiv, abs/2008.00137. <https://arxiv.org/abs/2008.00137>

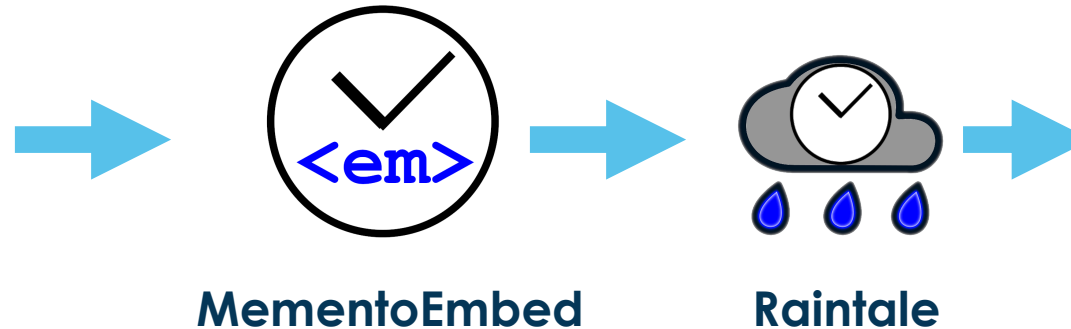
Now we can create stories, but how do we select mementos?

1. Select the pages
for the story

2. Gather information
to summarize
each memento

3. Summarize
all mementos together
and publish the story

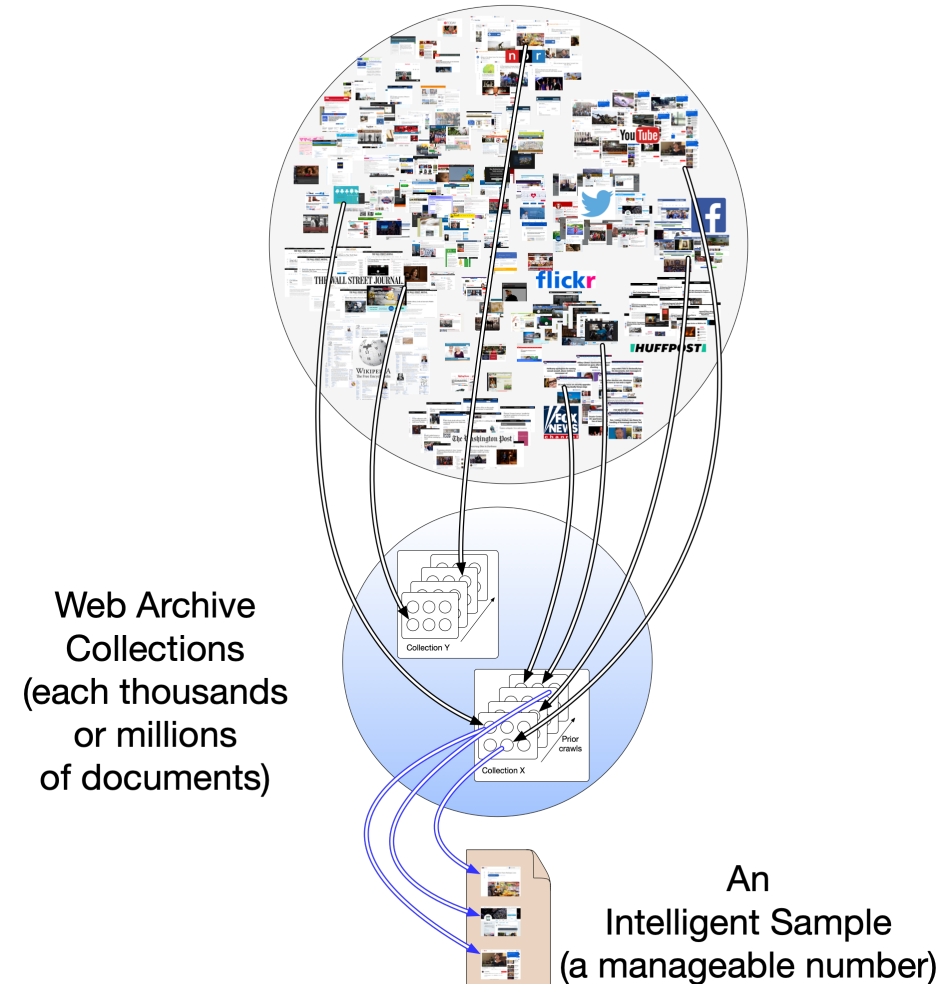
?



Hypercanes provides intelligent sampling of web archive collections

- Works with Memento-compliant archives
- Discovers mementos
- Can create new mementos
- Reports on:
 - collection metadata
 - named entities
 - terms
 - image analysis
- Provides summarization steps:
 - Clustering
 - Filtering
 - Scoring
 - Ordering

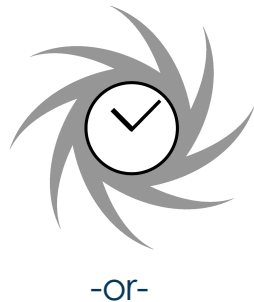
The Web in the “perpetual now”
(near ∞ documents)



The DSA Toolkit provides a solution for each stage of the storytelling lifecycle

1. Select the pages for the story

Automatically with
HyperCane

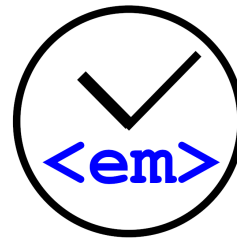


-or-



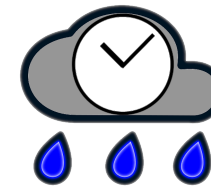
By People

2. Gather information to summarize each memento

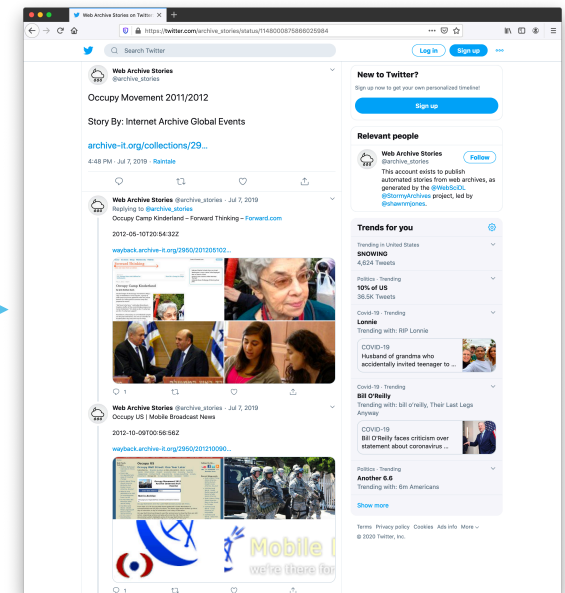


MementoEmbed

3. Summarize all mementos together and publish the story



Raintale

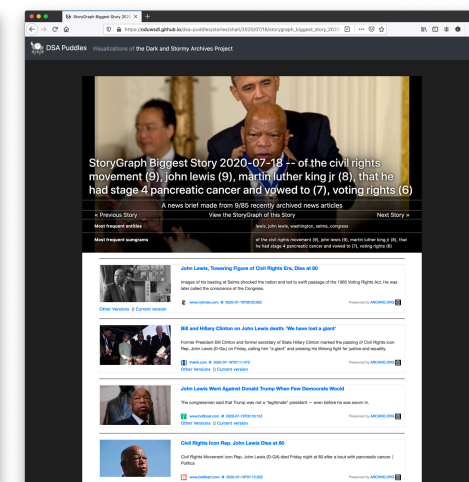
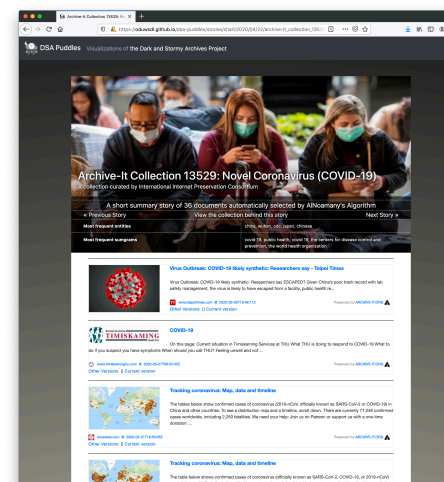




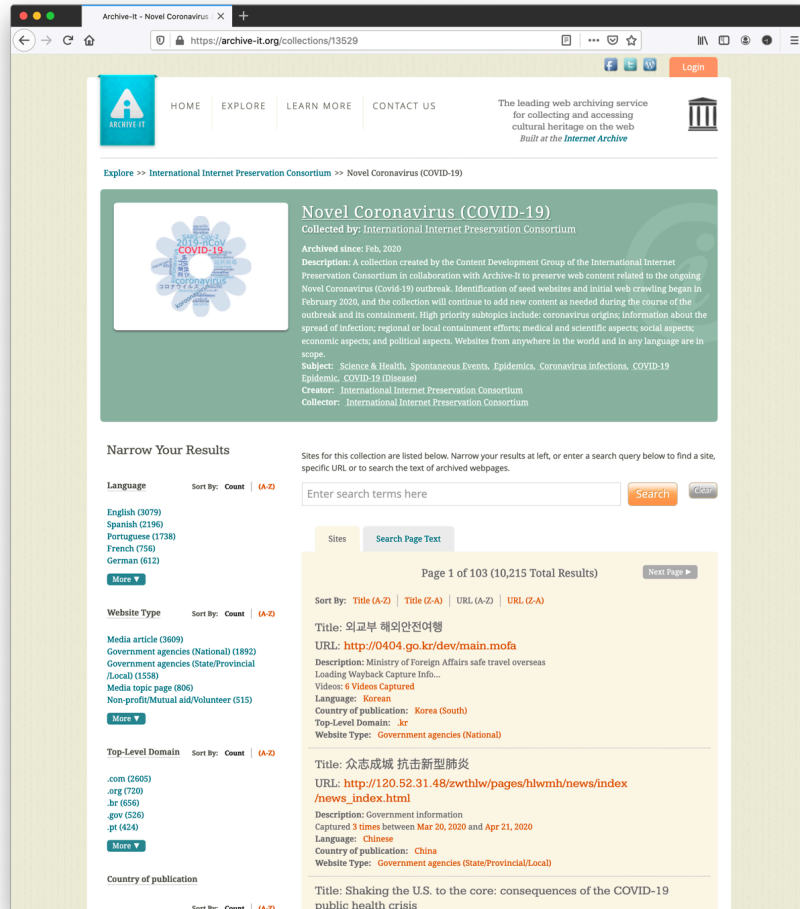
Examples of the DSA Toolkit at work



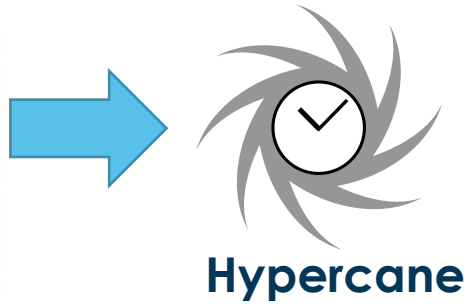
@shawnmjones @StormyArchives



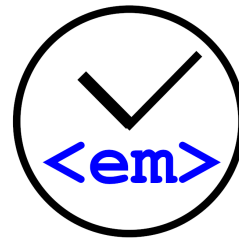
We automatically summarize public Archive-It collections



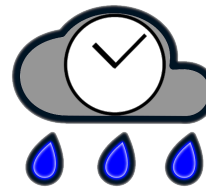
From this: 23,376 mementos



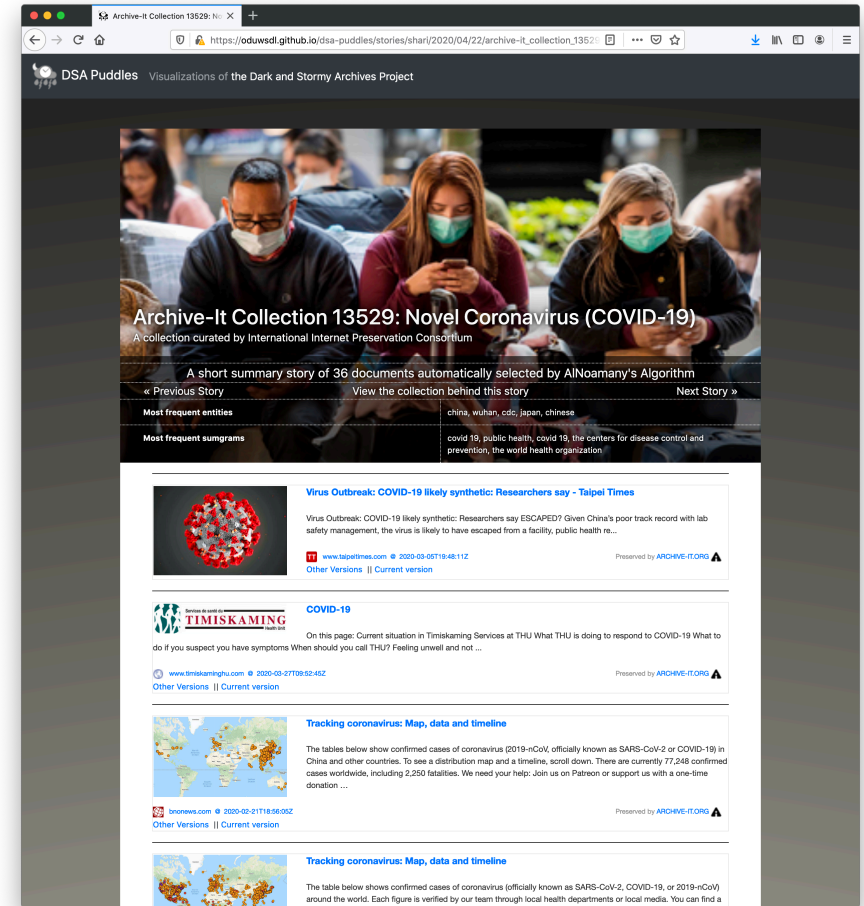
HyperCane



MementoEmbed



Raintale



To this: a sample of 36 mementos visualized as social cards, phrases, and images



@shawnmjones @StormyArchives

Jones, S.M., Klein, M., Weigle, M.C., & Nelson, M.L. (2020). MementoEmbed and Raintale for Web Archive Storytelling. In Web Archiving and Digital Libraries Workshop (WADL 2020). arXiv, abs/2008.00137. <https://arxiv.org/abs/2008.00137>

With the DSA Toolkit, we can compare Archive-It collections about mass shootings

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 AI/Noamany's Algorithm

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

Most frequent entities: virginia tech, virginia, norris hall, blackburg, the associated press

Most frequent sumgrams: virginia tech, the associated press, norris hall, 32 people, blackburg 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 year...

thelead.blogs.nytimes.com © 2008-04-27T00:21:40Z 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 © 2007-05-23T18:21:30Z 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 Files Funded in part by the National Science Foundation and produced by the graduate school of...

topics.nytimes.com © 2008-04-13T15:00:21Z 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 Lindley /

Virginia Tech

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 AI/Noamany's Algorithm

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

Most frequent entities: norway, oslo, norwegian, brevik, uleoya

Most frequent sumgrams: prime minister Jens Stoltenberg said, anders bethvig brevik, 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 Gjerfve and his family were drinking coffee outside their caravan and discussing the Oslo bombing when they began to hear shots from Uleoya between 5pm and 6pm on Friday. "I recognised the sound..."

www.telegraph.co.uk © 2011-07-31T00:30:33Z Preserved by ARCHIVE-IT.ORG

Other Versions | Current version

Norway shooting: politician describes witnessing Uleoya massacre on her blog - CyG NewsAgent

[open original article] [open original article] this article has been tweeted into @CyG_society @CyG_politics @CyG_experiment @SmartView : enhanced with bold keywords brands@breviknews...

cyp-newsagent.net © 2011-07-24T21:11:40Z 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 Uleoya

www.guardian.co.uk © 2011-07-20T20:38:40Z Preserved by ARCHIVE-IT.ORG

Other Versions | Current version

Police: 80 dead in Norway youth camp shooting - CBS News

Norway

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 AI/Noamany'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 teass

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.dailypress.com © 2019-08-08T18:19:00Z 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.dailypress.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.espn.com © 2019-08-13T17:08:30Z Preserved by ARCHIVE-IT.ORG

Other Versions | Current version

Centro Comercial El Paso

Jump to sections of this pageAccessibility HelpPress alt + / to open this menuFor a better

El Paso

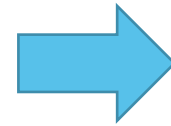
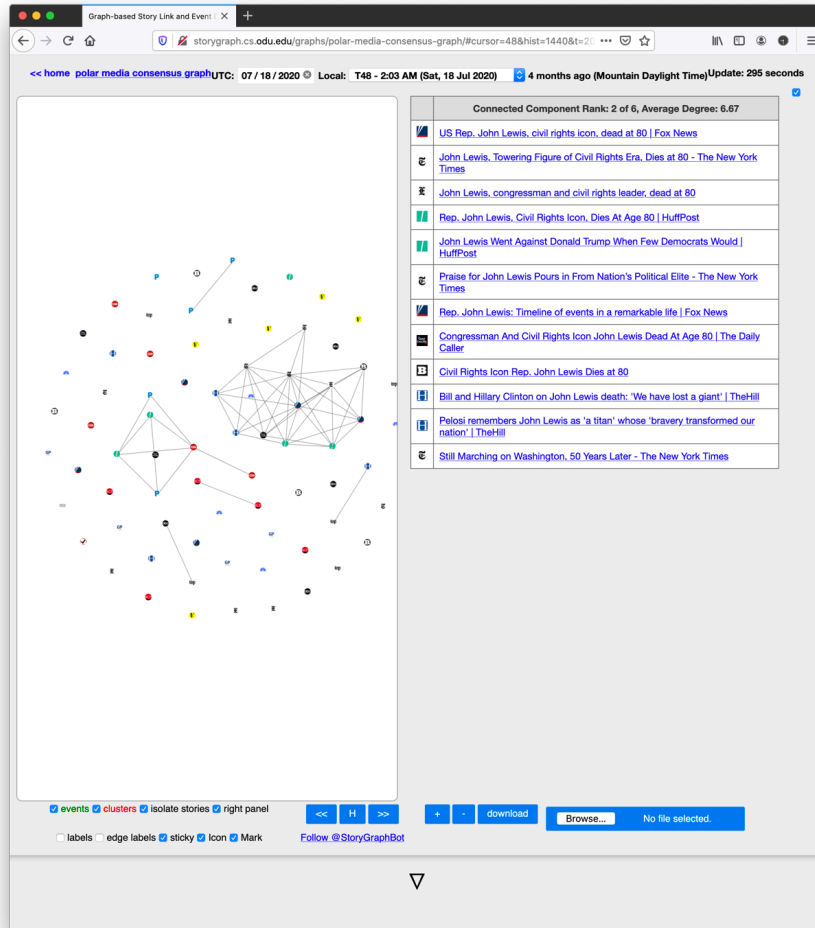


But the DSA Toolkit is NOT just for Archive-It collections...

The DSA Toolkit works with Memento-compliant Archives



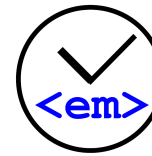
We automatically produce daily news summaries



StoryGraph
Toolkit



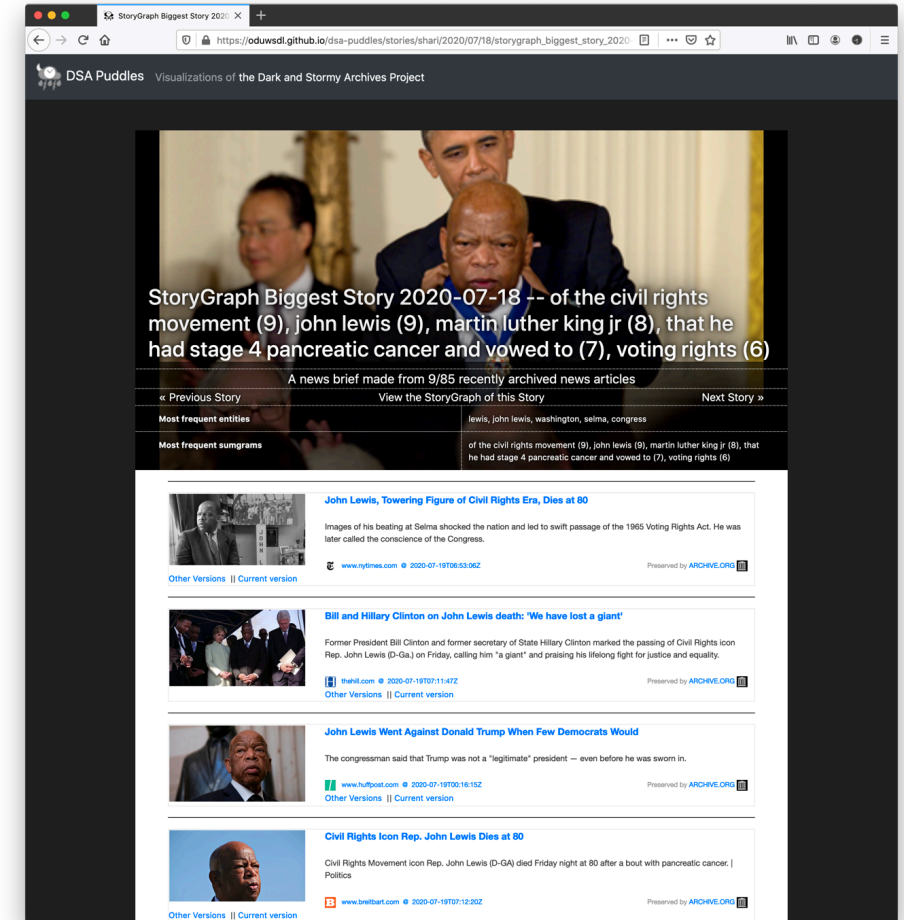
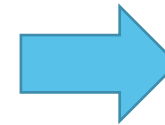
Hypercanon



MementoEmbed



Raintale



From this: news articles extracted by StoryGraph

To this: news articles of the biggest story of the day

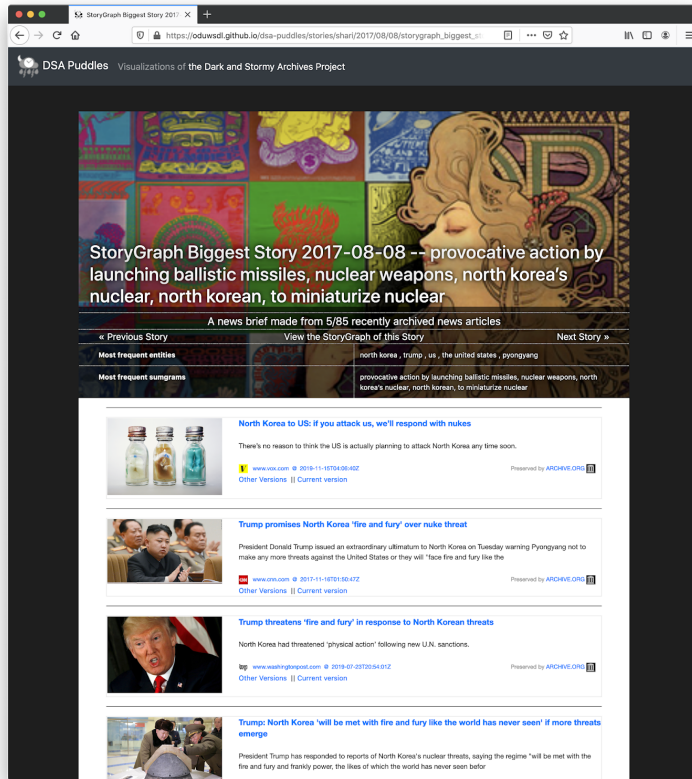


@shawnmjones @StormyArchives

Jones, S.M., Nwala, A.C., Klein, M., Weigle, M.C., & Nelson, M.L. (2020). SHARI -- An Integration of Tools to Visualize the Story of the Day. In Web Archiving and Digital Libraries Workshop (WADL 2020). arXiv, abs/2008.00137.

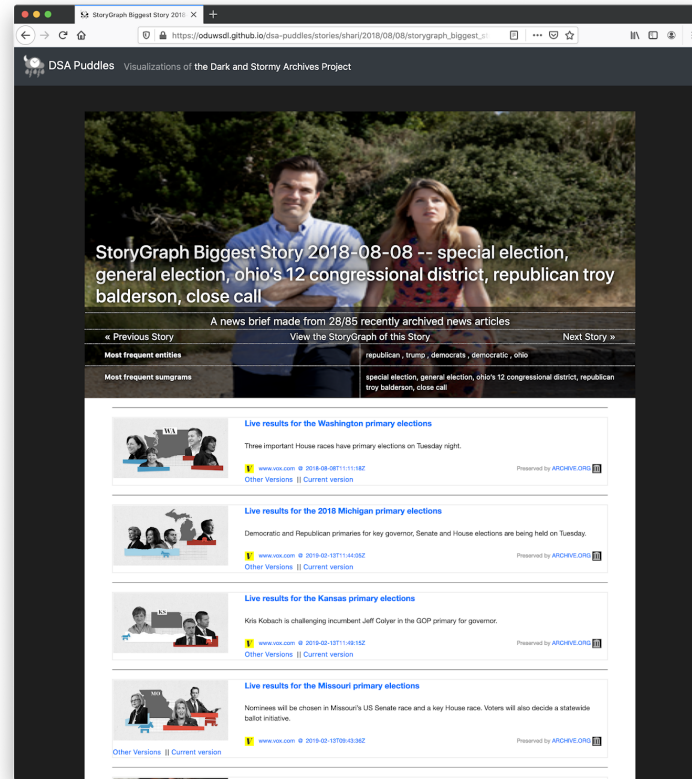
<https://arxiv.org/abs/2008.00137>

This allows us to review news from the same day in different years



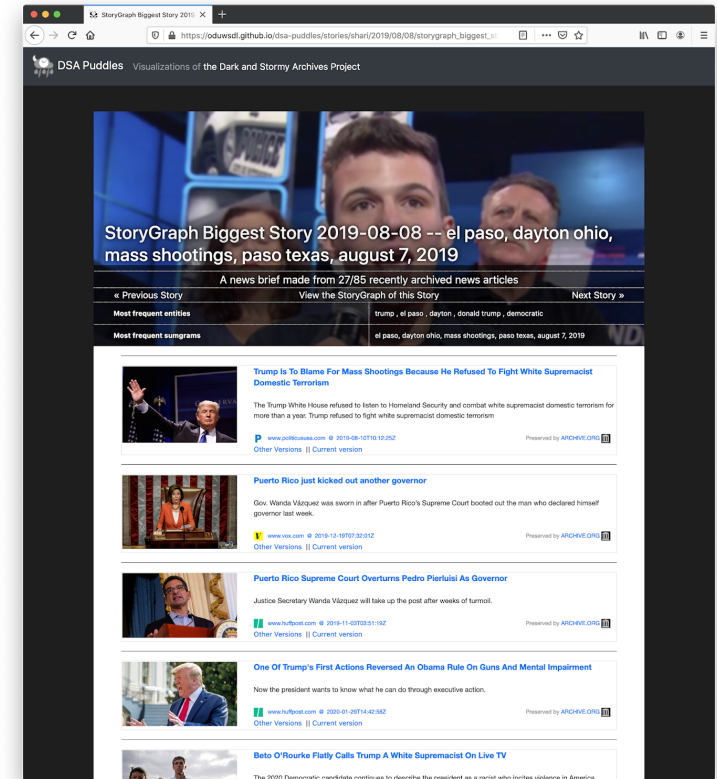
August 8, 2017

Nuclear Provocation by North Korea



August 8, 2018

US Primaries and Special Elections



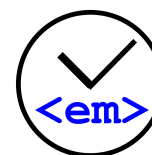
August 8, 2019

Aftermath of Shootings
at El Paso and Dayton



@shawnmjones @StormyArchives

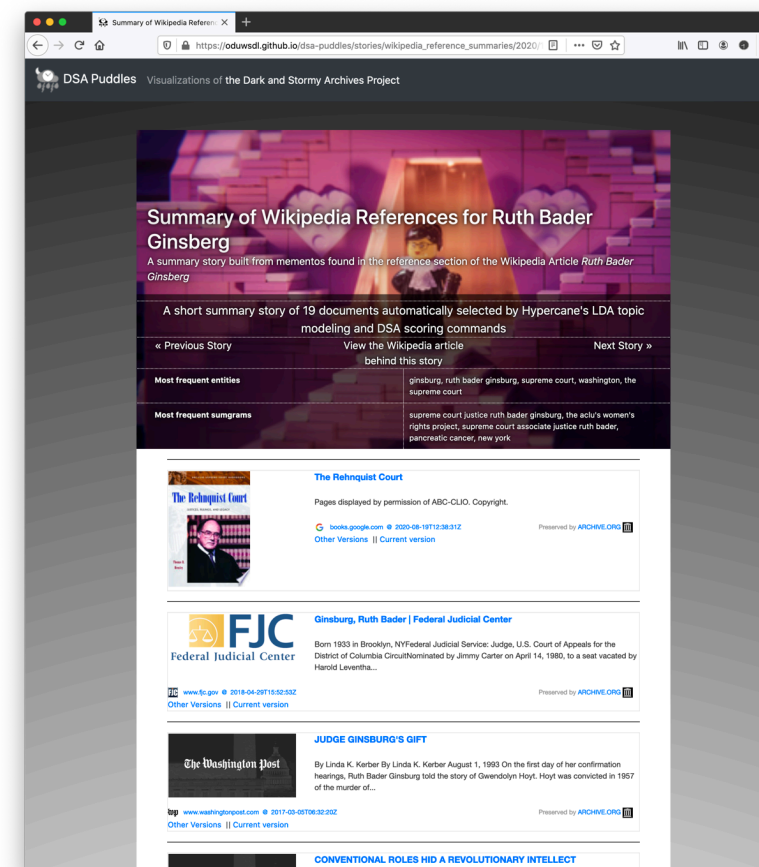
We can automatically summarize Wikipedia references



MementoEmbed



Raintale



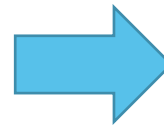
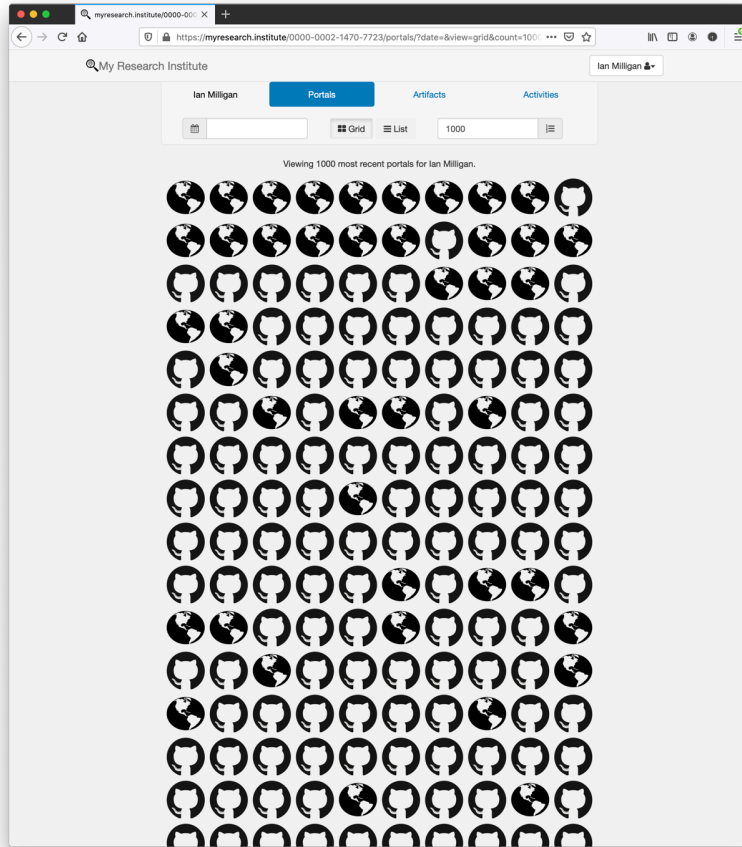
From this: Wikipedia references as mementos

To this: a sample providing new insight into the references not seen in the article



@shawnmjones @StormyArchives

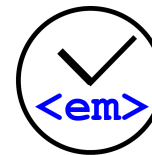
We can summarize a scholar's grey literature from the web archive collections of the Scholarly Orphans project, previously presented at CNI



myresearch.institute
Scraper



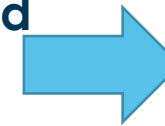
HyperCane



MementoEmbed



Raintale



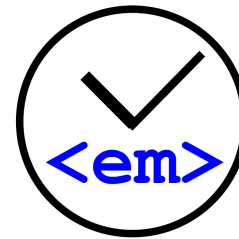
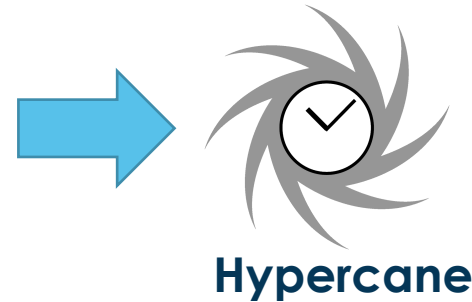
From this: 1000 mementos of grey literature from the Scholarly Orphans project

To this: 20 mementos representing a scholar's work

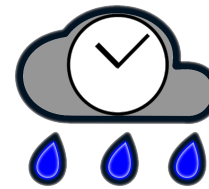


@shawnmjones @StormyArchives

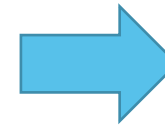
You can automatically summarize your web archive collection here



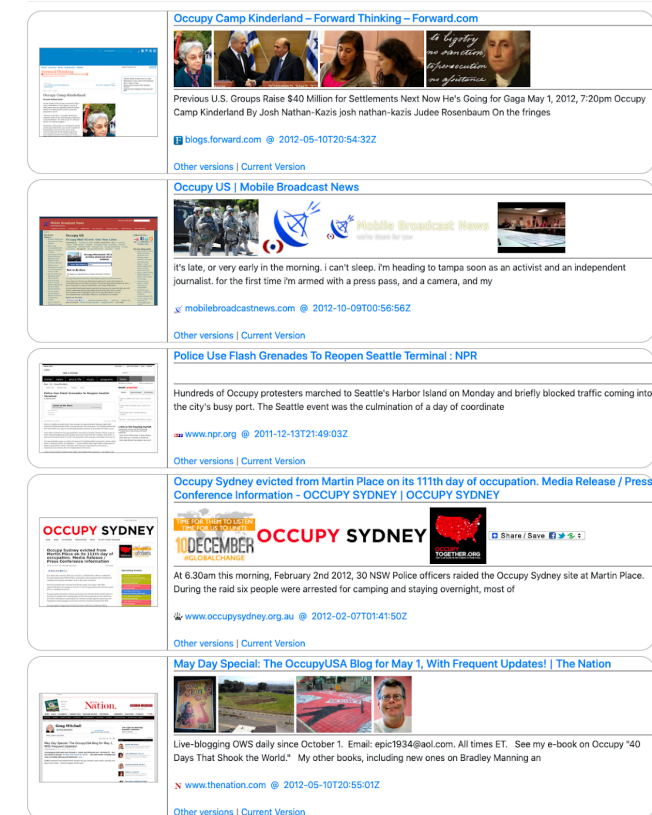
MementoEmbed



Raintale



Occupy Movement 2011/2012



From this: **your web archive collection**

To this: a small sample of mementos
visualized with your templates and features



@shawnmjones @StormyArchives

Jones, S.M., Klein, M., Weigle, M.C., & Nelson, M.L. (2020). MementoEmbed and Raintale for Web Archive Storytelling. In Web Archiving and Digital Libraries Workshop (WADL 2020). arXiv, abs/2008.00137. <https://arxiv.org/abs/2008.00137>

For More Information on Dark and Stormy Archives



- Dark and Stormy Archives Project
 - Website: <https://oduwsdl.github.io/dsa/>
 - Example visualizations: <https://oduwsdl.github.io/dsa-puddles/>



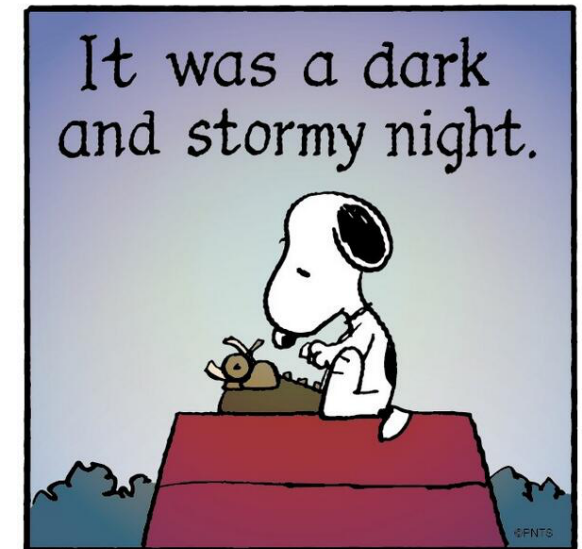
- Hypercane – Intelligently Sample Mementos
 - Website: <https://oduwsdl.github.io/hypercane/>
 - Documentation: <https://hypercane.readthedocs.io/en/latest/>
 - Source code: <https://github.com/oduwsdl/hypercane>



- Raintale – Tell Stories with Many Mementos
 - Website: <https://oduwsdl.github.io/raintale/>
 - Documentation: <https://raintale.readthedocs.io/en/latest/>
 - Report Issues: <https://github.com/oduwsdl/raintale/issues>
 - Technical Report: <https://arxiv.org/abs/2008.00137>



- MementoEmbed – Summarize Individual Mementos
 - Documentation: <https://mementoembed.readthedocs.io/en/latest/>
 - Report Issues: <https://github.com/oduwsdl/MementoEmbed/issues>
 - Technical Report: <https://arxiv.org/abs/2008.00137>





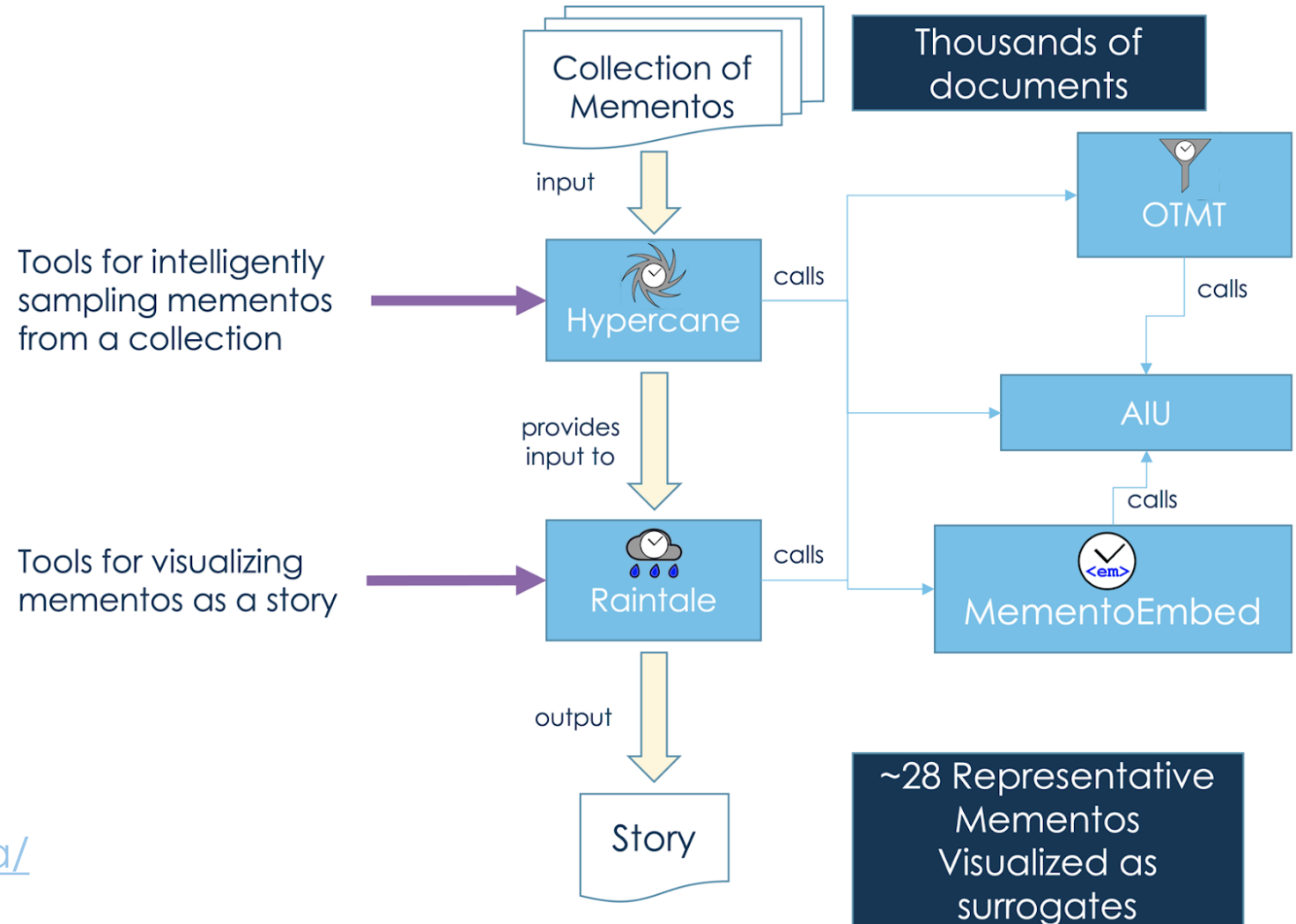
Backup slides



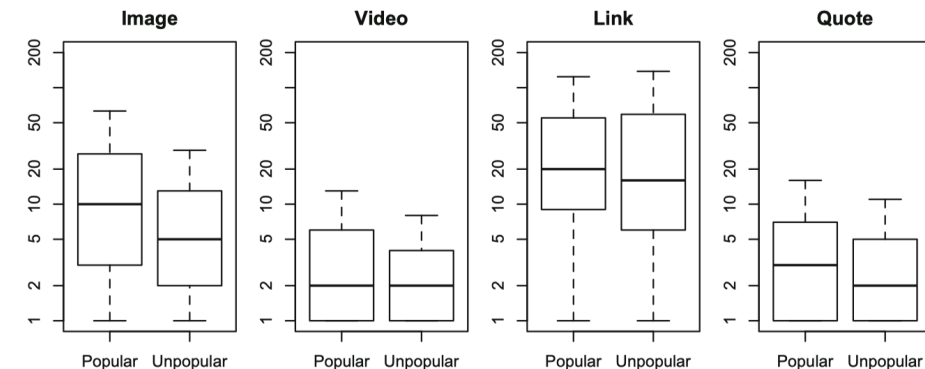
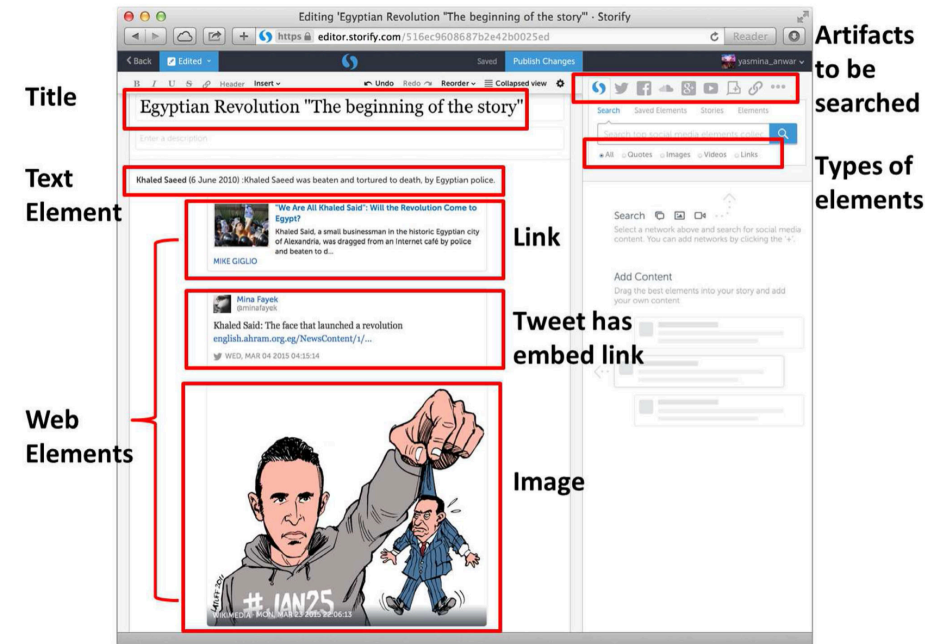
The Dark and Stormy Archives Toolkit



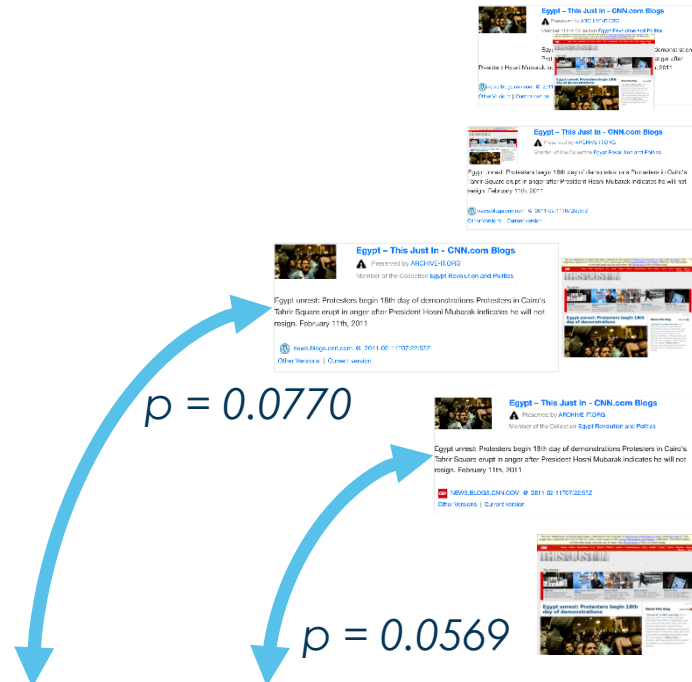
<https://oduwsdl.github.io/dsa/>



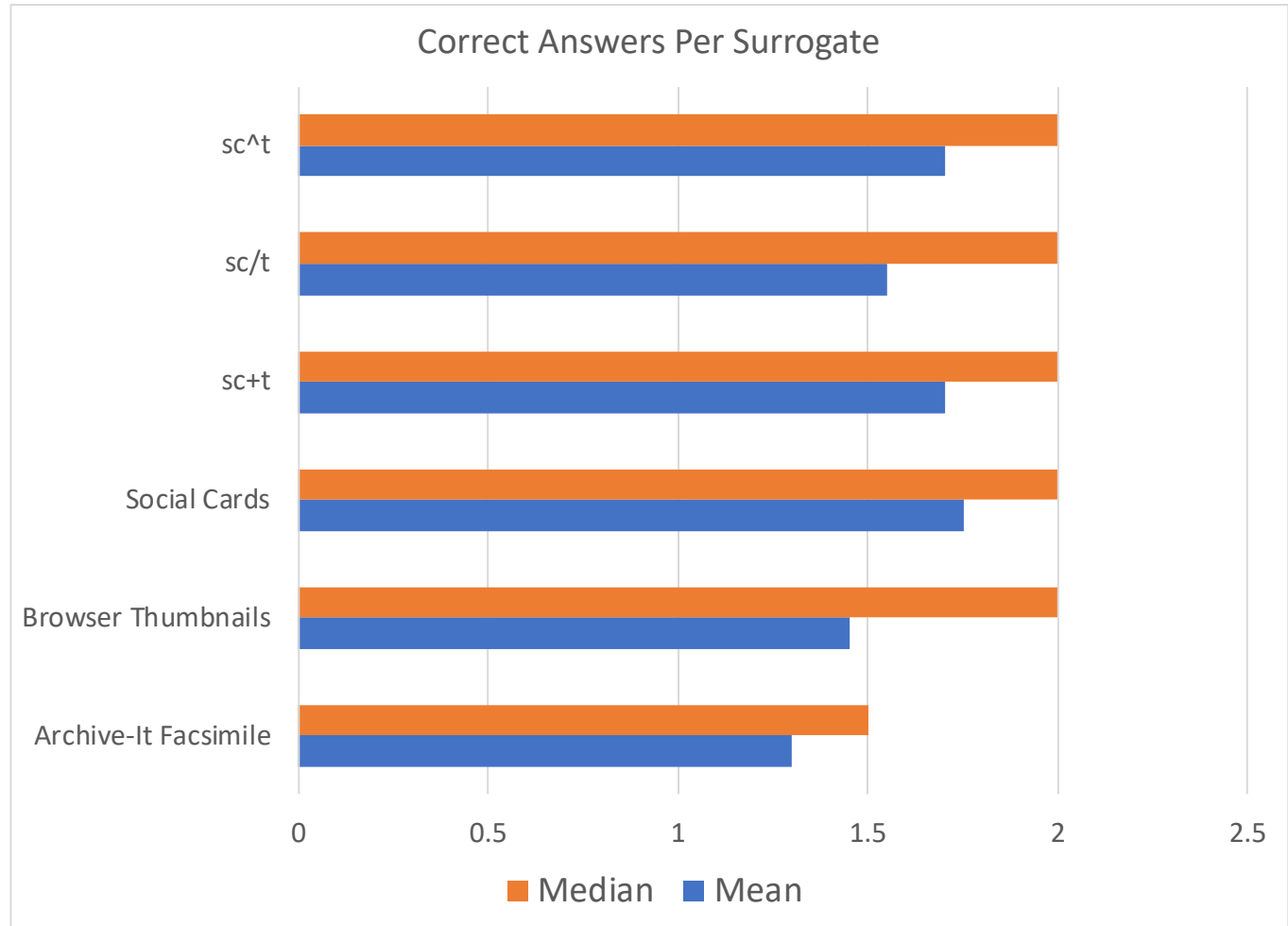
AlNoamany discovered that popular social media stories contain around 28 elements



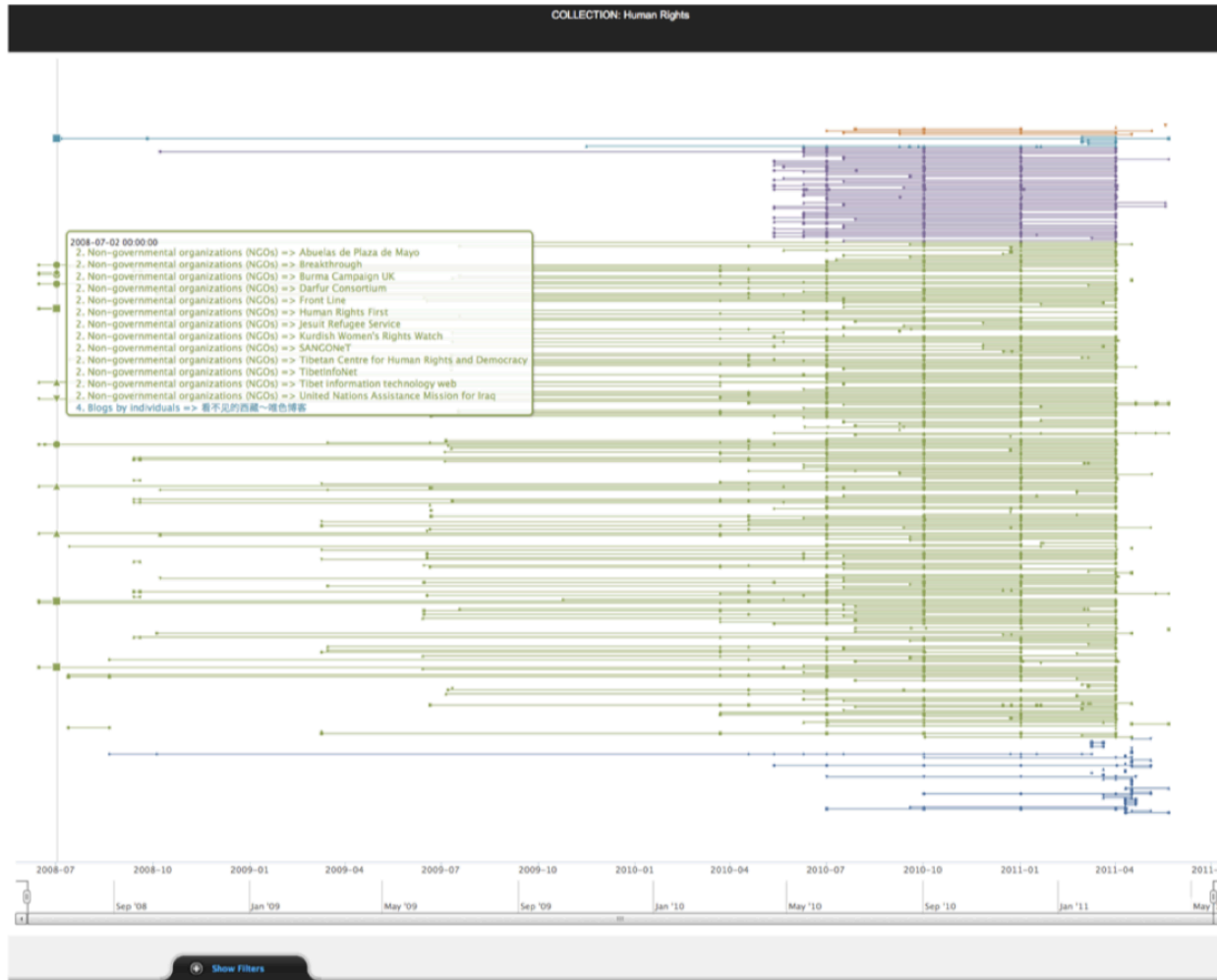
Through a user study we discovered that social cards probably provide for better understanding of web archive collections



URL: <http://news.blogs.cnn.com/category/world/egypt-world-latest-news/>
Captured 41 times between Feb 1, 2011 and Jan 4, 2012



Maybe timelines are best due to the temporal nature of web archives...



What do these colors mean?
If I hover over one, I get a group of titles?

How do I see across timelines?

What timeline do I pick first?

Which timelines are useful?

Why do so many stop at the same point?

Where are the seeds?

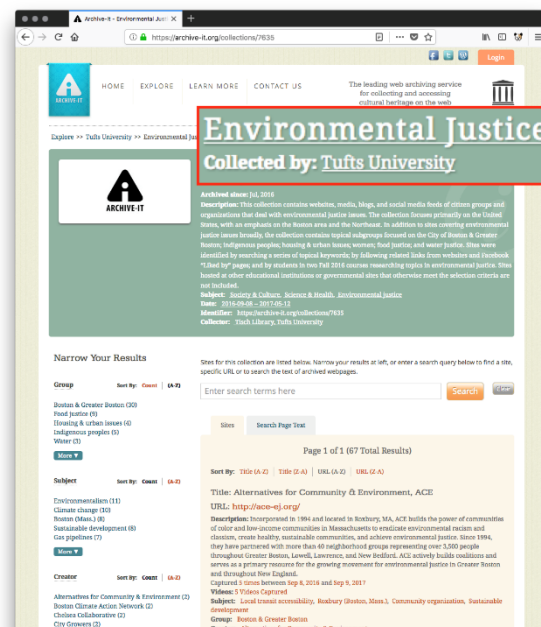
Where are the mementos?



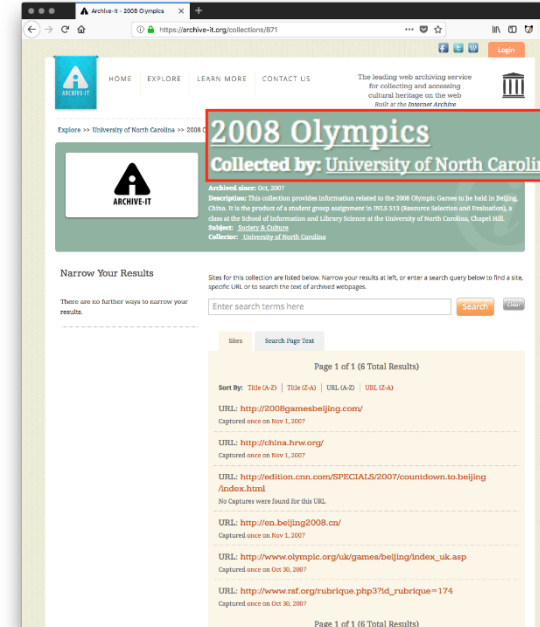
We discovered four categories in Archive-It collections...



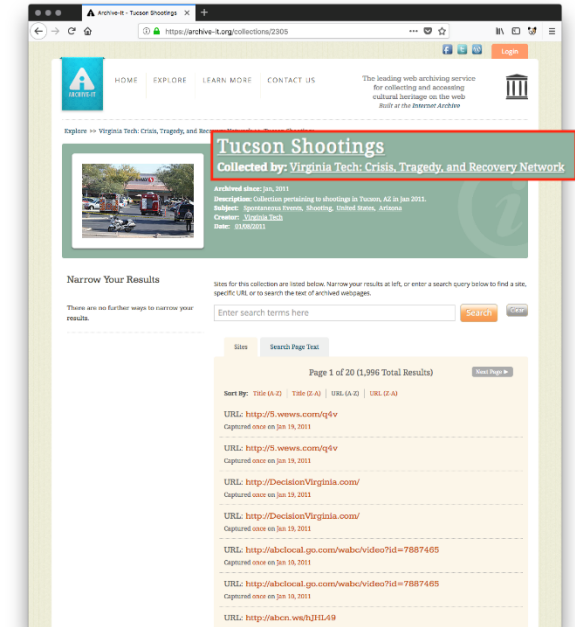
Self-Archiving
54.1% of collections



Subject-based
27.6% of collections



Time Bounded – Expected
14.1% of collections



Time Bounded – Spontaneous
4.2% of collections

In a study of 3,382 Archive-It collections



@shawnmjones @StormyArchives

S. M. Jones, A. Nwala, M. C. Weigle, and M. L. Nelson.
2018. The Many Shapes of Archive-It. In *International Conference on Digital Preservation (iPRES) 2018*.
<https://doi.org/10.17605/OSF.IO/EV42P>

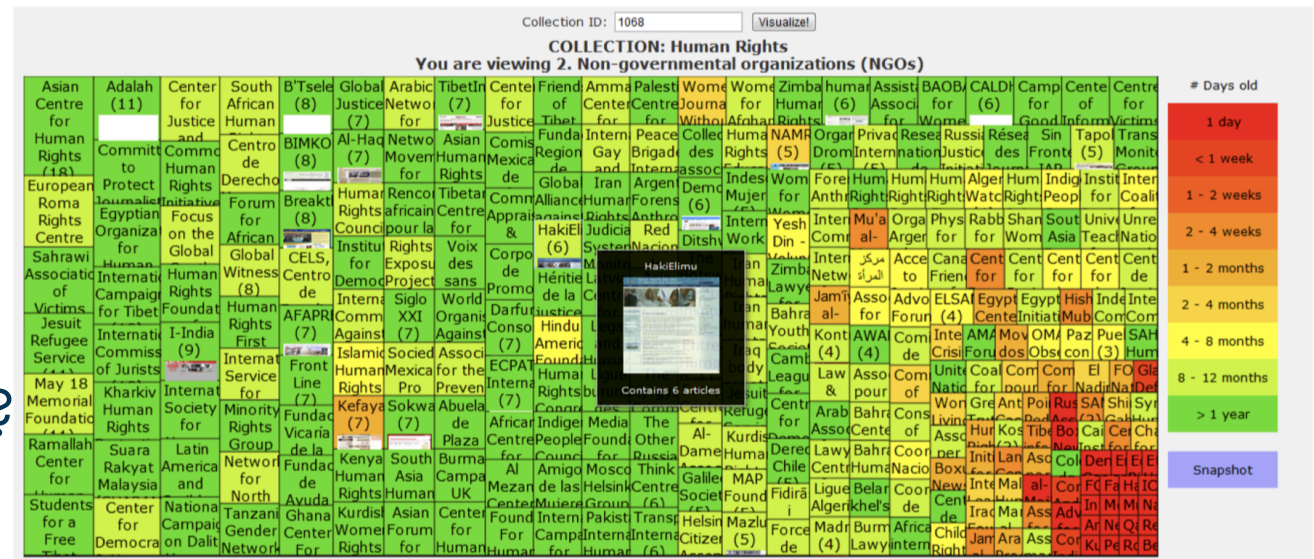
Thousands of mementos of different pages?

Maybe heatmaps with colors are best for conveying what this collection is about...

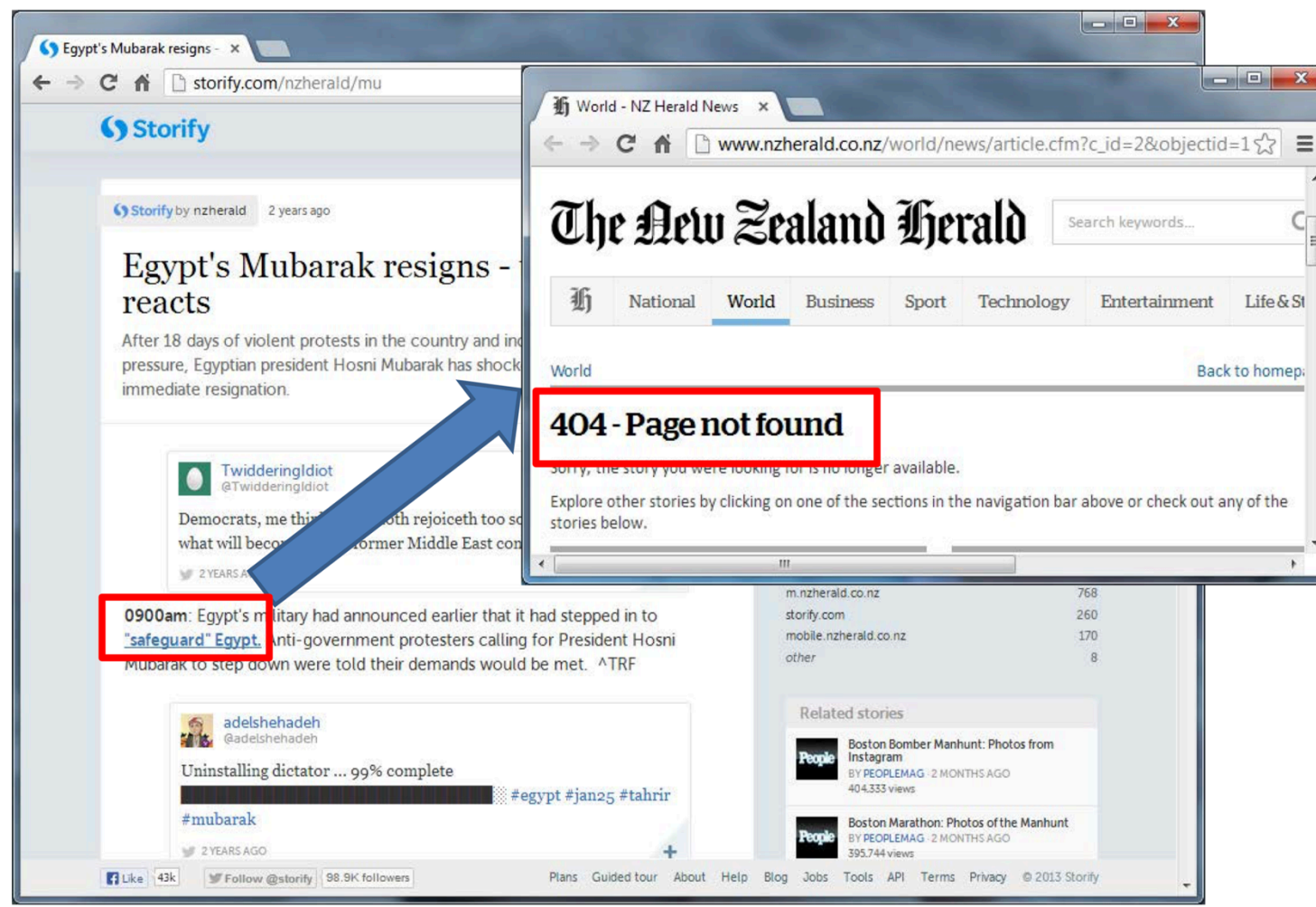


What do the colors mean?
Why are the differently sized blocks?
Do I need to read the thumbnail?

Are these times based on today?
Is this the whole collection?
Do I need to explore each block?



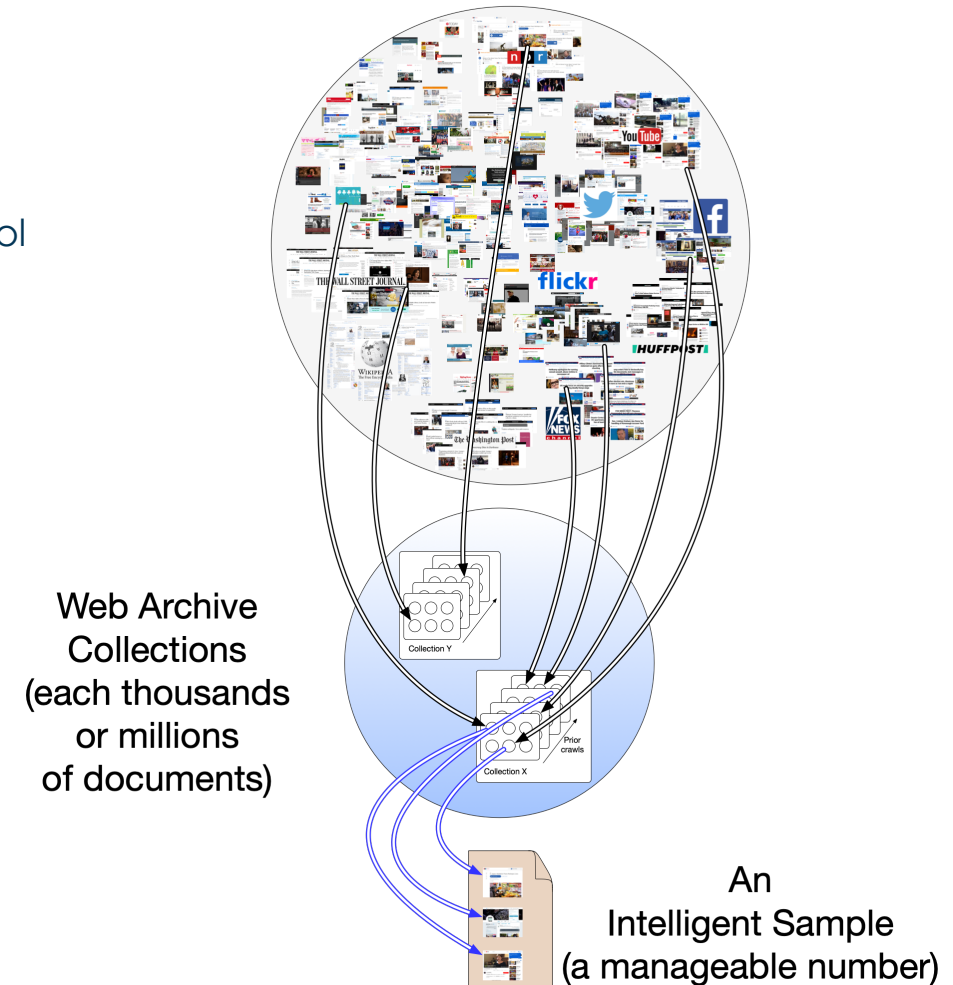
Existing storytelling services are bookmarking, not preserving!



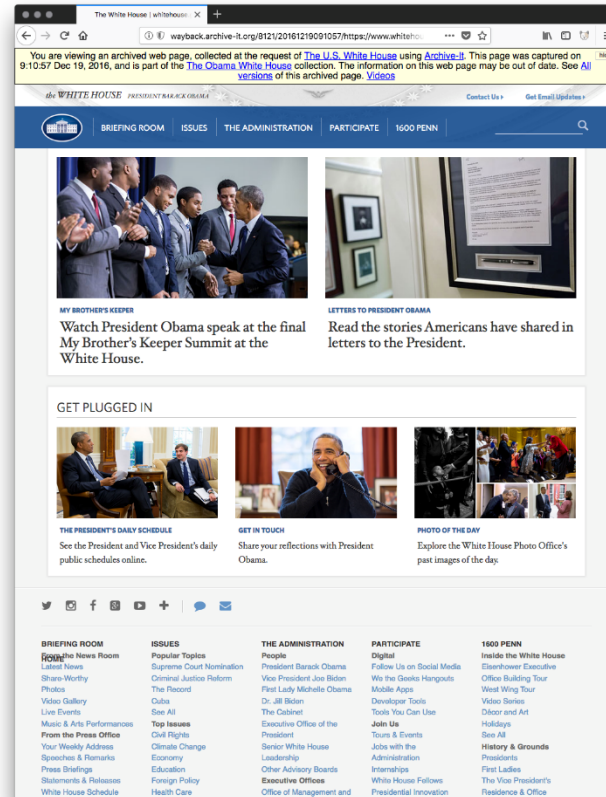
Hypercanes provides intelligent sampling of web archive collections

- **Works with Memento-compliant archives**
- Discovers mementos from TimeMaps and vice versa
- Creates new mementos with the help of ArchiveNow
- Can re-synthesize parts of web archives via the Memento Protocol
- Reports on:
 - collection metadata
 - named entities
 - terms
 - image analysis
- Synthesizing mementos into WARCs or other formats
- Clustering mementos by different algorithms
- Filtering mementos by off-topic, language, keyword matching, and more
- Scoring mementos by different functions
- Ordering mementos by features
- Currently a command-line application

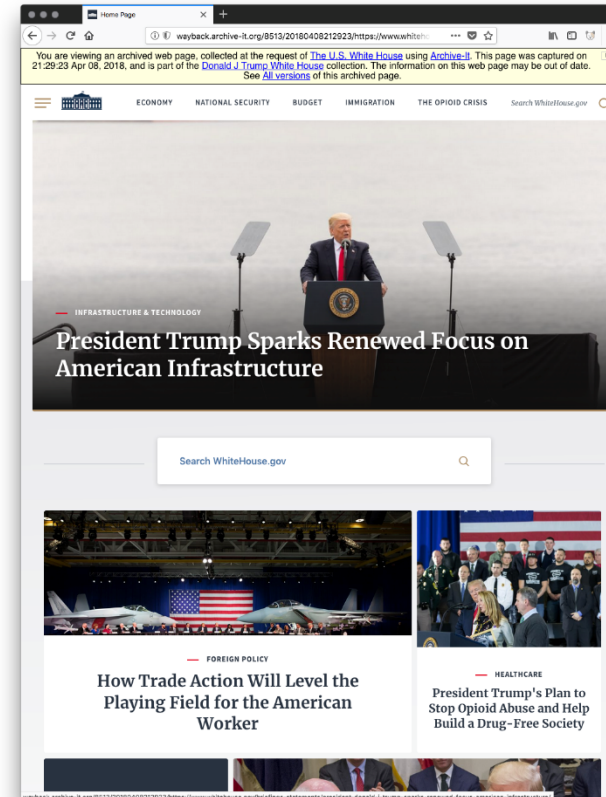
The Web in the “perpetual now”
(near ∞ documents)



Different versions allow us to see changes in an organization's web presence



The White House: 2016



The White House: 2018



Storytelling in social media consists of sampling and arranging web resources for summarization

