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
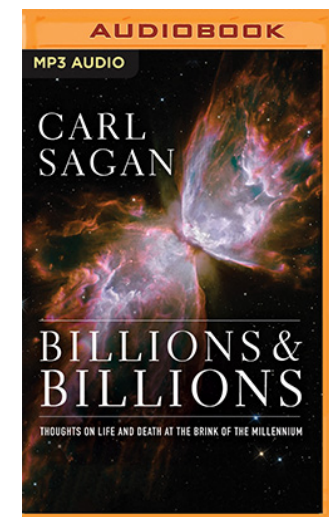
CNI Fall 2020 Meeting

Thanks to:

The abundance of data in web archives

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...)   Go @internetarchive !

> **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

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.

Government of Canada Publications
Collected by: Canadian Government Information PLN Web Archive
Archived since: Feb, 2013
No description.

ARCHIVE-IT

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.

Creator   Sort By: **Count** | (A-Z)
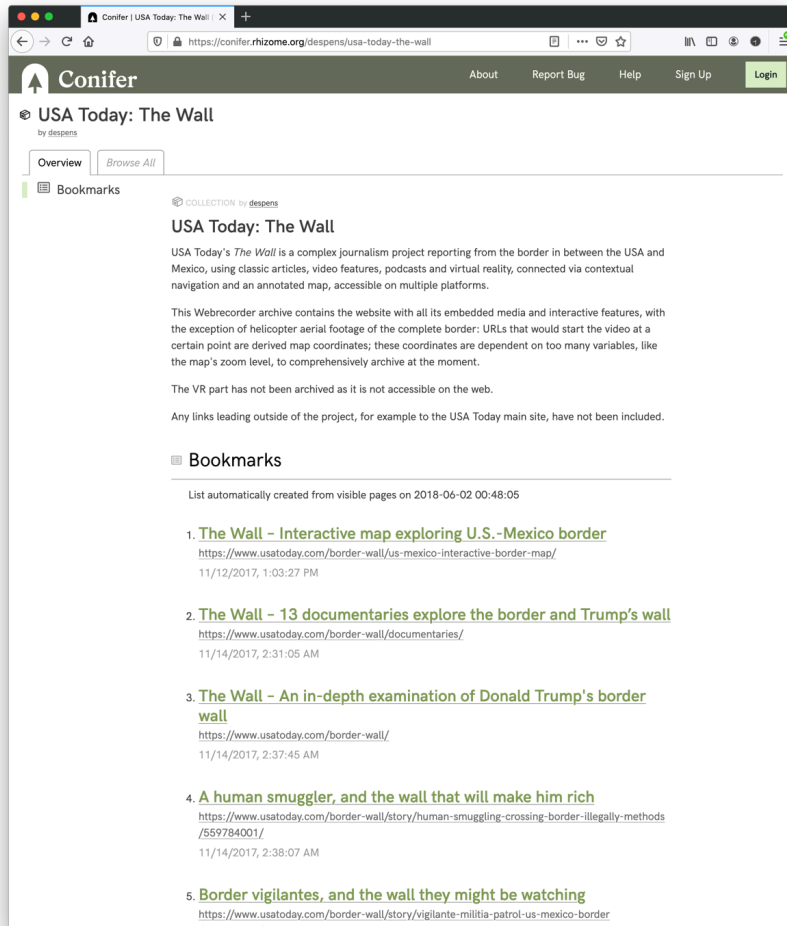
Enter search terms here   Search   Clear

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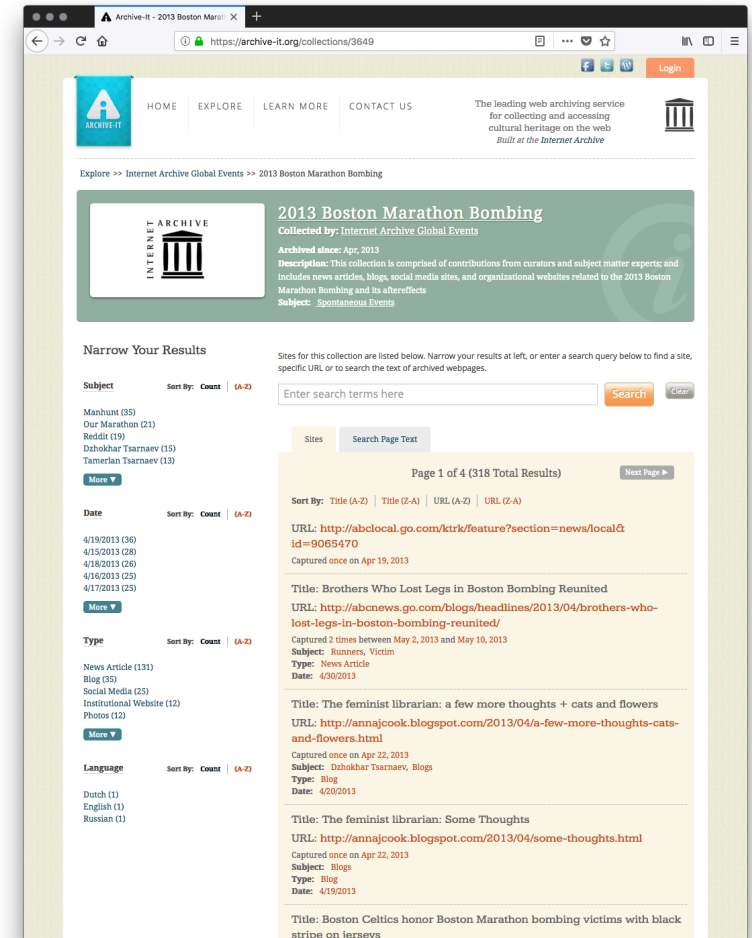More ▼

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.

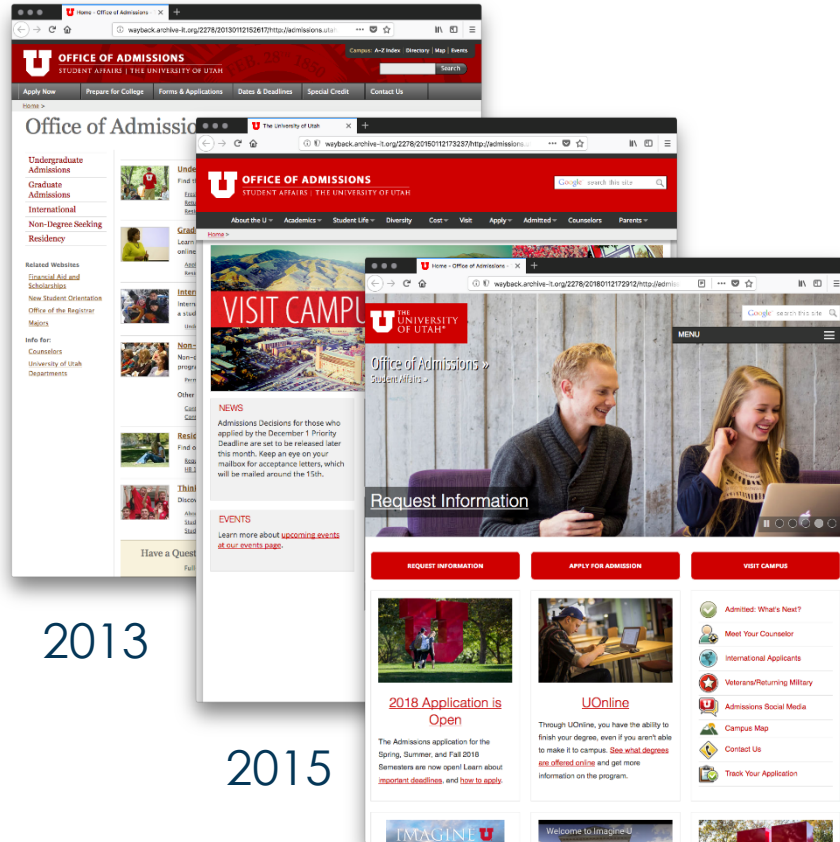The mementos are the documents in our collections.

Collections with Conifer/Webrecorder

Collections with Archive-It

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
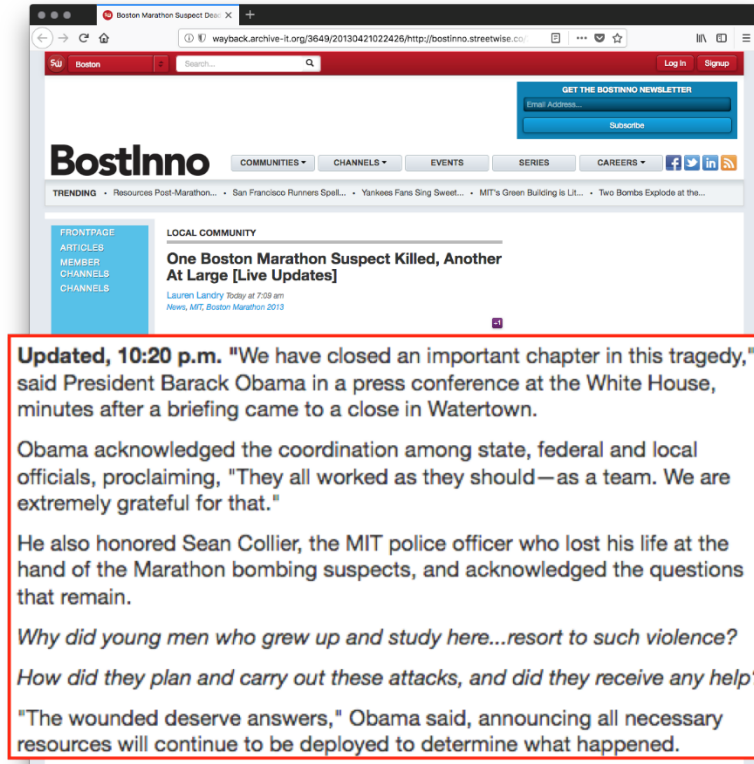from the #blacklivesmatter Collection

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

**Memento from
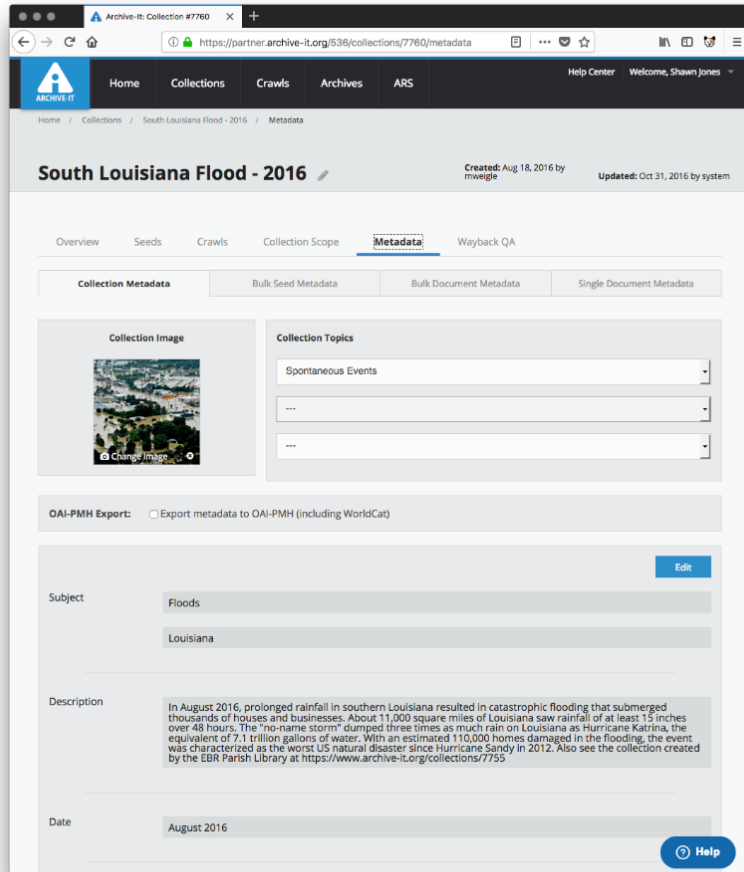April 19, 2013 17:12**
Searching for suspects,
City on lockdown
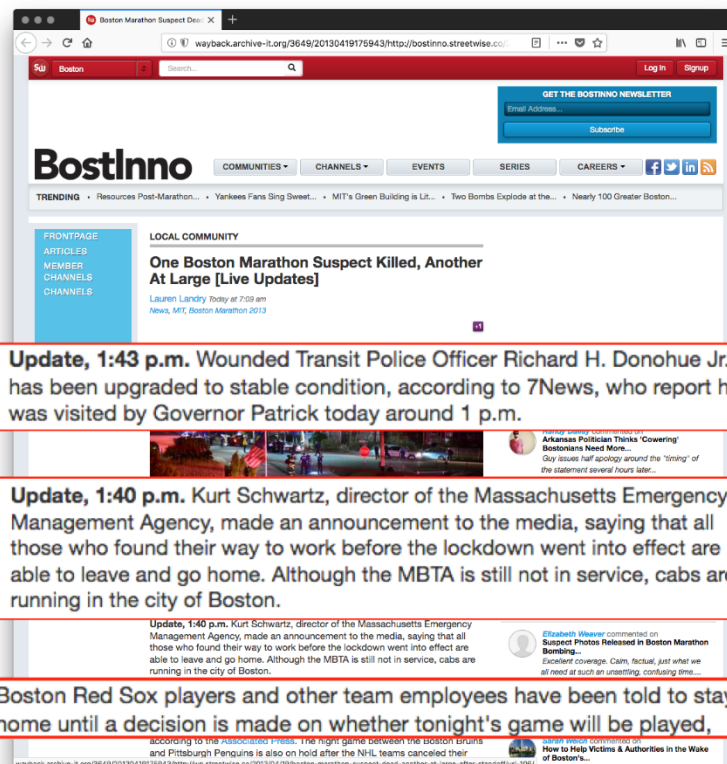
**Memento from
April 19, 2013 17:59**
Officer Donahue in hospital,
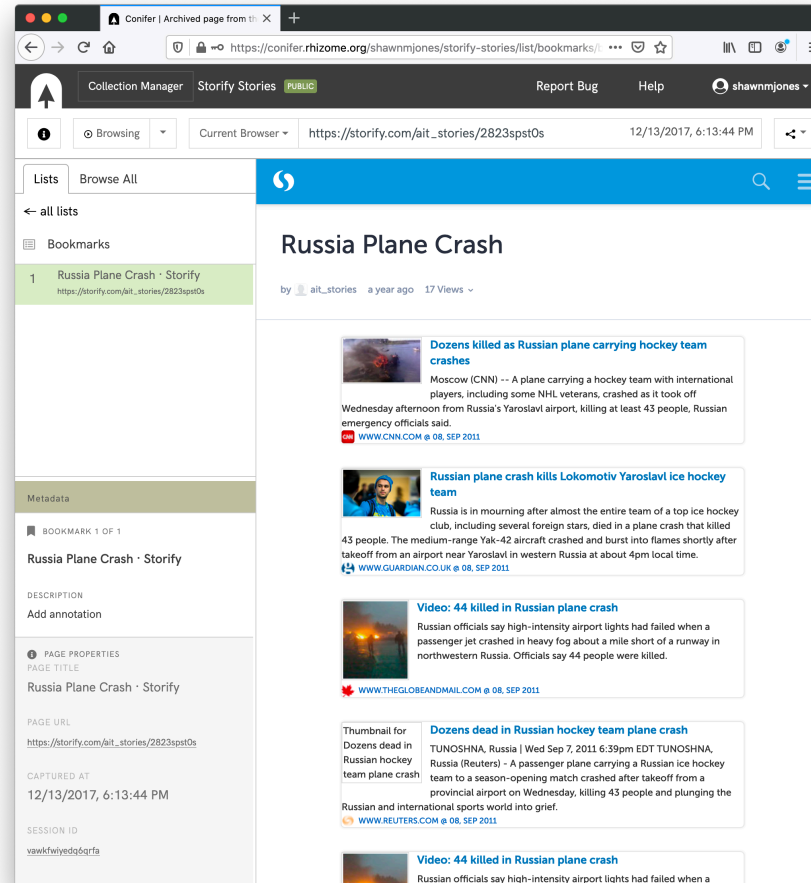Lockdown loosened,
Will the Red Sox game be cancelled?

**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

… 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?

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.
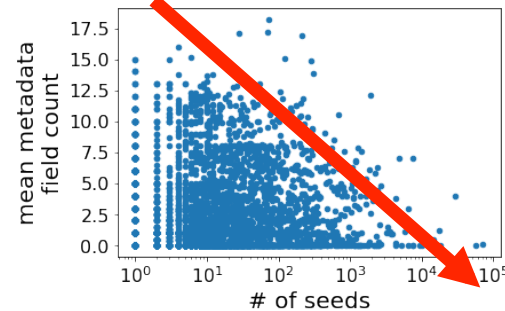
9 seeds
with metadata

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

11

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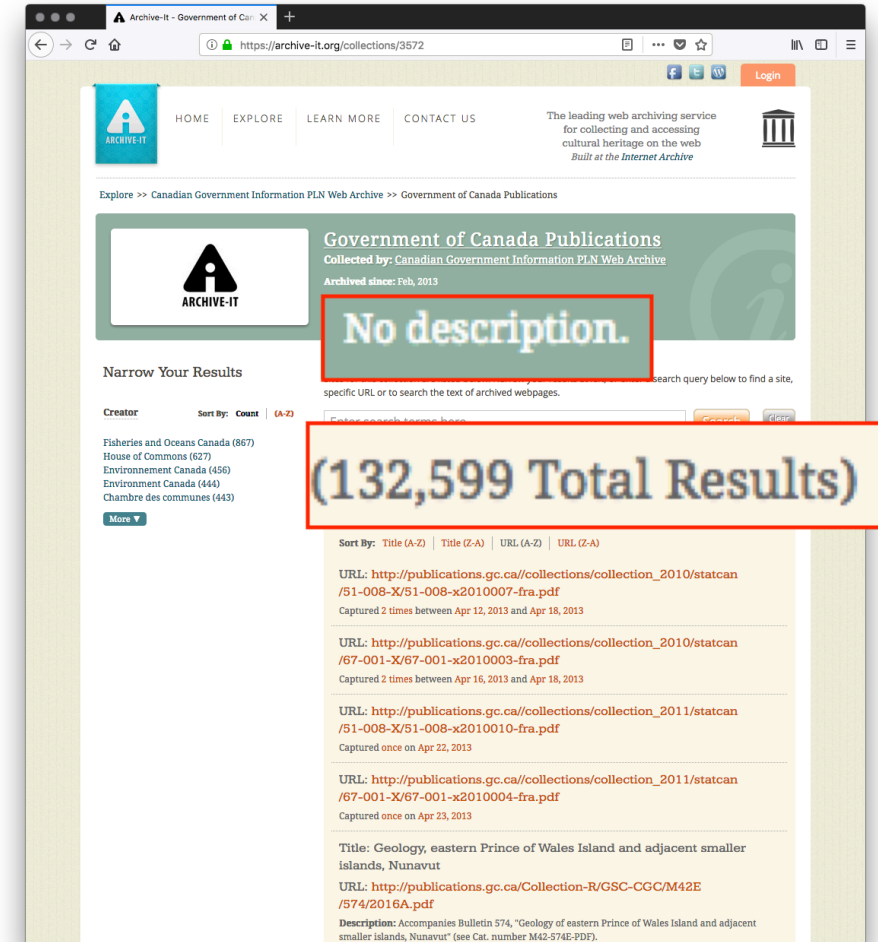
This collection has 132,599 seeds, many with multiple mementos

In addition to more mementos in web archives, more Archive-It collections are added every year
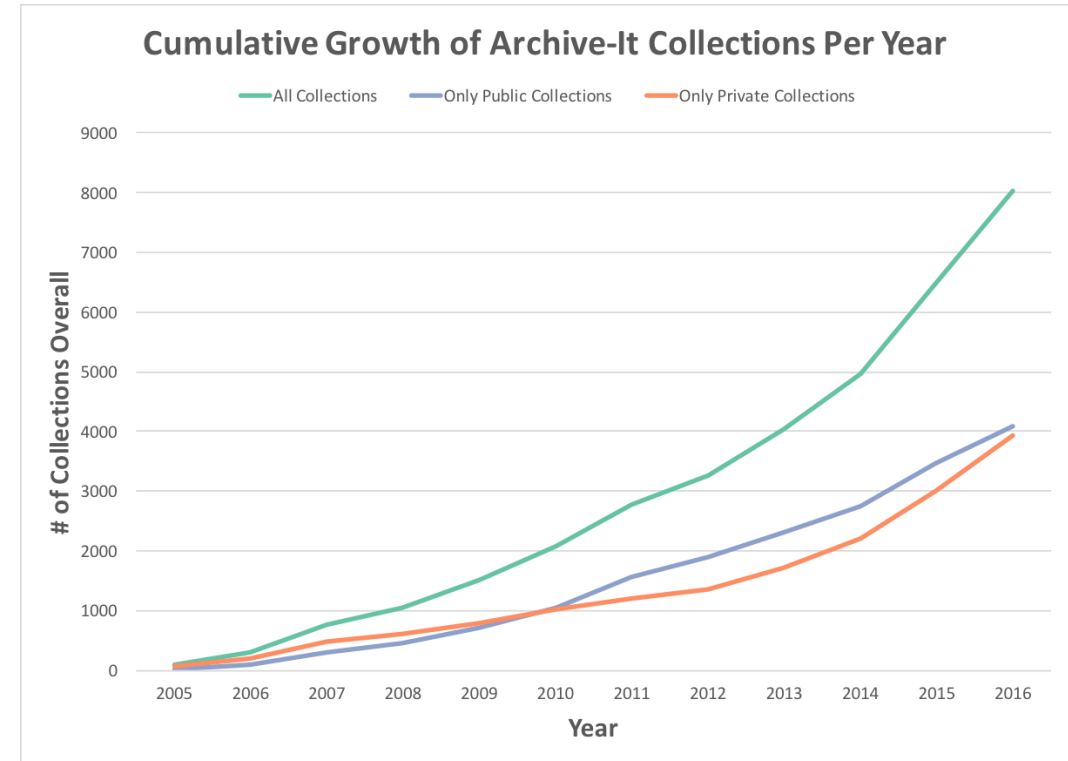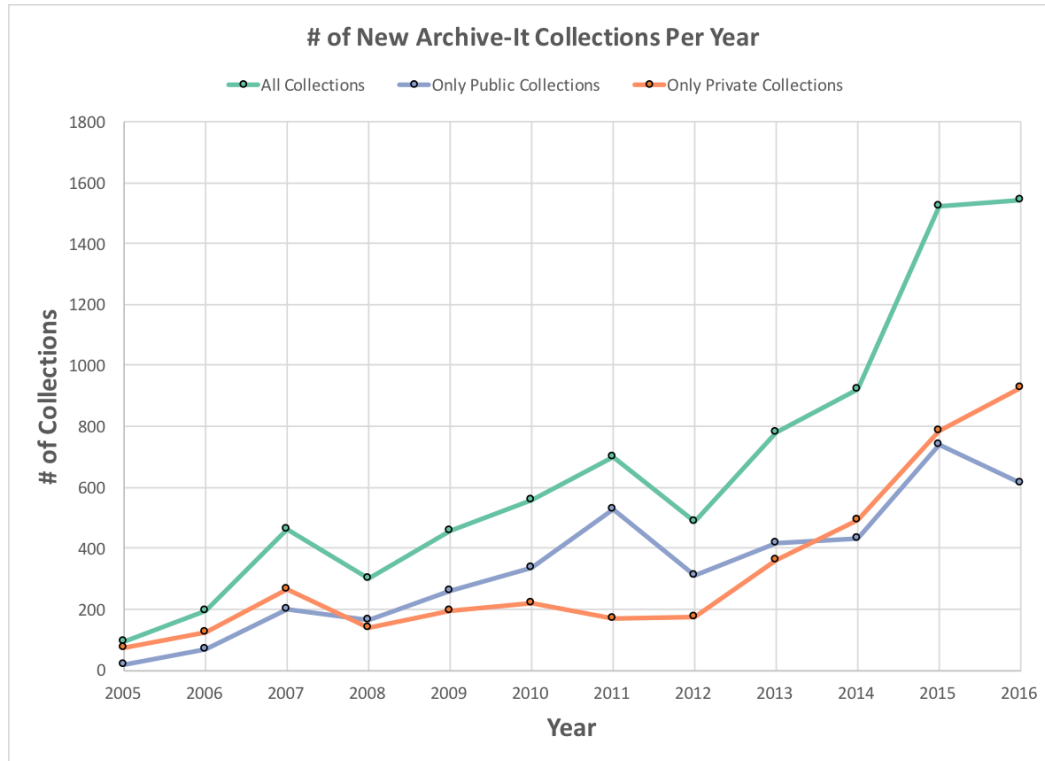
More than 15,000 collections exist

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.*
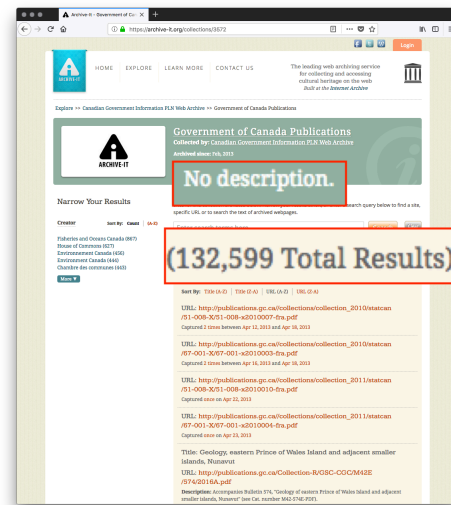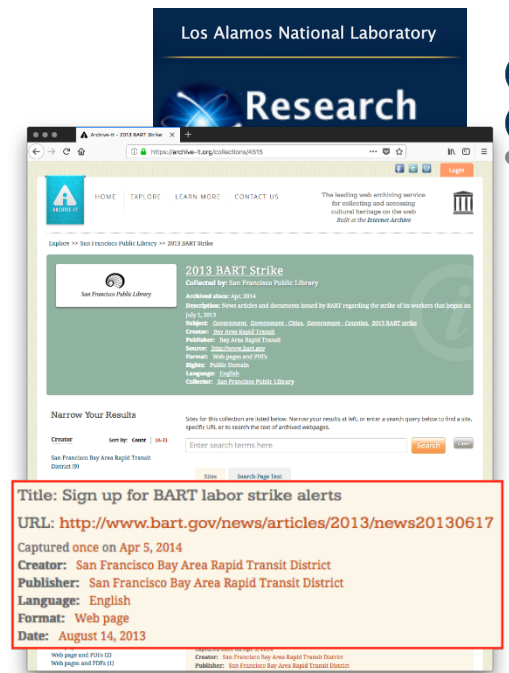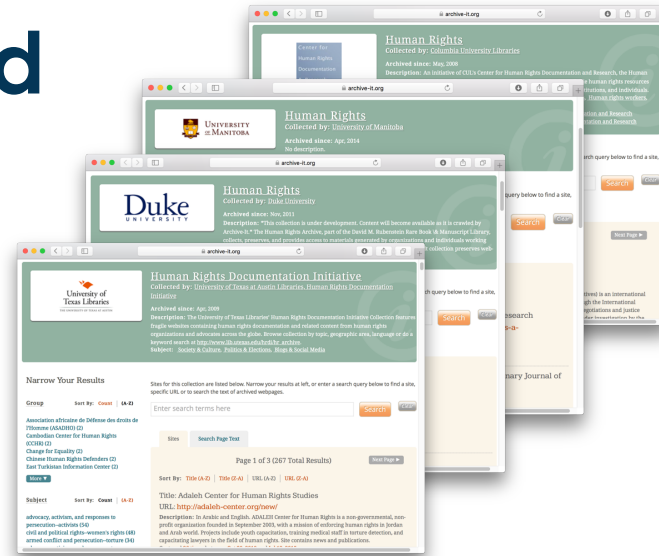
Existing Solutions
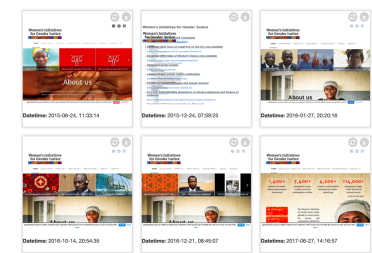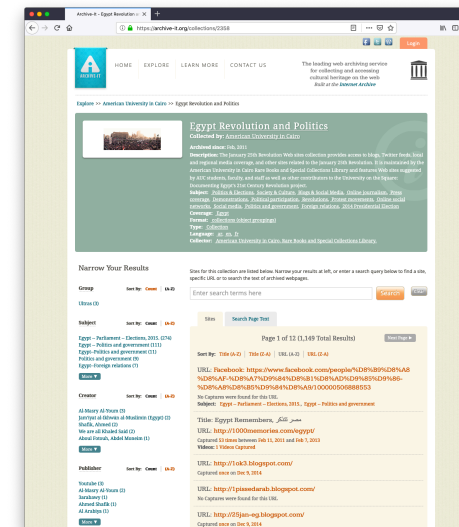
Ideally, a user would be able to glance at a visualization and gain understanding of the collection
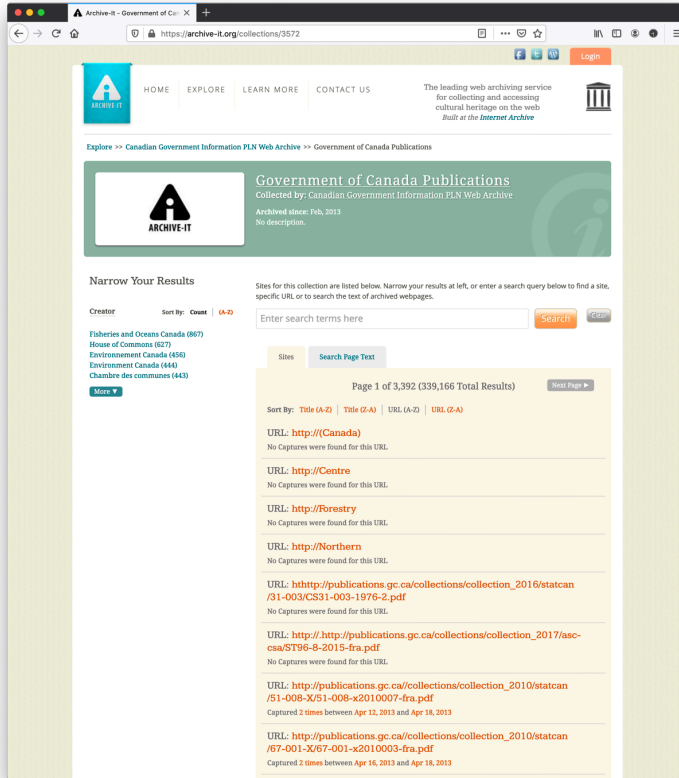
Web Archive Collection
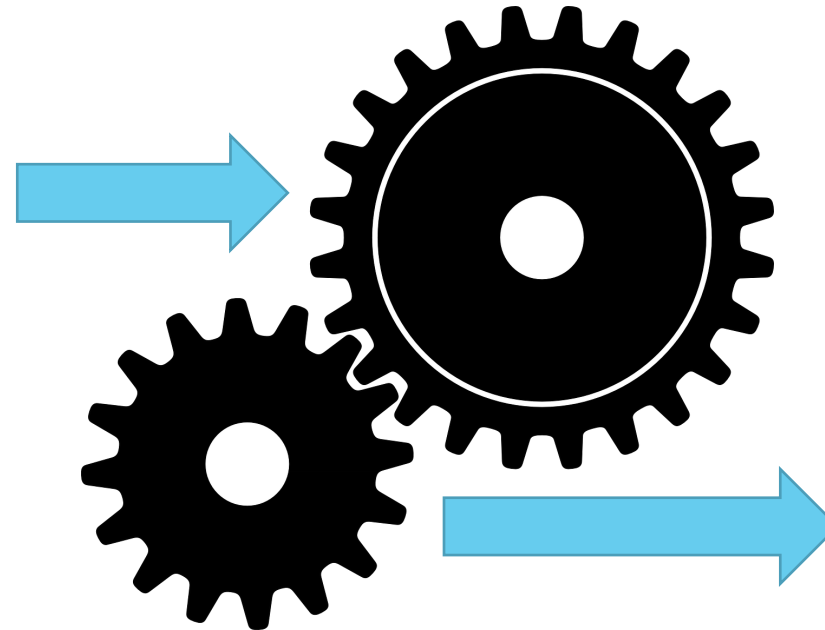
Automated Solution

Visualization that conveys understanding at a glance

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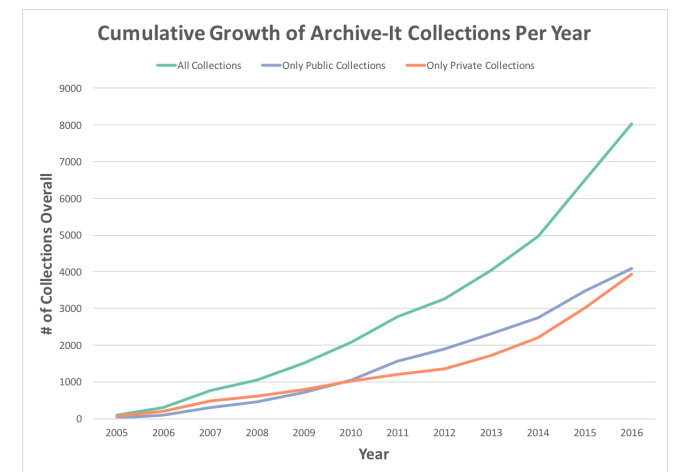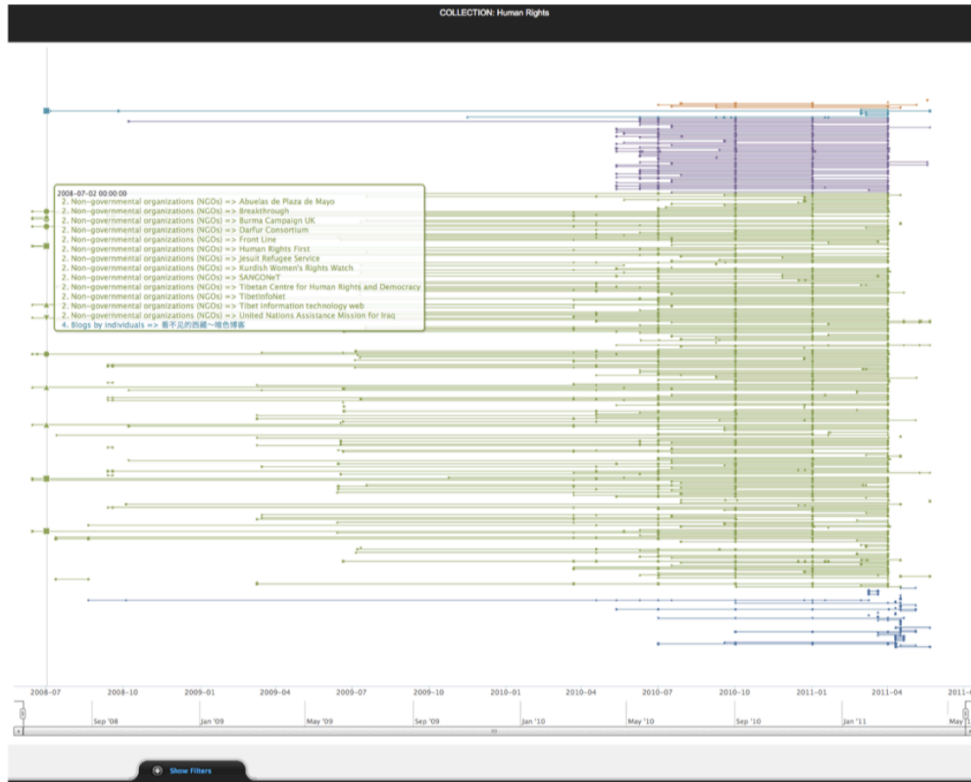

Timelines provide an overview, but do not provide understanding at a glance.
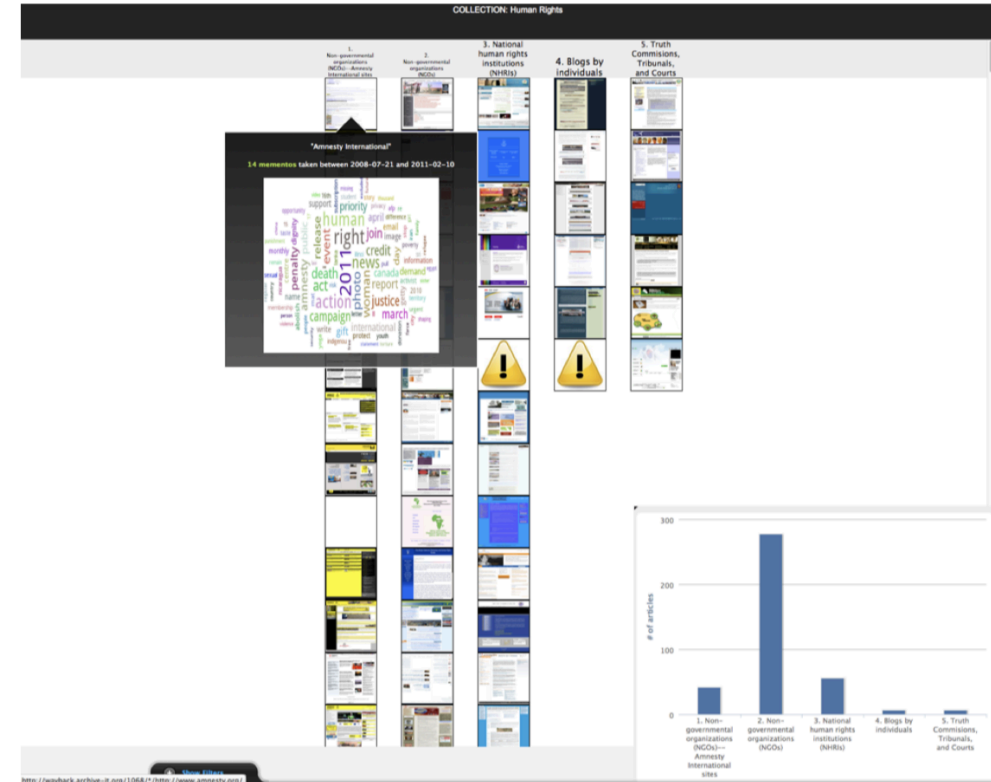
Thumbnails work well for small collections, but do not scale.

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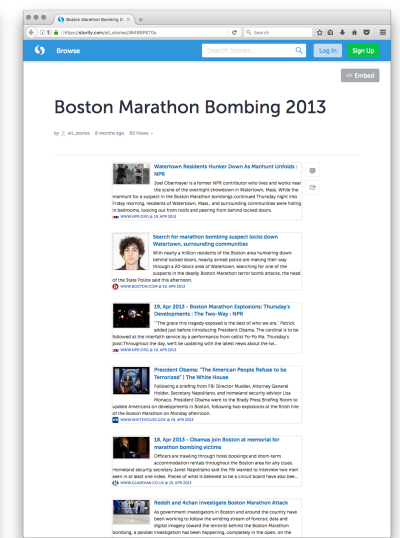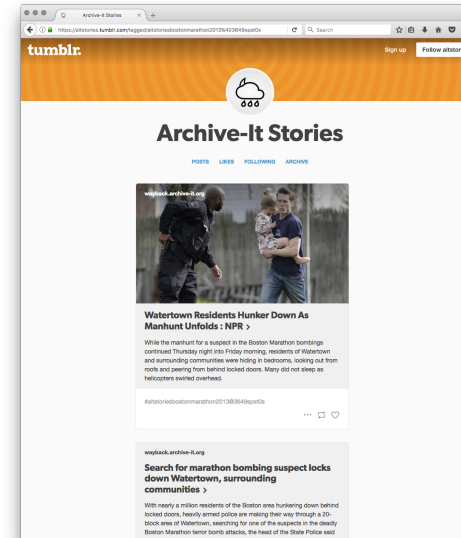
Web surrogates provide a visual summary of a web resource drawn from the content of the resource

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)

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

Archived 24 Oct 2008   Archived 24 Nov 2008   Archived 24 Dec 2008   Archived 24 May 2009   Archived 24 Jul 2009

Browser Thumbnail (example from UK Web Archive)

www.loc.gov
the library of congress
2,717,124   278,906   3,868   81
25,124,927 web captures from 1996 to 2017
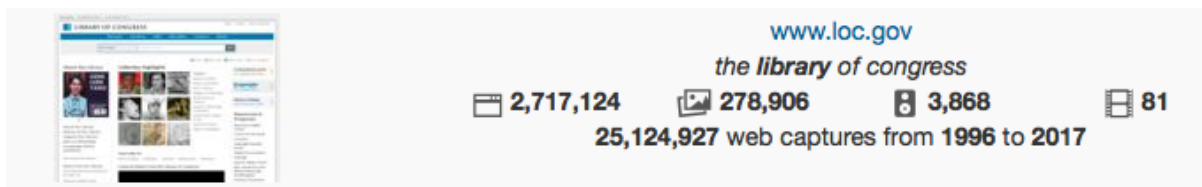
Text + Thumbnail (example from Internet Archive)

2017-08-11: Where Can We Post Stories Summarizing Web Archive Collections?
A social card generated by Facebook for my previous blog post. Rich links, snippet , social snippet, social media card , Twitter car...
WS-DL.BLOGSPOT.COM

Social Card (example from Facebook)

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:

Old Dominion University to Los Alamos National Laboratory

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

GOOGLE.COM

§ 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

Twitter

Tumblr

Bing

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**.
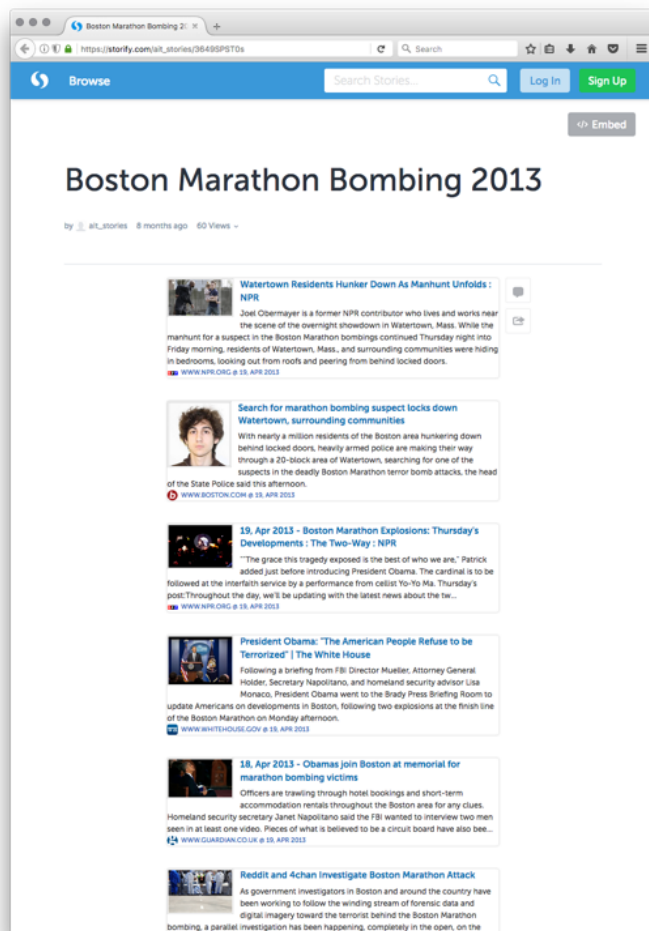
6 resources are shown in this Storify story
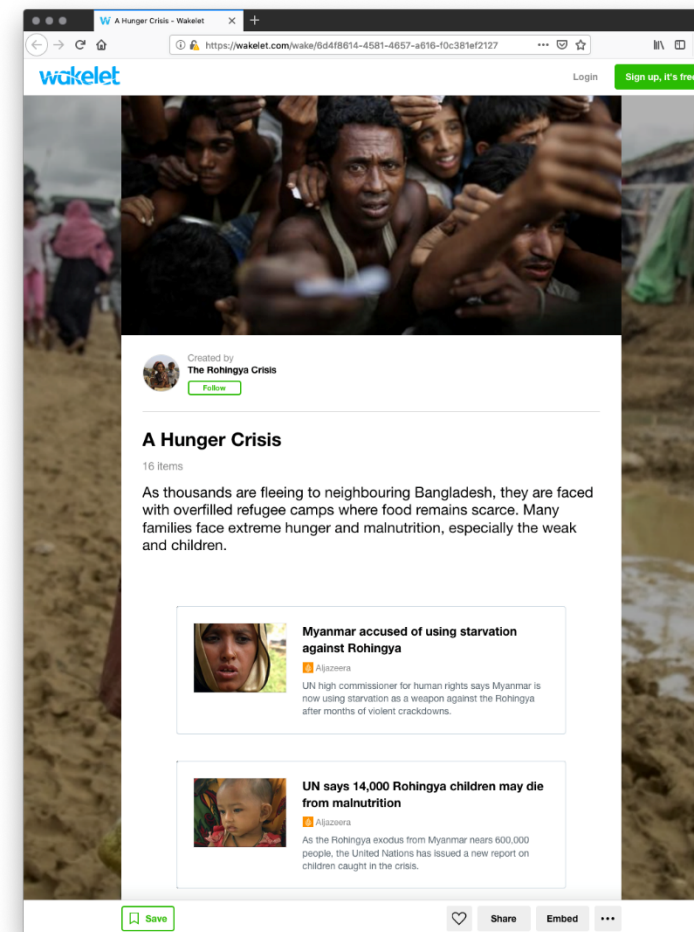
2 resources are shown in this Wakelet story

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:
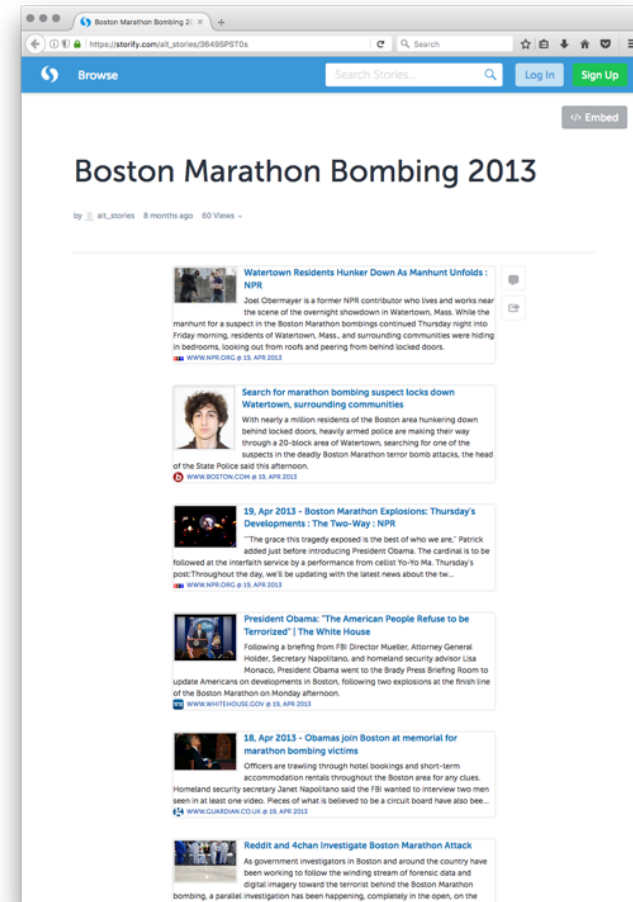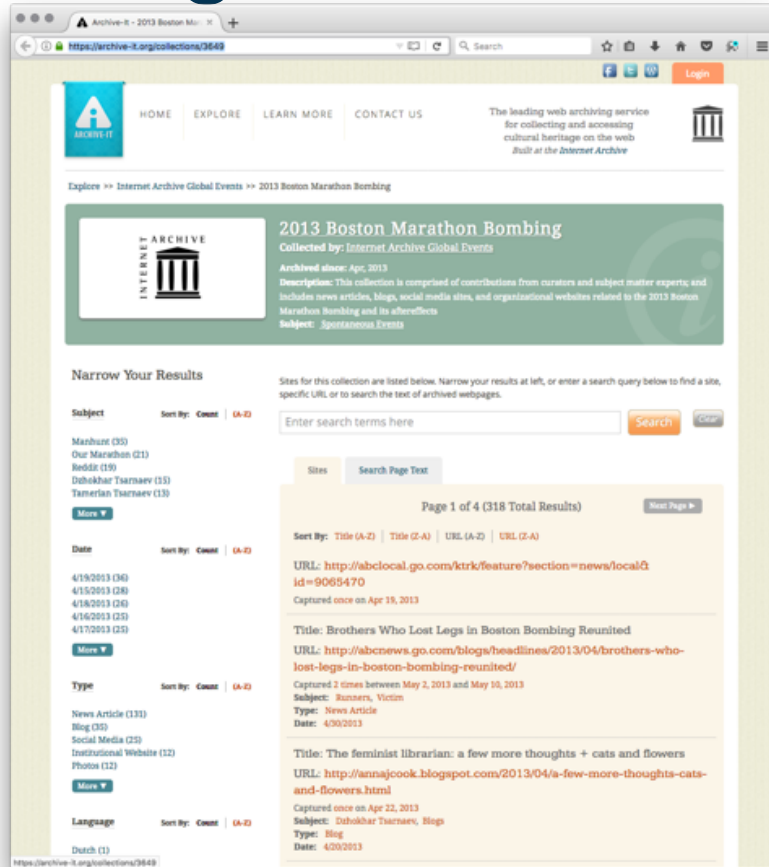318 seeds with
2421 mementos

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

Storytelling With Web Archives Has Multiple Use Cases…

1. **Promotion of the collection**
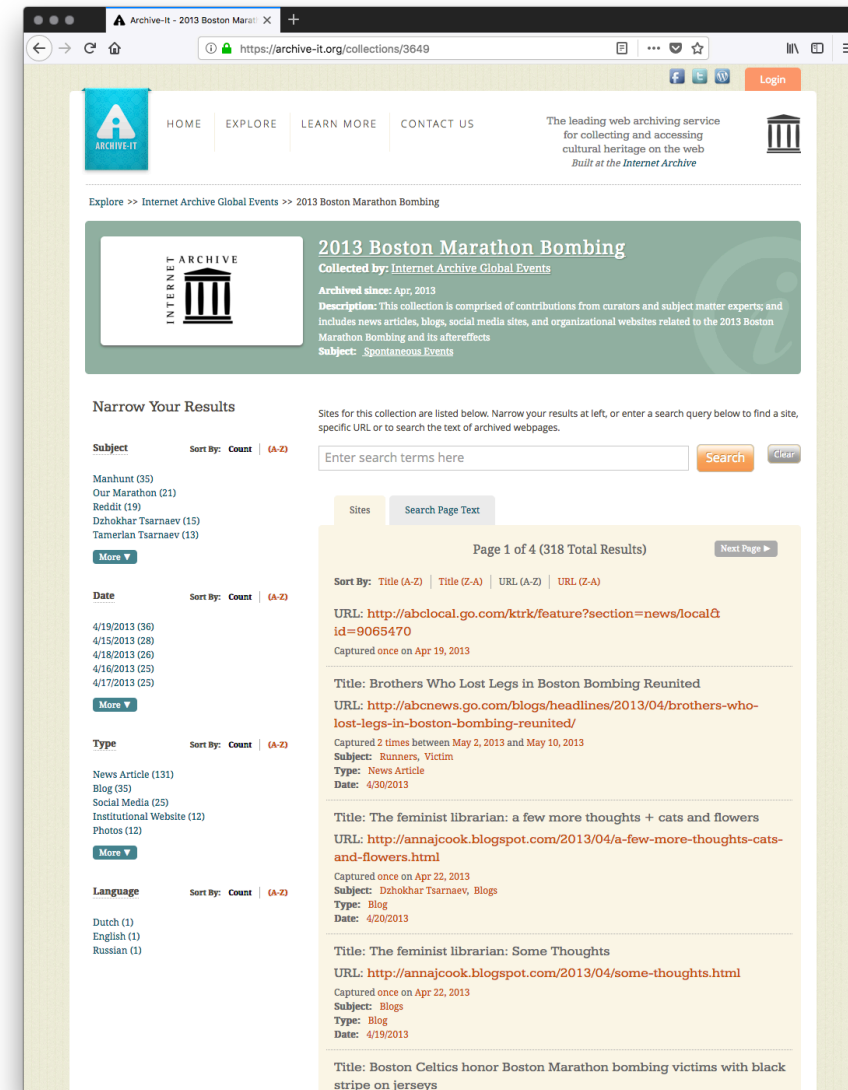   § Storytelling allows a curator to promote a collection, making others aware of it

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

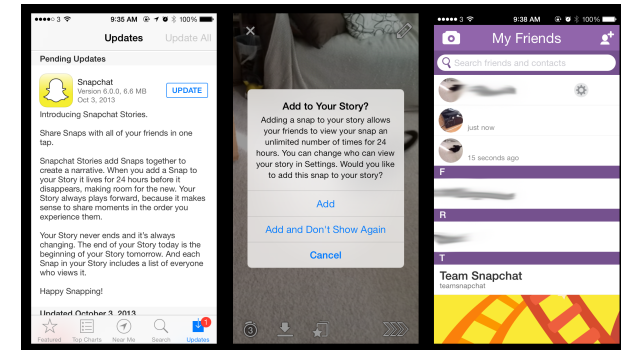# Some services equate "stories" with ephemera, which is not what we want

Facebook stories

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

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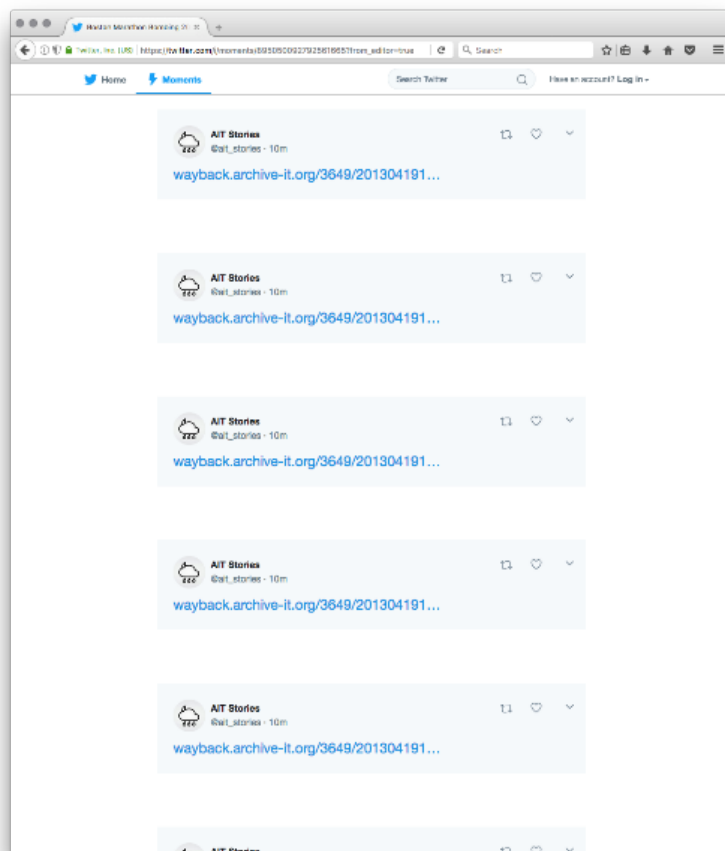
Image ref:
https://buffer.com/library/instagram-stories

Snapchat stories

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

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.

@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

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

Archive-It Stories http://wayback.archive-it.org/.../www.wcvb.com/news/local
August 3 at 5:26pm

Archive-It Stories http://wayback.archive-it.org/.../8uoVeyShelDu.../story.html

Patriot's Day heroes: Marathon spectators become first- responders - The Boston...

BOSTONGLOBE.COM

August 3 at 5:27pm

@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

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

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.

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

@shawnmjones @StormyArchives

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.

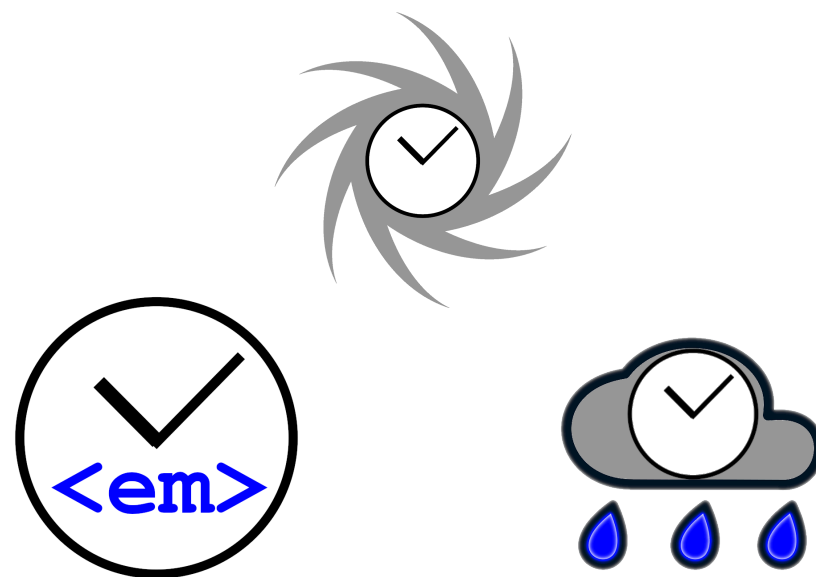Web archive collections are vast and hard to understand

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

Telling a story with web archives has 3 basic steps

1. Select the pages for the story

2. Gather information to summarize each memento

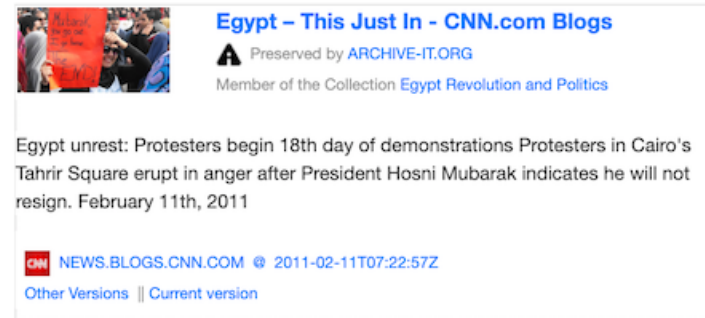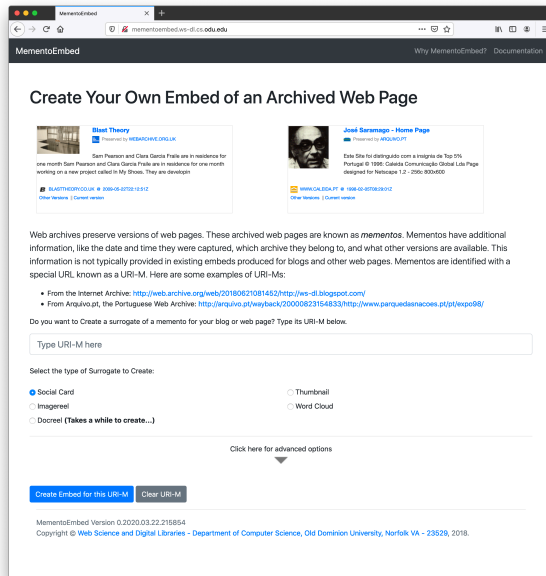3. Summarize all mementos together and publish the story
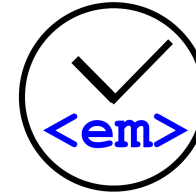
? → ? → ? →

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

Egypt – This Just In - CNN.com Blogs

Preserved by ARCHIVE-IT.ORG

Member of the Collection Egypt Revolution and Politics

Egypt unrest: Protesters begin 18th day of demonstrations Protesters in Cairo's Tahrir Square erupt in anger after President Hosni Mubarak indicates he will not resign. February 11th, 2011

CNN NEWS.BLOGS.CNN.COM @ 2011-02-11T07:22:57Z

Other Versions || Current version

Information about the archive

Information from the memento content

Information about other mementos for this resource

Link to current version of this resource

Information about the memento

Information in the MementoEmbed social card is separated to avoid issues of confusion about attribution.

MementoEmbed is archive-aware. It can locate information about the memento that is not available in other cards.

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 summarize *individual* pages with MementoEmbed in different ways

Cards

Browser Thumbnails

Word Clouds

Imagereels

Docreels

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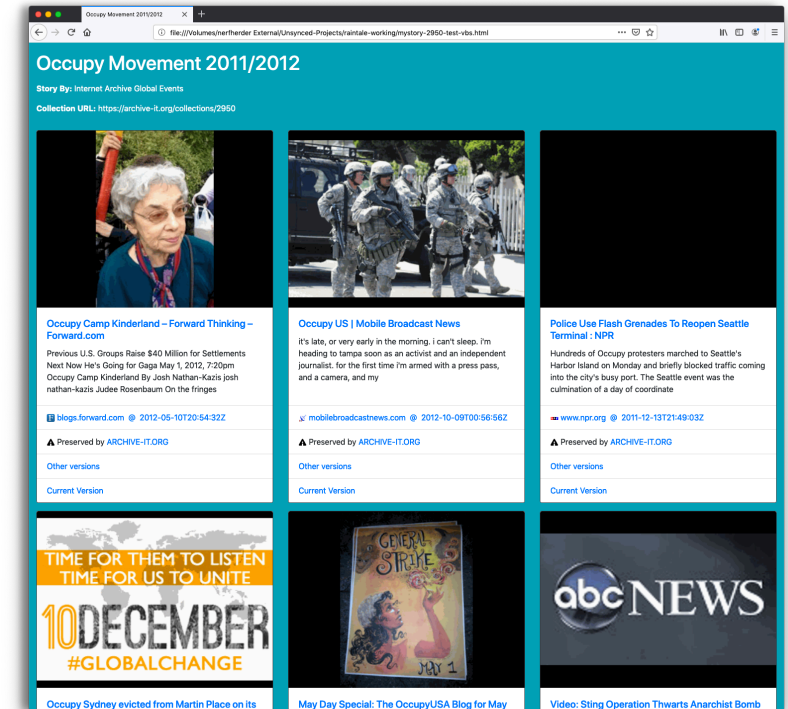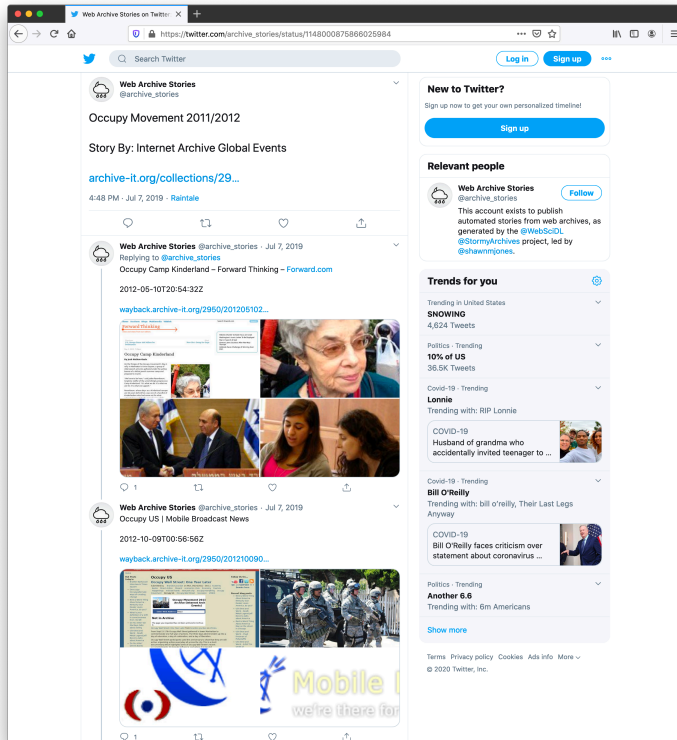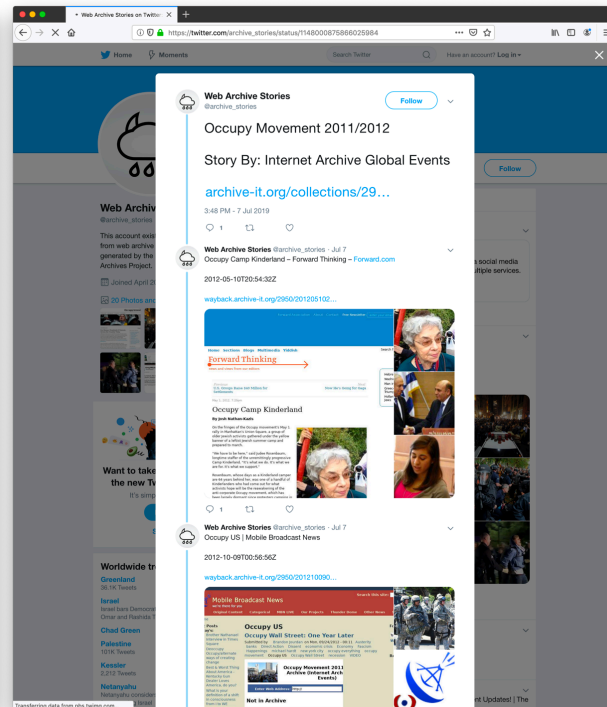the same page about the Egyptian Revolution.

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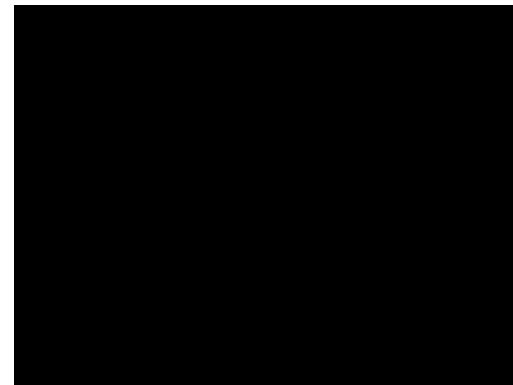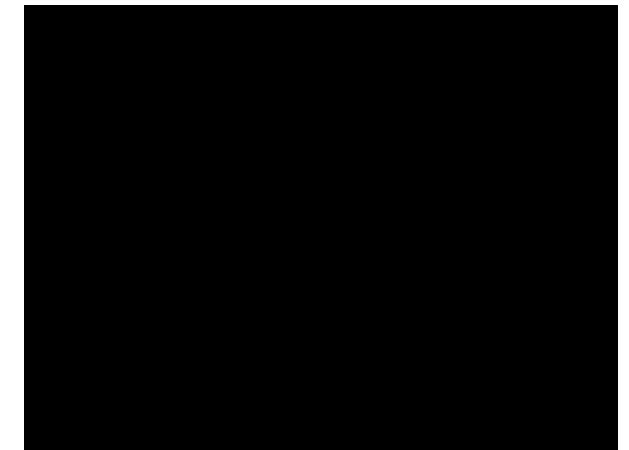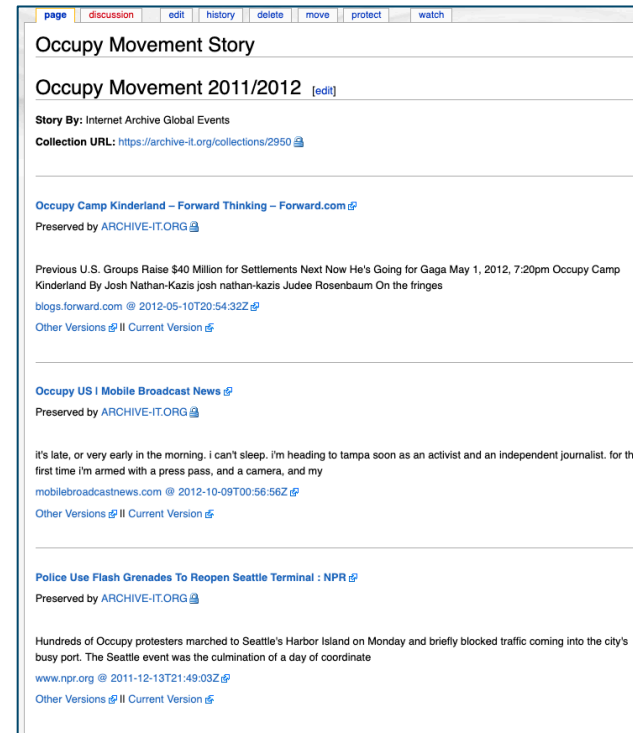
Now we can produce cards, but how do we create stories?

1. Select the pages for the story

2. Gather information to summarize each memento

3. Summarize all mementos together and publish the story

?

<em>

MementoEmbed

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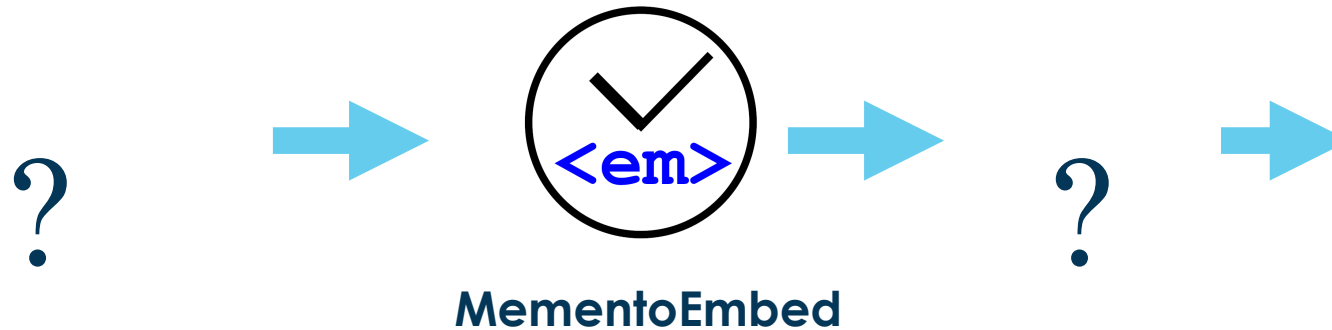
MementoEmbed just produces information for individual pages, so we created Raintale to tell stories…

§ 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/

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
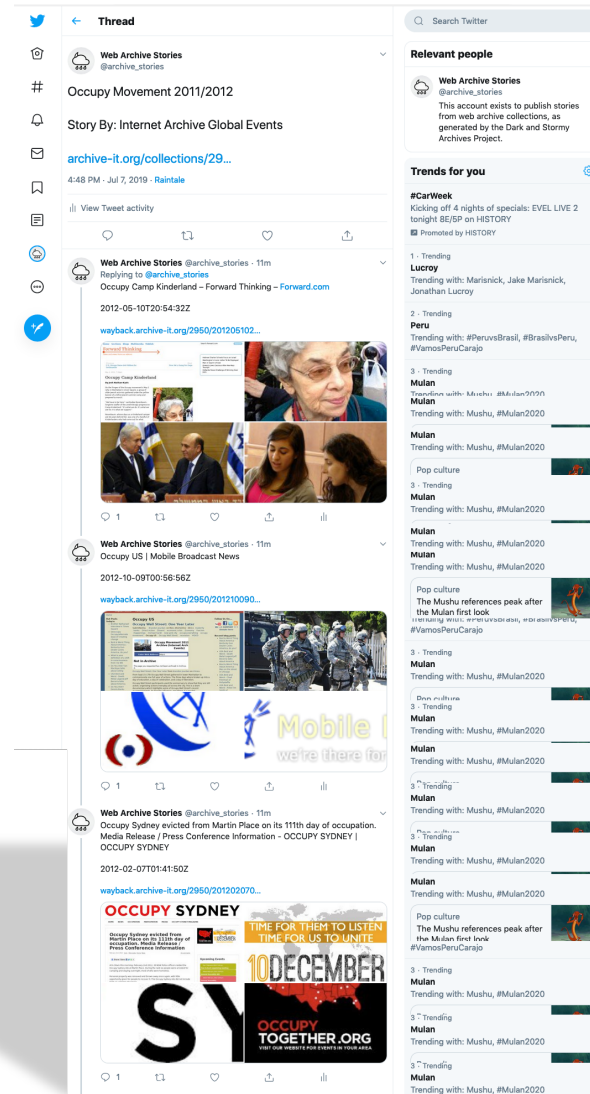
We can publish *many* Mementos with **Raintale** in a variety of formats
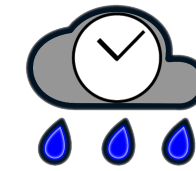
Twitter Threads
with screenshots,
images, titles, dates, links

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

38

Raintale uses templates to let you customize the look and destination media of your story
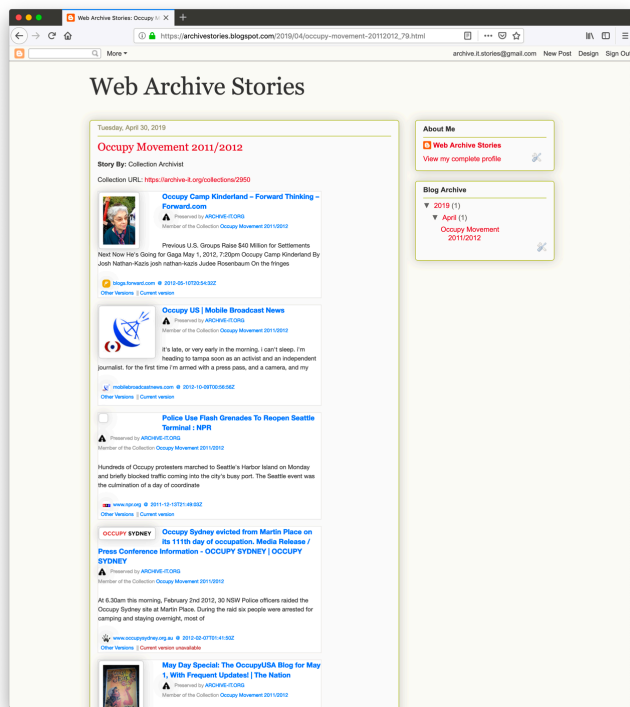
MementoEmbed
cards in Blogger

Customized Twitter
Threads

MediaWiki
Pages
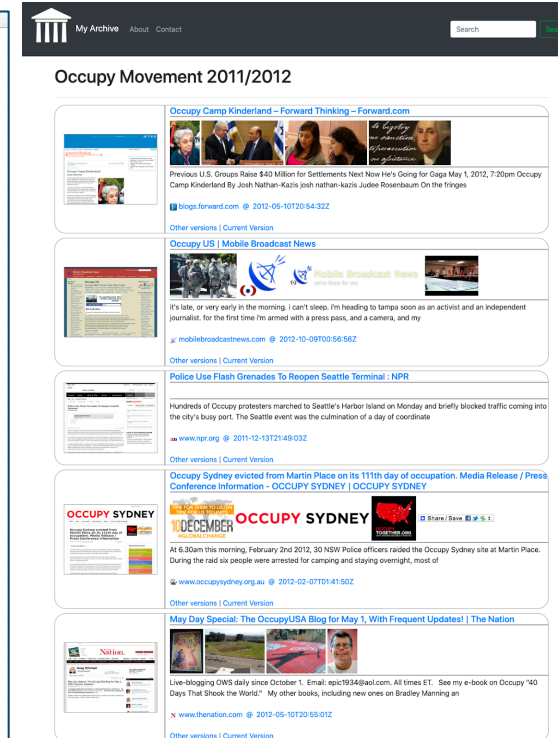
Brand your own stories for
your institution

@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

39

# 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

?
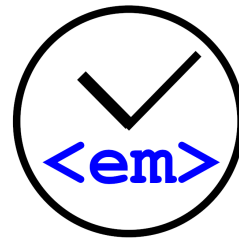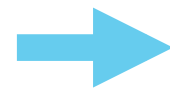
MementoEmbed
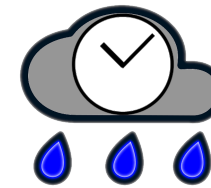
Raintale
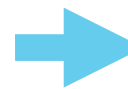
Hypercane provides intelligent sampling of web archive collections

§ **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)

Web Archive
Collections
(each thousands
or millions
of documents)

An
Intelligent Sample
(a manageable number)
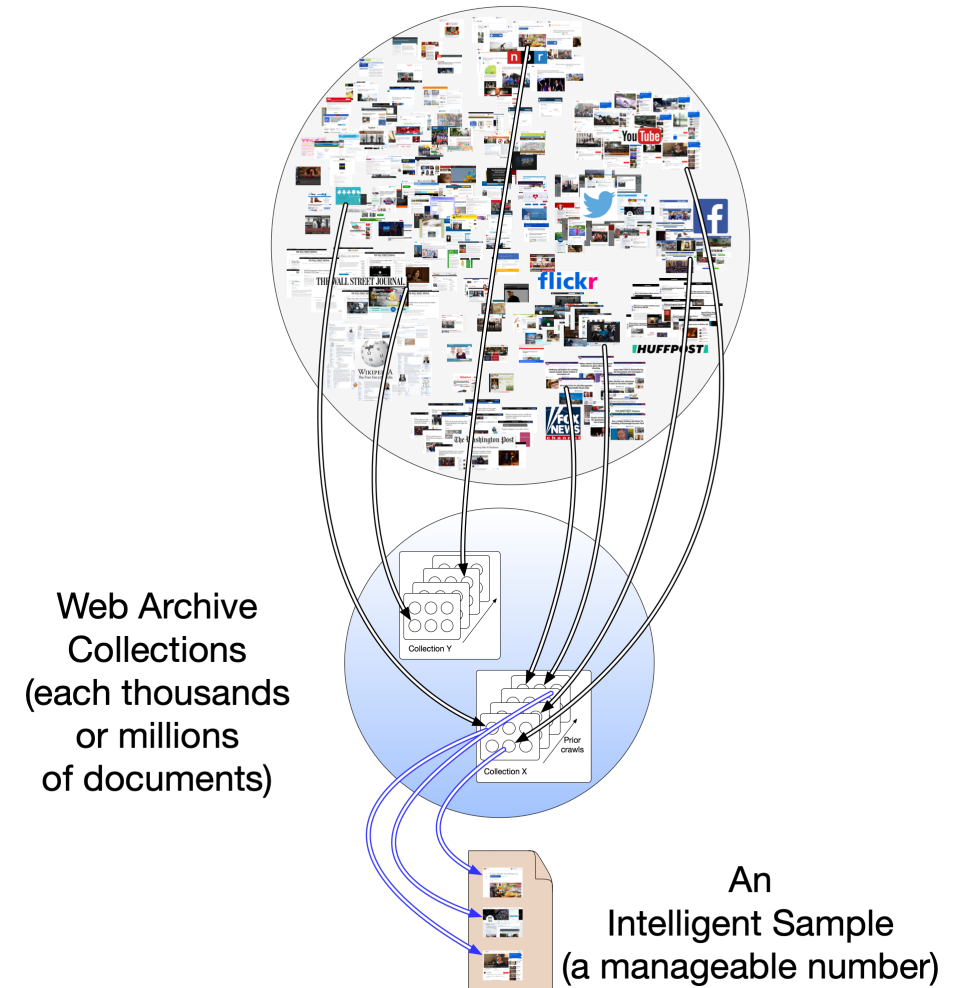
S. M. Jones. "Hypercane Part 1: Intelligent Sampling of Web Archive Collections"

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

1. Select the pages for the story

Automatically with **Hypercane**
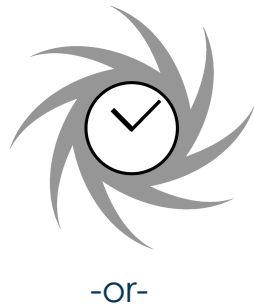
-or-

By People

2. Gather information to summarize each memento
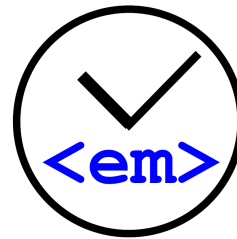
MementoEmbed

3. Summarize all mementos together and publish the story
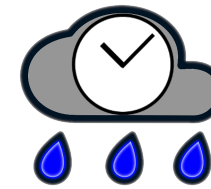
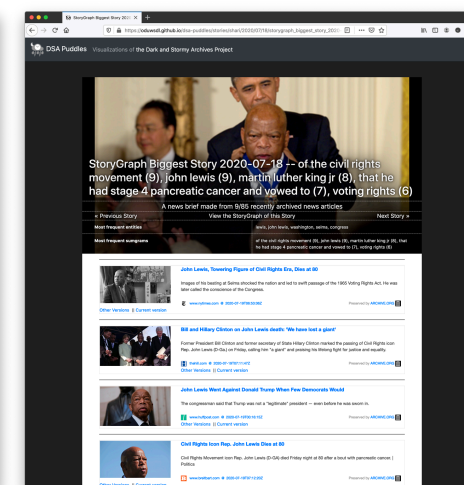**Raintale**
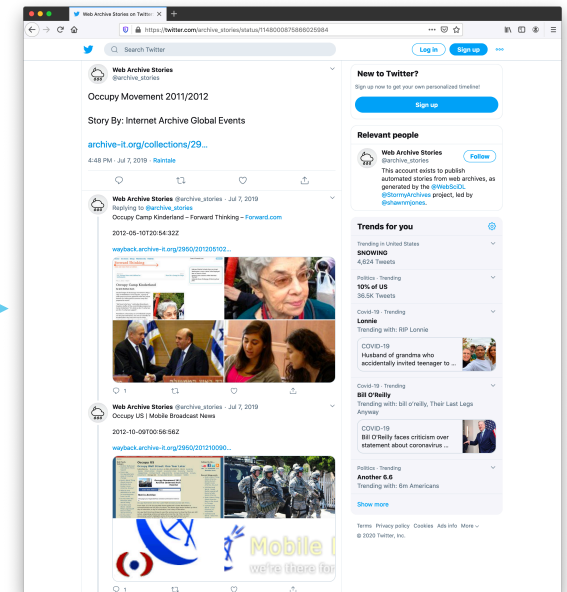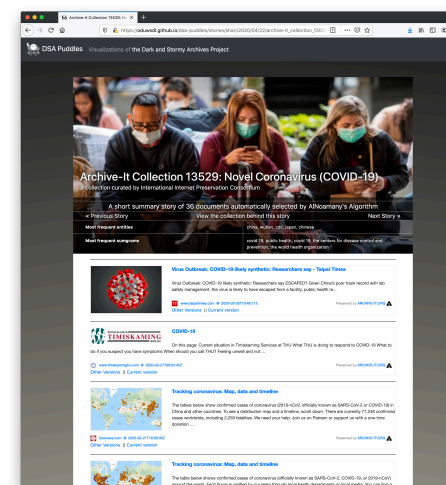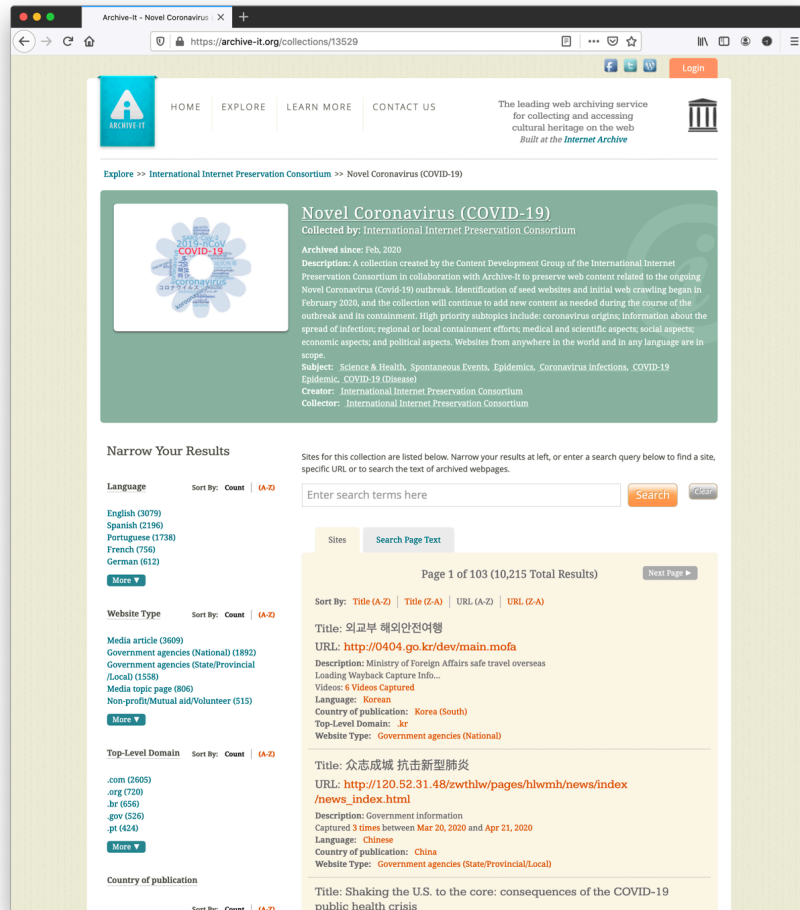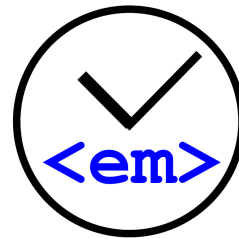
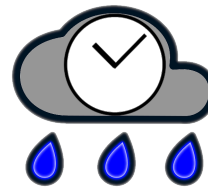# Examples of the DSA Toolkit at work

We automatically summarize public Archive-It collections
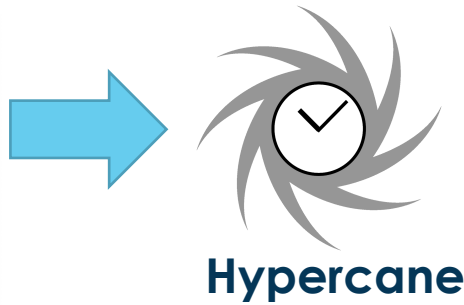
Hypercane

MementoEmbed

Raintale

From this: 23,376 mementos

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

44

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

Virginia Tech

Norway

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
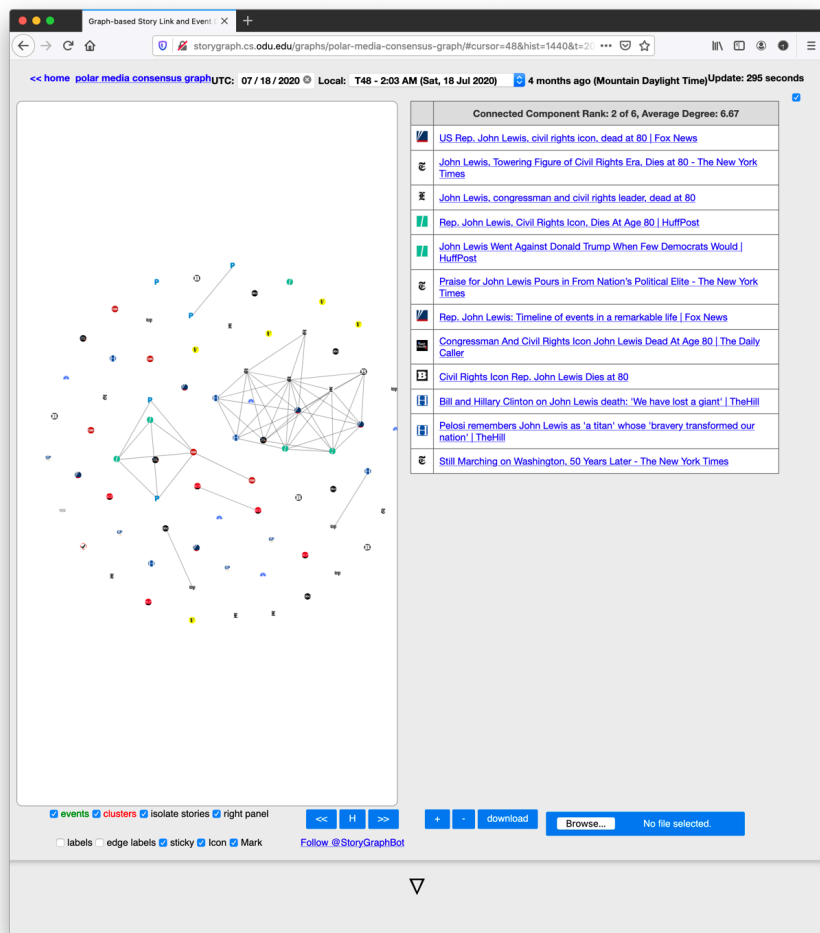
StoryGraph Toolkit

Hypercane
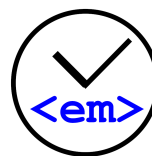
MementoEmbed

Raintale
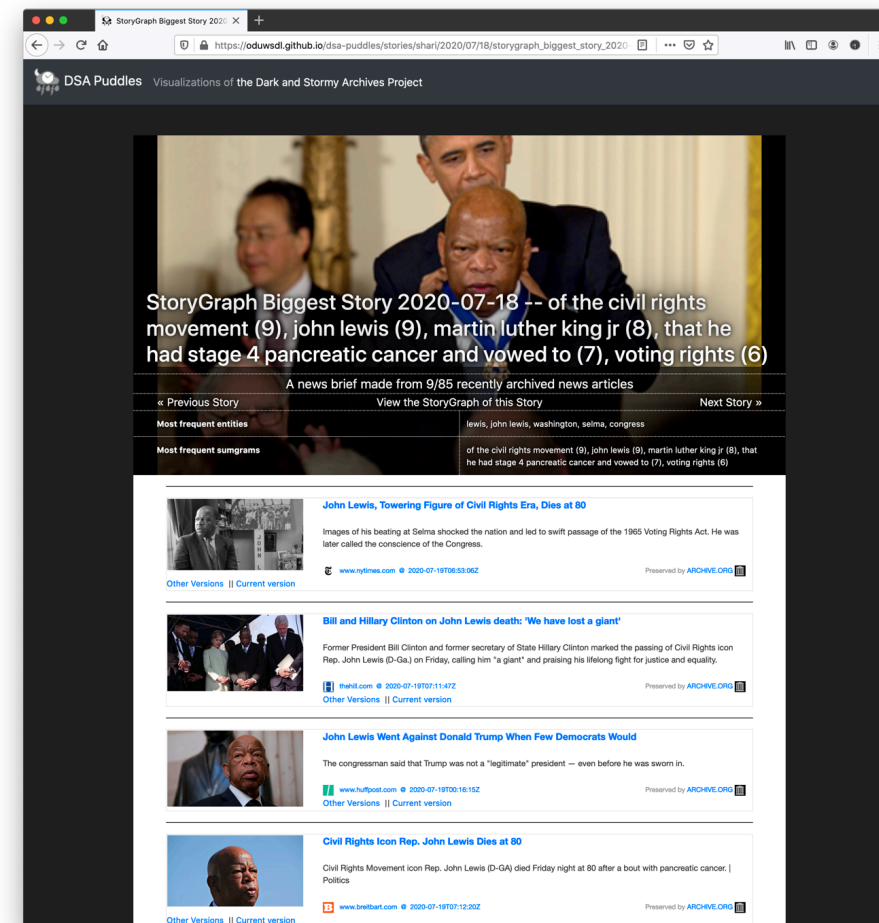
From this: news articles extracted by StoryGraph

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

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
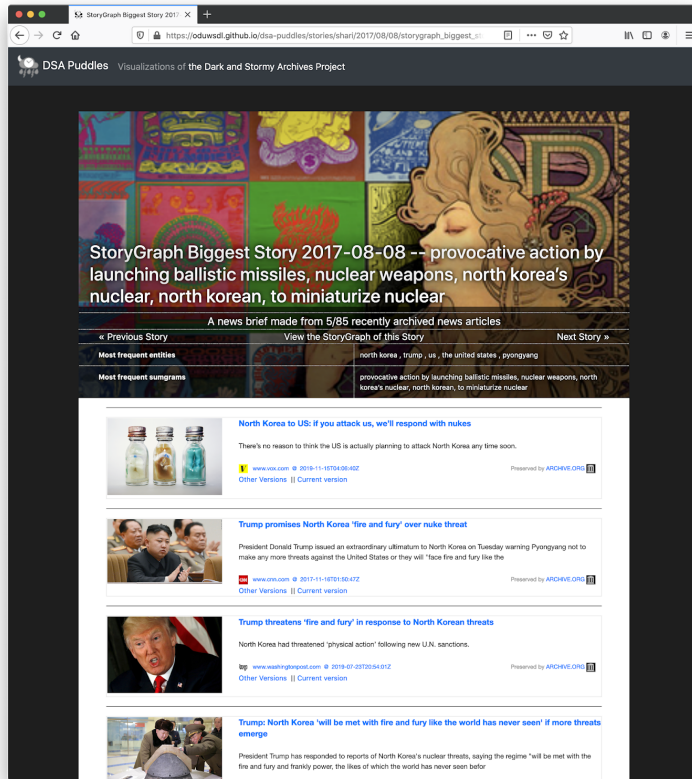
August 8, 2017
Nuclear Provocation by North Korea
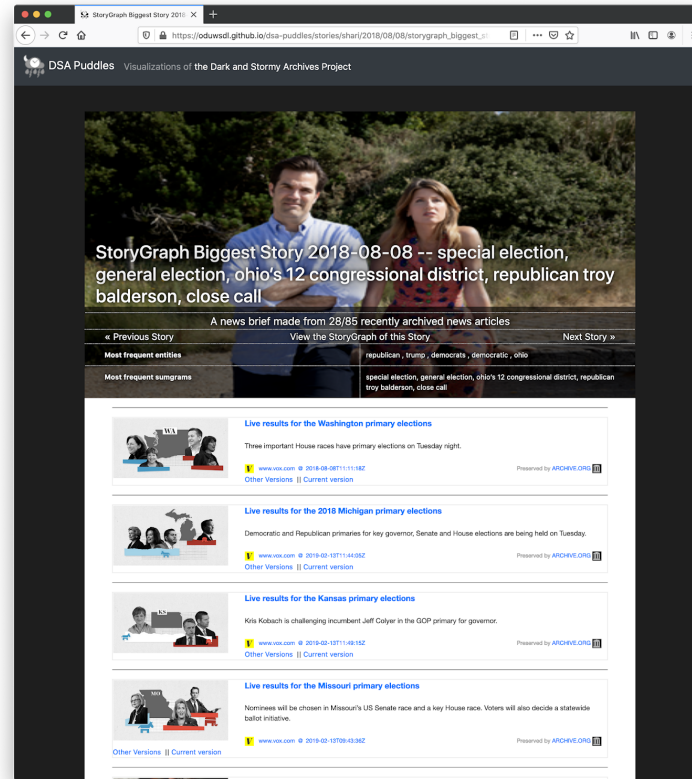
August 8, 2018
US Primaries and Special Elections
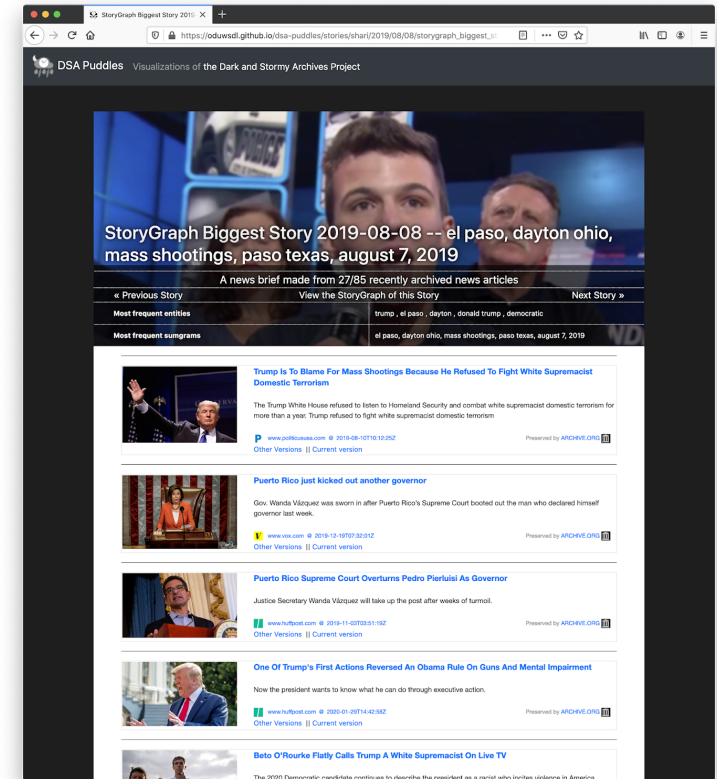
August 8, 2019
Aftermath of Shootings
at El Paso and Dayton

@shawnmjones @StormyArchives

We can automatically summarize Wikipedia references

From this: Wikipedia references as mementos

Wikipedia Scraper

Hypercane

MementoEmbed

Raintale

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

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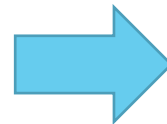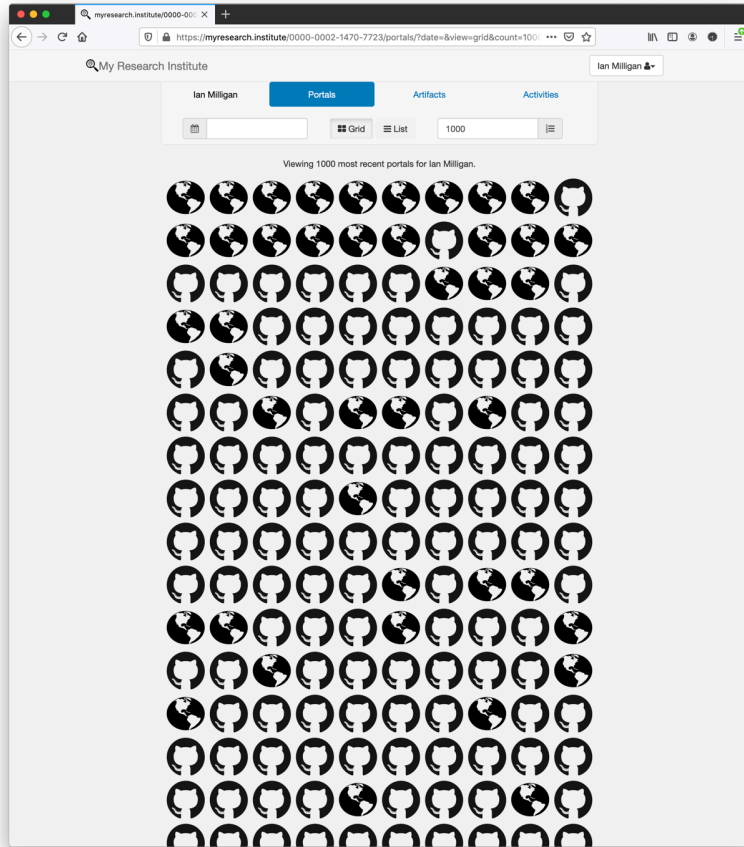
myresearch.institute Scraper

Hypercane
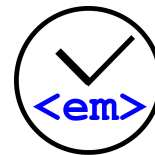
MementoEmbed
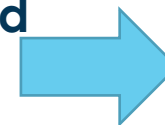
Raintale

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

To this: 20 mementos representing a scholar's work

You can automatically summarize
your web archive collection here
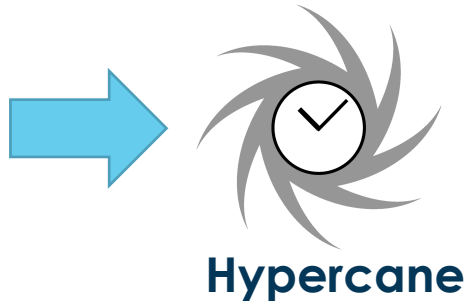
Hypercane
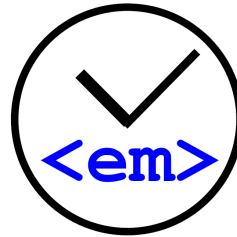
MementoEmbed

Raintale

Occupy Movement 2011/2012

Occupy Camp Kinderland – Forward Thinking – Forward.com

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

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

Police Use Flash Grenades To Reopen Seattle Terminal : NPR

Hundreds of Occupy protesters marched to Seattle's Harbor Island on Monday and briefly blocked traffic coming into the city's busy port. The Seattle event was the culmination of a day of coordinate

www.npr.org @ 2011-12-13T21:49:03Z

Other versions | Current Version

Occupy Sydney evicted from Martin Place on its 111th day of occupation. Media Release / Press Conference Information - OCCUPY SYDNEY | OCCUPY SYDNEY

At 6.30am this morning, February 2nd 2012, 30 NSW Police officers raided the Occupy Sydney site at Martin Place. During the raid six people were arrested for camping and staying overnight, most of

www.occupysydney.org.au @ 2012-02-07T01:41:50Z

Other versions | Current Version

May Day Special: The OccupyUSA Blog for May 1, With Frequent Updates! | The Nation

Live-blogging OWS daily since October 1. Email: epic1934@aol.com. All times ET. See my e-book on Occupy "40 Days That Shook the World." My other books, including new ones on Bradley Manning an

www.thenation.com @ 2012-05-10T20:55:01Z

Other versions | Current Version

From this: your web archive collection

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

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

@shawnmjones @StormyArchives

51

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

It was a dark and stormy night.