2020

SHARI- An Integration of Tools to Visualize the Story of the Day

Shawn M. Jones

Alexander C. Nwala
Old Dominion University

Martin Klein

Michele C. Weigle
Old Dominion University

Michael L. Nelson
Old Dominion University

Original Publication Citation

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. Web Archiving & Digital Libraries Workshop 2020, Virtual, August 5, 2020.

SHARI – An Integration of Tools to Visualize the Story of the Day

Shawn M. Jones
smjones@lanl.gov
Los Alamos National Laboratory
Los Alamos, New Mexico

Alexander C. Nwala
anwala@cs.odu.edu
Old Dominion University
Norfolk, Virginia

Martin Klein
mklein@lanl.gov
Los Alamos National Laboratory
Los Alamos, New Mexico

Michele C. Weigle
mweigle@cs.odu.edu
Old Dominion University
Norfolk, Virginia

Michael L. Nelson
mln@cs.odu.edu
Old Dominion University
Norfolk, Virginia

ABSTRACT

Tools such as Google News and Flipboard exist to convey daily news, but what about the news of the past? In this paper, we describe how to combine several existing tools and web archive holdings to convey the "biggest story" for a given date in the past. StoryGraph clusters news articles together to identify a common news story. Hypercane leverages ArchiveNow to store URLs produced by StoryGraph in web archives. Hypercane analyzes these URLs to identify the most common terms, entities, and highest quality images for social media storytelling. Raintale then takes the output of these tools to produce a visualization of the news story for a given day. We name this process SHARI (StoryGraph Hypercane ArchiveNow Raintale Integration). With SHARI, a user can visualize the articles belonging to a past date's news story.

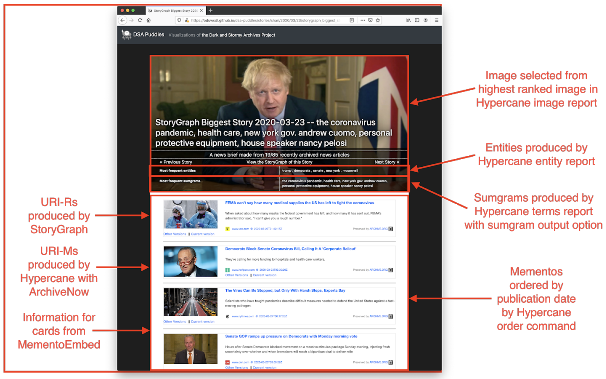
KEYWORDS

news, web archives, memento, storytelling, visualization, summarization

1 INTRODUCTION

AlNoamany et al. [1] introduced how to use social media storytelling to summarize web archive collections. Collections on specific topics exist at various web archives [6]. Klein et al. [7] have built collections from web archives by conducting focused crawls. Jones developed Hypercane [4] to intelligently sample mementos from larger collections. Jones also developed Raintale [3] for generating social media stories to summarize groups of mementos, providing visualizations that employ familiar techniques, like cards, that require no training for most users to understand. What if we want to tell stories from web archives with semi-current news articles?

Nwala et al. [9, 10] have focused on finding seeds within search engine result pages (SERPs), social media stories, and news feeds. As part of this research, Nwala et al. also developed StoryGraph [11], a tool that analyzes multiple news sources every ten minutes and automatically determines the news story or stories that dominate the media landscape at that time. Aturban et al. developed ArchiveNow [2], a tool that accepts live web URI-Rs and submits them to web archives to produce memento URI-Ms. We have tied StoryGraph together with tools from the Dark and Stormy Archives Toolkit[1] to produce visualizations summarizing the biggest StoryGraph story of a given day.


StoryGraph Biggest Story 2020
https://oduwsdl.github.io/dsa-puddles/stories/shari/2020/03/23/storygraph_biggest_s
DSA Puddles
Visualizations of the Dark and Stormy Archives Project
« Previous Story
View the StoryGraph of this Story
Next Story »
Other Versions || Current version
Preserved by ARCHIVE.ORG
Image selected from highest ranked image in Hypercane image report
Entities produced by Hypercane entity report
Sumgrams produced by Hypercane terms report with sumgram output option
Mementos ordered by publication date by Hypercane order command
Formatting of page by Raintale
URI-Rs produced by StoryGraph
URI-Ms produced by Hypercane with ArchiveNow
Information for cards from MementoEmbed

StoryGraph Biggest Story 2020-03-23 -- the coronavirus pandemic, health care, new york gov. andrew cuomo, personal protective equipment, house speaker nancy pelosi

A news brief made from 19/85 recently archived news articles

Most frequent entities	trump , democrats , senate , new york , mcconnell
Most frequent sumgrams	the coronavirus pandemic, health care, new york gov. andrew cuomo, personal protective equipment, house speaker nancy pelosi

FEMA can't say how many medical supplies the US has left to fight the coronavirus

When asked about how many masks the federal government has left, and how many it has sent out, FEMA's administrator said, "I can't give you a rough number."

www.vox.com @ 2020-03-22T21:42:17Z

Democrats Block Senate Coronavirus Bill, Calling It A 'Corporate Bailout'

They're calling for more funding to hospitals and health care workers.

www.huffpost.com @ 2020-03-23T03:30:26Z

The Virus Can Be Stopped, but Only With Harsh Steps, Experts Say

Scientists who have fought pandemics describe difficult measures needed to defend the United States against a fast-moving pathogen.

www.nytimes.com @ 2020-03-24T00:17:25Z

Senate GOP ramps up pressure on Democrats with Monday morning vote

Hours after Senate Democrats blocked movement on a massive stimulus package Sunday evening, injecting fresh uncertainty over whether and when lawmakers will reach a bipartisan deal to deliver relie

www.cnn.com @ 2020-03-23T03:09:29Z

Figure 1: The "biggest news story" of for the March 23, 2020 story produced by SHARI[2]. Annotations detail which components provide each part of the visualization.

2 THE SHARI PROCESS

The StoryGraph Hypercane ArchiveNow Raintale Integration (SHARI) [5] process automatically creates stories summarizing news for a day. Figure 1 details what each tool contributes to the story. Figure 2 shows the steps of the SHARI process. In step 1, with the StoryGraph Toolkit, we query the StoryGraph service for the URI-Rs belonging to the biggest story of the day. In step 2, Hypercane converts these URI-Rs to URI-Ms by first querying the LANL Memento Aggregator via the Memento Protocol [12]. For each URI-M that does not have a memento, Hypercane creates a memento by calling ArchiveNow [2]. In step 3, Hypercane runs the mementos through spaCy[3] to generate a list of named entities, sorted by frequency. In step 4, Hypercane runs the mementos through sumgram [8] and generates a list of sumgrams, sorted by frequency. In step 5, Hypercane scores all of the mementos' embedded images. In step 6, Hypercane runs the mementos through newspaper3k[4] to extract each article's publication date and orders the URI-Ms by that date. In step 7, Hypercane consolidates the entities, terms, image scores, and ordered URI-Ms into a JSON file containing the structured data for the summary. During this step, Hypercane uses the highest scoring image as the striking image for the summary. In Figure 1,

[1] https://oduwsdl.github.io/dsa/software.html

[2] https://oduwsdl.github.io/dsa-puddles/stories/shari/2020/03/23/storygraph_biggest_story_2020-03-23/

[3] https://spacy.io/

[4] https://newspaper.readthedocs.io/en/latest/

the highest-ranking image is the UK Prime Minister addressing his country about the COVID-19 pandemic. In step 8, Raintale renders the output as Jekyll HTML based on the contents of this JSON file, a template file, and information on each memento provided by MementoEmbed. In step 9, the SHARI script publishes the summary story to GitHub Pages for distribution. Figure 3 shows the output of our *dsa_tweeter* bot which announces the story after publication.

3 SUMMARY AND FUTURE WORK

SHARI produces a familiar yet novel method of viewing news for a day in the past. SHARI can create stories for today, yesterday, and back to StoryGraph's creation on August 8, 2017. It is different from other storytelling services like Wakelet[5] because SHARI is entirely automated. The stories produced by SHARI are different from services like Google News[6] or Flipboard[7] because those tools focus on current events and personalized topics. Because StoryGraph samples content from multiple sides of the political spectrum, the SHARI process can provide a visualization of articles not tied to one interest area or even a single side's terminology. This process works because each component is loosely coupled, has high cohesion, has explicit interfaces, and engages in information hiding. Each command passes data in the expected format to the next.

We are exploring how to produce and render other news stories for a given day and any given period of time. We are researching how to best visualize significant events that span substantial periods of time, like the entire COVID-19 news story. Though StoryGraph is an existing service that gathers current news, we also want to apply its algorithm directly to mementos and tell the news stories of past events like the Hurricane Katrina disaster. One day, through SHARI, historians, journalists, and other researchers may glance at the news for any date.

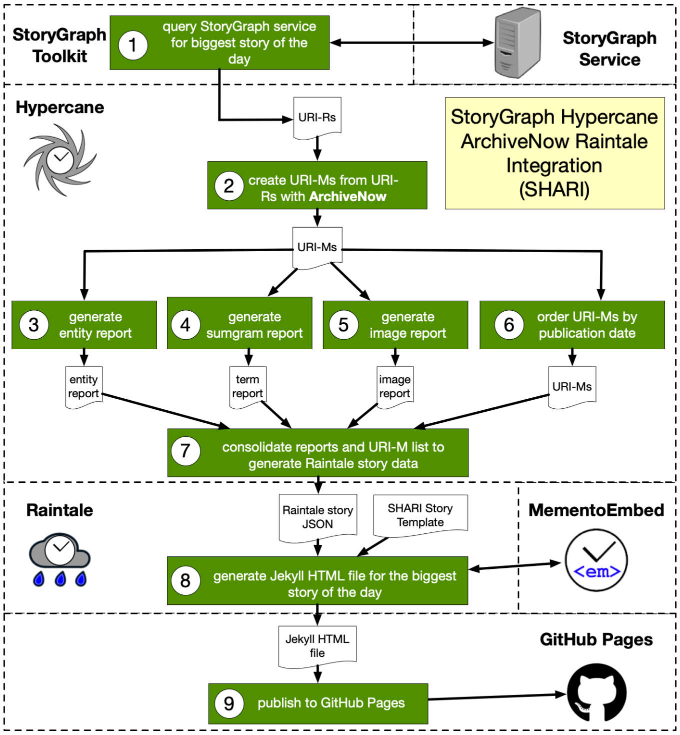
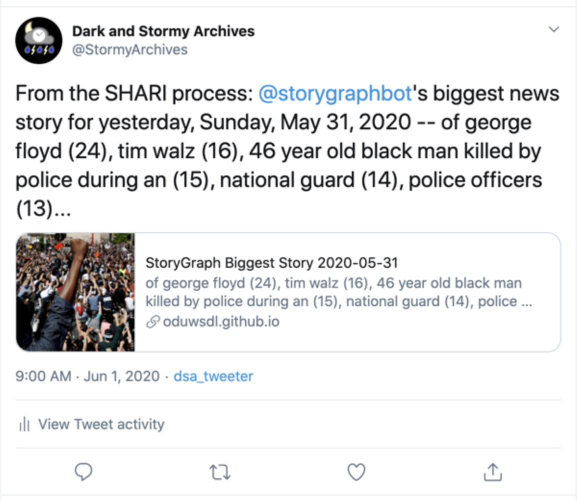

StoryGraph Toolkit
1 query StoryGraph service for biggest story of the day
StoryGraph Service
Hypercane
URI-Rs
2 create URI-Ms from URI-Rs with ArchiveNow
StoryGraph Hypercane ArchiveNow Raintale Integration (SHARI)
URI-Ms
3 generate entity report
4 generate sumgram report
5 generate image report
6 order URI-Ms by publication date
entity report
term report
image report
URI-Ms
7 consolidate reports and URI-M list to generate Raintale story data
Raintale
Raintale story JSON
SHARI Story Template
MementoEmbed
<em>
8 generate Jekyll HTML file for the biggest story of the day
Jekyll HTML file
GitHub Pages
9 publish to GitHub Pages

Figure 2: SHARI process for creating a visualization of the biggest news story for a given day


Dark and Stormy Archives
@StormyArchives

From the SHARI process: @storygraphbot's biggest news story for yesterday, Sunday, May 31, 2020 -- of george floyd (24), tim walz (16), 46 year old black man killed by police during an (15), national guard (14), police officers (13)...

StoryGraph Biggest Story 2020-05-31
of george floyd (24), tim walz (16), 46 year old black man killed by police during an (15), national guard (14), police ...
oduwsdl.github.io

9:00 AM · Jun 1, 2020 · dsa_tweeter

Figure 3: The *dsa_tweeter* bot announces the availability of new SHARI stories each day.

REFERENCES

[1] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. 2017. Generating Stories From Archived Collections. In *WebSci 2017*. Troy, New York, USA, 309–318. https://doi.org/10.1145/3091478.3091508

[2] Mohamed Aturban, Mat Kelly, Sawood Alam, John A. Berlin, Michael L. Nelson, and Michele C. Weigle. 2018. ArchiveNow: Simplified, Extensible, Multi-Archive Preservation. In *JCDL 2018*. Fort Worth, Texas, USA, 321–322. https://doi.org/10.1145/3197026.3203880

[3] Shawn M. Jones. 2019. Raintale – A Storytelling Tool For Web Archives. https://ws-dl.blogspot.com/2019/07/2019-07-11-raintale-storytelling-tool.html

[4] Shawn M. Jones. 2020. Hypercane Part 1: Intelligent Sampling of Web Archive Collections. https://ws-dl.blogspot.com/2020/06/2020-06-03-hypercane-part-1-intelligent.html

[5] Shawn M. Jones. 2020. SHARI: StoryGraph Hypercane ArchiveNow Raintale Integration – Combining WS-DL Tools For Current Events Storytelling. https://ws-dl.blogspot.com/2020/04/2020-04-01-shari-storygraph-hypercane.html

[6] Shawn M Jones, Alexander Nwala, Michele C Weigle, and Michael L Nelson. 2018. The Many Shapes of Archive-It. In *iPres 2018*. Boston, Massachusetts, USA, 1–10. https://doi.org/10.17605/OSF.IO/EV42P

[7] Martin Klein, Lyudmila Balakireva, and Herbert Van de Sompel. 2018. Focused Crawl of Web Archives to Build Event Collections. In *WebSci 2018*. Amsterdam, Netherlands, 333–342. https://doi.org/10.1145/3201064.3201085

[8] Alexander C. Nwala. 2019. Introducing sumgram, a tool for generating the most frequent conjoined ngrams. https://ws-dl.blogspot.com/2019/09/2019-09-09-introducing-sumgram-tool-for.html

[9] Alexander C. Nwala, Michele C. Weigle, and Michael L. Nelson. 2018. Bootstrapping Web Archive Collections from Social Media. In *Hypertext 2018*. Baltimore, Maryland, USA, 64–72. https://doi.org/10.1145/3209542.3209560

[10] Alexander C. Nwala, Michele C. Weigle, and Michael L. Nelson. 2018. Scraping SERPs for Archival Seeds: It Matters When You Start. In *JCDL 2018*. Fort Worth, Texas, USA, 263–272. https://doi.org/10.1145/3197026.3197056

[11] Alexander C. Nwala, Michele C. Weigle, and Michael L. Nelson. 2020. *365 Dots in 2019: Quantifying Attention of News Sources.* Technical Report arXiv:2003.09989. https://arxiv.org/abs/2003.09989 arXiv: 2003.09989.

[12] Herbert Van de Sompel, Michael Nelson, and Robert Sanderson. 2013. RFC 7089 - HTTP Framework for Time-Based Access to Resource States – Memento. https://tools.ietf.org/html/rfc7089

[5] https://wakelet.com/
[6] https://news.google.com/
[7] https://flipboard.com/