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

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 someone's 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://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.
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...
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.

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

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.
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

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

Instances

Text snippet (example from Bing)

Browser Thumbnail (example from UK Web Archive)

Text + Thumbnail (example from Internet Archive)

Social Card (example from Facebook)

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!1s0x89ba99ad24ba3945:0xcd2bdc432c4e4bac12m21d-76.30677612d36.88555151m51m11s0x87181246af22e765:0x7f5a90170c5df1b42m21d-106.28716212d35.8440582

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!
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.

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)

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

This is confusing.

Who published these resources?
Archive-It?
CNN?

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

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…

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.

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

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

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
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

[Image]

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

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)

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

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

3. Summarize all mementos together and publish the story

MementoEmbed
Raintale

@shawnmjones @StormyArchives
Examples of the DSA Toolkit at work
We automatically summarize public Archive-It collections

From this: 23,376 mementos

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


@shawnmjones @StormyArchives
With the DSA Toolkit, we can compare Archive-It collections about mass shootings.

Virginia Tech

Norway

El Paso

@shawnmjones @StormyArchives
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

From this: news articles extracted by StoryGraph

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

This allows us to review news from the same day in different years.
We can automatically summarize Wikipedia references

From this: Wikipedia references as mementos

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.

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

From this: your web archive collection

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

@shawnmjones @StormyArchives

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

§ MementoEmbed – Summarize Individual Mementos
  § Documentation: https://mementoembed.readthedocs.io/en/latest/
  § Report Issues: https://github.com/oduwsdl/MementoEmbed/issues