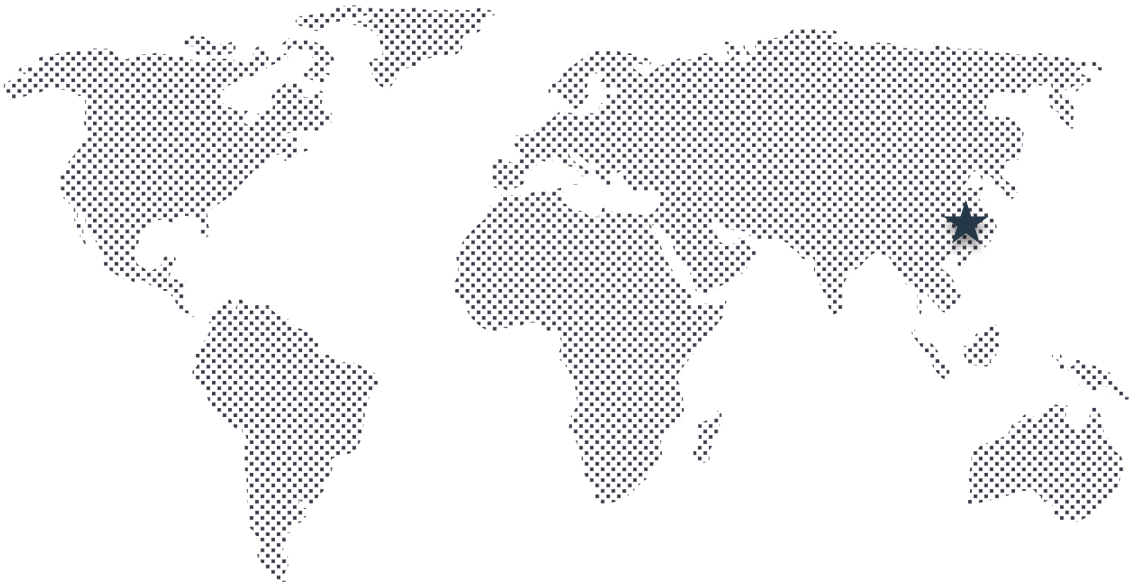
Proceedings *of the*



2020

Web Archiving & Digital Libraries

Workshop

August 5, 2020

Virtual Event

Zhiwu Xie

Martin Klein

Edward A. Fox

Editors

WADL 2020 Homepage

Web Archiving and Digital Libraries,
a Virtual Workshop of JCDL 2020 (http://2020.jcdl.org),
Aug. 5, 2020, according to the WADL 2020 schedule.
We welcome broad attendance; contact the co-chairs if you have questions.

Please see the approved WADL 2020 workshop description from the JCDL proceedings.

Please also refer to past WADL homepages: 2019, 2018, 2017, and 2016. Past workshop proceedings can be found from: WADL 2017-19, Pre 2016. Prior workshops have led in part to a special issue of International Journal on Digital Libraries.

Invited Talk: Dr. Tian Xia

- *Title:* The practice and inspiration of Web Archiving in China
- *Abstract:* Much of information published on the Web is unique and historically valuable, therefore archiving Web content is an important task for social memory. This talk will introduce some typical Web archiving projects carried out by libraries, archives and scientific research institutions in China. The contributions and valuable lessons from these projects will be discussed, so we can get the key factors affecting the success of Web archiving projects. It can be expected that institutional Web Archiving with limited objectives, clear responsibilities and utilization-driven characteristics will become the mainstream of Web Archiving in China.
- *CV:* Dr. Xia is an Associate Professor at School of Information Resource Management, Renmin University of China, and a research fellow at Key Laboratory of Data Engineering and Knowledge Engineering, MOE, China. His research interests include information retrieval, Web mining, electronic record management and semantic Web. He has published more than 40 papers and 6 books in the fields of LIS and archival science.

Invited Panel:

- *Title:* Making, Using, and Exploring Web Archives: Tales from Scholars & Practitioners
- *Moderator:* Vicky Steeves, New York University
- Alexander Nwala, Old Dominion University
- Genevieve Milliken, New York University
- Emily Maemura, University of Toronto
- Karen Hansen, Ithaka/Portico
- Meghan Lyon, Library of Congress
- *Abstract:* Web archiving has contributed immensely to the digital preservation landscape, allowing for a wider range of materials to be saved for greater reuse. A remaining and bustling area of work in web archiving includes capturing dynamic and complex content hosted on the web with high fidelity. This type of preservation activity not only has major technical barriers, but also social and legal ones too. This panel will bring together practitioners from computer, library, and information sciences to discuss how web archiving has been applied to save complex digital objects, such as eBooks, source code and its contextual ephemera, and many more. Panelists will discuss not only the current state of the art in web archiving complex material, but also the interoperability of web archiving technology and how that might help facilitate the reuse of web archives.

Accepted Papers

- **Format:** 20 min. presentation + 10 min. Q&A

- "125 Databases for the Year 2080" by Kai Naumann
- "Improving the Quality of Web Harvests Using Web Curator Tool" by Ben O'Brien, Andrea Goethals, Jeffrey van der Hoeven, Hanna Koppelaar, Trienka Rohrbach, Steve Knight, Frank Lee, and Charmaine Fajardo
- "SHARI – An Integration of Tools to Visualize the Story of the Day" by Shawn Jones, Alexander Nwala, Martin Klein, Michele Weigle, and Michael Nelson
- "MementoEmbed and Raintale for Web Archive Storytelling" by Shawn Jones, Martin Klein, Michele Weigle, and Michael Nelson
- "TMVis: Visualizing Webpage Changes Over Time" by Abigail Mabe, Dhruv Patel, Maheedhar Gunnam, Surbhi Shankar, Mat Kelly, Sawood Alam, Michael Nelson, and Michele Weigle

Description:

Due to COVID-19, JCDL 2020 will be held virtually with online sessions and discussions. WADL 2020 will also be moved entirely online.

WADL 2020 will continue the WADL tradition to provide a forum and collaboration platform for international leaders from academia, industry, and government to discuss challenges, and share insights, in designing and implementing concepts, tools, and standards in the realm of web archiving. Together, we will explore the integration of web archiving and digital libraries, over the complete digital resource life cycle: creation/authoring, uploading, publishing on the web, crawling/collecting, compressing, formatting, storing, preserving, analyzing, indexing, supporting access, etc.

WADL 2020 will cover all topics of interest, including but not limited to:

- Archival metadata, description, classification
- Archival standards, protocols, systems, tools
- Community building
- Crawling of dynamic, online art, and mobile content
- Discovery of archived resources
- Diversity in web archives
- Ethics in web archiving
- Event archiving and collection building
- Extraction and analysis of archival records
- Interoperability of web archiving systems
- National and international perspectives on web archiving
- Social media archiving

Objectives:

- Continue to build the diverse community of people integrating web archiving with digital libraries
- Help attendees learn about useful methods, systems, and software in this area
- Help chart future research and practice in this area, to enable more and higher quality web archiving
- Promote synergistic efforts including collaborative projects and proposals
- Produce an archival publication that will help advance technology and practice

Workshop Co-chairs:

- Chair: Zhiwu Xie, Professor & Chief Strategy Officer, Virginia Tech Libraries, zhiwuxie@vt.edu,
- Co-chair: Edward A. Fox, Professor and Director Digital Library Research Laboratory, Virginia Tech, fox@vt.edu http://fox.cs.vt.edu,
- Co-chair: Martin Klein, Los Alamos National Laboratory Research Library, mklein@lanl.gov

Program Committee:

- Brunelle, Justin F., The MITRE Corporation, jbrunelle@cs.odu.edu
- Duncan, Sumitra, Frick Art Reference Library, duncan@frick.org
- Finnell, Joshua, Colgate University, jfinnell@colgate.edu
- Goethals, Andrea, National Library of New Zealand, Andrea.Goethals@dia.govt.nz
- Jones, Shawn, Old Dominion University, sjone@cs.odu.edu
- Ko, Lauren, UNT Libraries, lauren.ko@unt.edu
- McCown, Frank, Harding University, fmccown@harding.edu
- Nelson, Michael, Old Dominion University, mln@cs.odu.edu
- Risse, Thomas, University Frankfurt, University Library J. C. Senckenber, t.risse@ub.uni-frankfurt.de
- Taylor, Nicholas, Los Alamos National Laboratory, taylorn@gmail.com
- Weber, Matthew, Rutgers University, matthew.weber@rutgers.edu
- Weigle, Michele, Old Dominion University, mweigle@cs.odu.edu
- Wrubel, Laura, George Washington University, lwrubel@gwu.edu

Submissions (please provide contact and supporting info in <= 2 pages using the ACM Proceedings template):

- Due: June 13, 2020; Notifications: June 22, 2020. Completed - see results above.
- EasyChair submission page: https://easychair.org/conferences/?conf=wadl2020
- Please use the ACM Proceedings template.

Updated time: July 21, 2020

Date	Beijing Time (UTC+8)	New York Time (UTC-4)	London Time (UTC+1)	Room 1 (New York Time, UTC-4)
Day 5 (August 5 Beijing Time, August 5 New York Time)	23:00-0:30	11:00-12:30	16:00-17:30	11:00 - 11:15 Welcome & Self Introduction 11:15 - 12:00 Invited Talk “The practice and inspiration of Web Archiving in China”, Tian Xia, School of Information Resource Management, Renmin University of China 12:00 - 12:30 "TMVis: Visualizing Webpage Changes Over Time" by Abigail Mabe, Dhruv Patel, Maheedhar Gunnam, Surbhi Shankar, Mat Kelly, Sawood Alam, Michael Nelson, and Michele Weigle
	1:00-2:30	13:00-14:30	18:00-19:30	13:00 - 13:30 "125 Databases for the Year 2080" by Kai Naumann 13:30 - 14:00 "SHARI – An Integration of Tools to Visualize the Story of the Day" by Shawn Jones, Alexander Nwala, Martin Klein, Michele Weigle, and Michael Nelson 14:00 - 14:30 "MementoEmbed and Raintale for Web Archive Storytelling" by Shawn Jones, Martin Klein, Michele Weigle, and Michael Nelson
	3:00-4:30	15:00-16:30	20:00-21:30	15:00 - 16:00 Invited Panel “Making, Using, and Exploring Web Archives: Tales from Scholars & Practitioners” Moderator: Vicky Steeves, New York University Alexander Nwala, Old Dominion University Genevieve Milliken, New York University Emily Maemura, University of Toronto Karen Hansen, Ithaka/Portico Meghan Lyon, Library of Congress 16:00 - 16:30 "Improving the Quality of Web Harvests Using Web Curator Tool" by Ben O'Brien, Andrea Goethals, Jeffrey van der Hoeven, Hanna Koppelaar, Trienka Rohrbach, Steve Knight, Frank Lee, and Charmaine Fajardo
	5:00-6:30	17:00-18:30	22:00-23:30	17:00 - 18:30 Project briefing, workshop continuation, and open discussion.

125 Databases for the Year 2080

A technology challenge and how it can be met

Kai Naumann
Archival Policy Department
Landesarchiv Baden-Württemberg
Stuttgart, Germany
kai.naumann@la-bw.de

ABSTRACT

In March 2020, the author issued a challenge in a number of mailing lists about how best to preserve a selection of databases (including their proper user interfaces) for future use from the year 2080 onwards. The mission is to prepare about 125 databases in such a way that they can be used in as many ways as possible in 2080. In the following 60 years a) no costs should be incurred apart from the secure storage of the data and b) the database contents must not be publicly accessible. This challenge should be met in the most robust and cost-efficient way.

The classical strategy for this task is choosing a well-known way of encoding the database contents. Being an experienced preservation practitioner, the author suggested that participants make an informed decision not to preserve some of the databases' content or capabilities for the sake of lower cost.

Progress in virtualisation and emulation technology might persuade a participant to put trust into just encapsulating content and software and storing it for the emulators of 2080. But will reviving hibernated 60 year-old databases in 2080 be an everyday routine?

This presentation sums up the answers we got up to this point and sets out the further steps the Landesarchiv will take in order to collaboratively shape a vision for long-term database preservation.

KEYWORDS

Web database, long-term preservation, emulation, technology watch, archival appraisal, E-ARK, SIARD2, relational database, database preservation

The background

Like other memory institutions, the Baden-Württemberg State Archives (Landesarchiv) has the duty to preserve its holdings for an indefinite term. Its most significant holdings are government records from all branches like police, jurisdiction, education, construction, agriculture, surveying and others. The proportion of this knowledge stored on paper today gets smaller every year. Databases, associated with records management systems, form the backbone of knowledge management in all domains of adminstration.

Since 2002, the State Archives has adopted the strategy of storing database content in a combination of CSV and XML files, screenshots of the graphical user interface (GUI), and handbooks or tutorials. The source databases had many aspects that were chosen not to survive : the database management system (DBMS) with its full contents, the business logic layer and the GUI itself. Those aspects were deemed not important enough for the users of the decades and centuries to come.

Geographic vector data was treated in a similar way, but relying on ESRI Shapefiles instead of CSV tables. As with textual databases, some aspects of Geographic Information Systems, like topological relations, were discarded. Others features like signature encodings were not preserved as coded information but only as textual documents and manuals encoded in PDF.

However, in the past 10 years, progress has been made in a special domain of virtualisation, the emulation business. In alliance with researchers worldwide, the bwFLA and EMiL (Steinke 2016) projects have gathered enough insights to see the outlook of software stacks that revive obsolete databases and also geoinformation systems in all aspects deemed necessary.

At the same time, the Digital Information LifeCycle Interoperability Standards Board (DILCIS.eu) and the E-ARK project are investing effort into making the SIARD 2.1 format a standard way of preserving database content in the European Union.This was why we have set up a worldwide challenge in order to estimate how we should think about database preservation in the next 20 years.

The challenge still goes on. The State Archives will organize a workshop at Stuttgart on this issue in 2021.

Details on the challenge

The term database as defined for the challenge comprises the graphical user interface (GUI) developed for the concrete database, the business logic and the contents of the database management system (DBMS). DB2, MariaDB, MS Access 2007, MySQL and Oracle are represented at DBMS. Non-relational database concepts are not part of the challenge. Future use should

range from simple querying by non-specialists to advanced re-use on the DBMS technology of 2080 or later.

We stated that it may be wise to limit the preservation of certain components of the databases in the interest of cost efficiency. In papers published during the last 25 years, there have been several approaches of this kind:

- limit the extent of the preserved data to the essential data objects by de-normalizing (Keitel 2004, Ohnesorge et al. 2016) or modeling key objects or processes and their dimensions (Ur Rahman et al. 2010)
- only preserve manuals on the GUI or record snapshots or videos in well documented formats on how the GUI works, not the GUI itself (Keitel 2004, Ur Rahman 2010a)
- migrate the DBMS contents into a common SQL format that lacks certain properties of the original DBMS. This SQL conceptual model is wrapped in XML (Faria et al. 2016, Ferreira 2016, Fitzgerald 2013, SIARD 2019).

Solutions should make use of existing tools, processes and standards. There should be mechanisms to document the steps that were taken. The challenge is aimed at web archivists, computer scientists, information scientists, librarians, archivists and researchers of all kinds who work with databases.

Political and legislative issues

Intellectual property (IP) legislation is poorely prepared for obsolesence. Unlike orphaned books, a piece of orphaned, formerly commercial software cannot be used openly in most parts of the world because of risks looming from unclear IP claims. Also, there is little legislation about who is responsible for preserving software. The EU DSM directive has recently moved into a good direction, but work has to continue in order to assure a risk-free environment for the emulation approaches described in this paper.

Solutions presented so far or found in literature

The challenge first issued in April 2020 did not receive many answers yet, but we found out that researchers are keen to learn about the outcome. A very handy input was a paper for the DH2020 conference by C. Neuefeind et al. that describes the ways researchers deal with preservation of their research databases. In this paper, let us only remember two classic solutions and have a look at three novel solutions sent to me. All of them will be evaluated for cost-efficiency and robustness.

The CSV solution

This solution already has been detailed in the first chapter of this paper. It needs the biggest investment before ingest and relies on the least promises about what future users and future machines can do. It doesn't require any regular maintenance during storage. The investment for reviving the data is rather small since it requires the least specialized software.

The XML solution

This solution has been developed independently at the Swiss Federal Archives and the German company CSP and is fully described by Lindley (2013). It needs an investment before ingest for transfer of the database. It may be combined with appraisal and preparation of simplifyed archival database objects. During storage, the holding institution needs to verify whether the unpacking solution for the XML dialect chosen is still available for transfer to a living system, e.g. every five years.

The disk image solution

This solution is described by Cochrane et al. (2013). Disk images are made of the server and the client computers. They are supposed to be running on two separate emulated software and hardware stacks composed of the computer hardware, BIOS, operating system, and DBMS. If documentary material like manuals comes along, there must also be viewers for them. This method needs an investment before ingest, and regular checks for availability of the complete emulative stack. Investment for use is supposed to be low as long as the stack is working. There is, however, a risk involved. The loss of stack components can make the database object practically lost because nobody is willing to pay for re-programming the lacking component.

The Docker image solution

This solution is similar to the disk image solution, but it adopts it to the Docker technology. The solution provider suggests that both client and server software be migrated to Docker containers. Once again, server and client rely on the availability of the whole emulative stack for 60 years.

This way of using emulation might reduce the cost of the associated services since system images have easier ways of interaction. Cost and risks are similar to the disk image solution, unless the Docker container technology becomes obsolete by itself.

The web crawler solution

This solution will only work for databases that come with a web-based frontend that shows a complete list of their content objects. It requires a web crawler that can harvest deeply and broadly enough to gather all the datasets the database has to offer. The result, composed of HTML and JavaScript, will be revisited every twenty years in order to assure accessibility. The participant suggests that alongside this method, the frontend is harvested by another independent institution like the Internet Archive and a comprehensive dataset is exported as CSV for quality assurance.

This solution poses little intellectual property risks since only open source technologies are involved. Investment for maintenance is rather low, but there is a need to regularly assure software availability. Expected investment for revival is rather low as well. However, the databases the challenge targets are supposed to be confidential and thus will normally lack an HTML/JavaScript interface.

Literature

This list does not only show the cited works but every item that seemed of interest during the investigations for the paper.

Anderson, B., Braxton, S., Imker, H., Popp, T. (2018), The Art of Preserving Scientific Data: Building Collaboration into the Preservation of a Legacy Database (iPRES 2018 Paper), https://doi.org/10.17605/OSF.IO/RH9SU.

Bewertung eletronischer Fachverfahren (2015). Arbeitspapier des VdA-Arbeitskreises Archivische Bewertung. In: Archivar 2015(1), p. 90-92. https://www.archive.nrw.de/archivar/hefte/2010/index.html.

Cha, S.-J., Choi, Y. J., Lee, K.-C. (2015), Development of Preservation Format and Archiving Tool for the Long-Term Preservation of the Database (IMCOM 2015 Paper) https://dl.acm.org/doi/10.1145/2701126.2701192.

Cochrane E., Suchodoletz, D., Crouch, M. (2013). Database Preservation Using Emulation – a Case Study. In: Archifacts 2013, p. 80-95.

Christophides, V., Buneman, P. (2007): Report on the First International Workshop on Database Preservation (PresDB'07), https://doi.org/10.1145/1324185.1324197

Däßler, R.; Schwarz, K. (2010). Archivierung und dauerhafte Nutzung von Datenbanken aus Fachverfahren – eine neue Herausforderung für die digitale Archivierung. In: Archivar 2010(1), p. 6-18. https://www.archive.nrw.de/archivar/hefte/2010/index.html.

Dekeyser, K. (2012), User story: archiving a FoxPro database, OPF Blog, https://openpreservation.org/blogs/user-story-archiving-foxpro-database/

Dorendorf, S., (2007). Kosten und Nutzen von Datenbankreorganisationen: Grundlagen, Modelle, Leistungsuntersuchungen. In: Kemper, A., Schöning, H., Rose, T., Jarke, M., Seidl, T., Quix, C. & Brochhaus, C. (Hrsg.), Datenbanksysteme in Business, Technologie und Web (BTW 2007) – 12. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS). Bonn: Gesellschaft für Informatik e. V. (S. 397-416). https://dl.gi.de/handle/20.500.12116/31812

Faria, L., Büchler, M., Aas, Kuldar (2016), Workshop on Relational Database Preservation Standards and Tools. (iPRES 2016 Paper), https://doi.org/11353/10.502816.

Ferreira, B., Faria L., Ferreira M., Ramalho J. C. (2016), Database Preservation Toolkit. A relational database conversion and normalisation tool (iPRES 2016 Paper). https://hdl.handle.net/11353/10.503182

Fitzgerald, N. (2013), Using data archiving tools to preserve archival records in business systems - a case study (iPRES 2013 Paper), https://hdl.handle.net/11353/10.378094

Heuscher, S., Jährmann, J., Keller-Marxer, P., Möhle, F. (2004). Providing authentic long-term archival access to complex relational data. In: European Space Agency Symposium ”Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data”, October, Frascati, Italy, 2004.

Keitel, C. (2004), Erste Erfahrungen mit der Langzeitarchivierung von Datenbanken. Ein Werkstattbericht. In: Hering, R., Schäfer, U. (Ed.), Digitales Verwalten – Digitales Archivieren. 8. Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen“. https://dx.doi.org/10.15460/HUP.STAHH.19.82

Klopprogge, M. R., Lockemann, P. C. (1983) Modelling Information Preserving Databases: Consequences of the Concept of Time, in: VLDB '83: Proceedings of the 9th International Conference on Very Large Data Bases, S. 399–416.

Lindley, A. (2013) : Database Preservation Evaluation Report - SIARD vs. CHRONOS. Preserving complex structures as databases through a record centric approach? iPRES 2013 Proceedings, https://dblp1.uni-trier.de/db/conf/ipres/ipres2013.html

Moore, R. W. (2018) : Archiving Experimental Data. Encyclopedia of Database Systems (2nd ed.) 2018.

Müller, H. (2009) Archiving and Maintaining Curated Databases, Rostock: Universität, 2009, https://doi.org/10.18453/rosdok_id00002203

Neuefeind, C., Schildkamp, P., Mathiak, B., Karadkar, U., Stigler, J., Steiner, E., Vasold, G., Tosques, F., Ciula, A., Maher, B., Newton, G., Arneil, G., Holmes, M. (2020), Sustainability Strategies for Digital Humanities Systems, DH2020 Panel Paper https://dh2020.adho.org/wp-content/uploads/2020/07/565_SustainabilityStrategiesforDigitalHumanitiesSystems.html

Ohnesorge K. W., Aas K., Delve J., Lux Z., Tømmerholt P. M., Nielsen A. B., Büchler M. (2016), Tutorial on Relational Database Preservation (iPRES 2016 Paper), https://doi.org/11353/10.502822.

Olson, J. E. (2010). Database archiving: how to keep lots of data for a very long time. ISBN 978-0123747204

Olson, J. E. (2011). Data Quality and Database Archiving: The Intersection of Two Important Data Management Functions (5. MIT Information Quality Industry Symposium 2011 Presentation), http://mitiq.mit.edu/IQIS/Documents/CDOIQS_201177/Papers/01_05_T2B_Olson.pdf

Potthoff, J., (2012). Beweiswerterhaltendes Datenmanagement im elektronischen Forschungsumfeld. In: Müller, P., Neumair, B., Reiser, H. & Rodosek, G. D. (Ed.), 5. DFN-Forum Kommunikationstechnologien – Verteilte Systeme im Wissenschaftsbereich. Bonn: Gesellschaft für Informatik e.V.. (S. 109-118). https://dl.gi.de/handle/20.500.12116/18172

Rumianek, Michael (2013). Archiving and Recovering Database-driven Websites. D-Lib Mag. 19(1/2) (2013) http://www.dlib.org/dlib/january13/rumianek/01rumianek.html

SIARD-2.1.1 Format Specification (2019), https://www.bar.admin.ch/dam/bar/en/dokumente/kundeninformation/siard_formatbeschreibung.pdf.download.pdf/siard_format_descriptioning.pdf

Steinke, T., Padberg, F., Schoger, A., Rechert, K. (2016), Project EmiL – Emulation of Multimedia Objects (iPRES 2016 Paper), https://hdl.handle.net/11353/10.503170.

Ur Rahman, A., David, G., Ribeiro C. (2010). Model Migration Approach for Database Preservation, 12th International Conference on Asia-Pacific Digital Libraries ICADL, Proceedings, https://dx.doi.org/10.1007/978-3-642-13654-2_10

Whitt, R. S. (2017), "Through A Glass, Darkly" Technical, Policy, and Financial Actions to Avert the Coming Digital Dark Ages, 33 Santa Clara High Tech. L.J. 117 (2017). https://digitalcommons.law.scu.edu/chtlj/vol33/iss2/1

Zeller, B., Herbst, A. & Kemper, A., (2003). XML-Archivierung betriebswirtschaftlicher Datenbank-Objekte. In: Weikum, G., Schöning, H. & Rahm, E. (Hrsg.), BTW 2003 – Datenbanksysteme für Business, Technologie und Web, Tagungsband der 10. BTW Konferenz. Bonn: Gesellschaft für Informatik e.V.. (S. 127-146). https://dl.gi.de/handle/20.500.12116/30094

Improving the Quality of Web Harvests Using Web Curator Tool

Ben O'Brien
National Library of New Zealand
Wellington, NZ
Ben.obrien@dia.govt.nz

Hanna Koppelaar
National Library of the Netherlands
The Netherlands
Hanna.Koppelaar@KB.nl

Trienka Rohrbach
National Library of the Netherlands
The Netherlands
Trienka.Rohrbach@KB.nl

Andrea Goethals
National Library of New Zealand
Wellington, NZ
Andrea.goethals@ dia.govt.nz

Jeffrey van der Hoeven
National Library of the Netherlands
The Netherlands
Jeffrey.vanderHoeven@KB.nl

Steve Knight
National Library of New Zealand
Wellington, NZ
Steve.knight@ dia.govt.nz

Frank Lee
National Library of New Zealand
Wellington, NZ
Frank.lee@ dia.govt.nz

Charmaine Fajardo
National Library of New Zealand
Wellington, NZ
Charmaine.fajardo@ dia.govt.nz

ABSTRACT

In the field of web archiving, the quality of captures is important to assess soon after a crawl to allow the opportunity to import missing content before resources disappear from the live web. For this reason, current work on the Web Curator Tool (WCT) is focused on enhancing its quality assurance functions to be more efficient, scalable and usable.

The WCT is a free open source workflow management tool for selecting, crawling websites, performing quality assurance and preparing websites for ingest into a preservation system. Through close collaboration between the National Library of New Zealand and the National Library of the Netherlands the WCT has already undergone several important uplifts in the past two years that have updated the underlying technical underpinning. With that work done, the project team is now focused on functional enhancements desired not only by the KBNL and NLNZ, but also by the wider web archiving community.

Feedback from WCT users at recent International Internet Preservation Consortium (IIPC) conferences has led to the current program of work focused on improvements to the tool to help with quality review. Increasing and improving this capability within Version 4 of the WCT will be achieved through enhancements in four core areas - improved import and prune functionality, crawl patching using Web Recorder, integration with the Pywb viewer, and improved screenshot generation.

WADL'20, August, 2020, Beijing, China
 978-1-4503-0000-0/18/06...$15.00
https://doi.org/10.1145/1234567890

CCS CONCEPTS

• **Information systems → Digital libraries and archives**
• General and reference → Evaluation

KEYWORDS

Web archiving, quality assurance, Web, Web collections, data preservation, digital preservation

1 Introduction

As online presence is indispensable in our fast-changing world, web archives are an invaluable source of factual information about what was online at a certain moment. Online content is short lived however, so it is important to gain the highest "quality" when the site is crawled or soon after before the live content has changed or disappeared. Quality in the context of a web archive has been defined as the completeness of material archived within a target perimeter, and the ability to render the original form of the site [1]. Because of the scale of web archives, quality assurance (QA) would not be possible without automated techniques and tools. It is therefore that the work scheduled on the Web Curator Tool (WCT) for 2020, is focused on enhancing its QA functions to be more efficient, scalable and usable.

The WCT is a free open source [2] workflow management tool for selecting, crawling websites, performing QA and preparing websites for ingest into a preservation system. Through close collaboration between the National Library of New Zealand (NLNZ) and the National Library of the Netherlands (KBNL), the WCT has undergone several important uplifts in the past two years. Version 2 added support for Heritrix 3[1] and improved the project documentation adding new tutorials, installation and administration guides. Version 3 addressed a large volume of

[1] See https://github.com/internetarchive/heritrix3

technical debt, in which the underpinning frameworks were upgraded, creating a stable foundation to take the development of WCT forward. With that work done, the project team began to focus on functional enhancements desired not only by the KBNL and NLNZ, but also by the wider IIPC community.

Through tutorials and workshops [3] at recent IIPC conferences, the project team demonstrated the new versions of WCT and asked for feedback on the features other organisations would like to see incorporated into the WCT. The Hungarian National Library in particular contributed many enhancement ideas, several related to improving the WCT's quality control features. These dovetailed nicely with the WCT QA improvements already planned for the next release.

2 WCT Version 4 Enhancements

Increasing and improving the quality review capability within Version 4 of the WCT will be achieved through enhancements in four core areas - improved import and prune functionality, crawl patching using Web Recorder, integration with the Pywb viewer, and improved screenshot generation.

2.1 Improved Import and Prune Functionality

Prior to version 4, WCT displayed a harvest as a tree for QA review. In version 4, the visualization is replaced with an interactive network graph, supporting a deeper exploration of the harvest, while retaining the import and prune functionality provided previously. The visualization was designed to handle relatively large harvests (10-40 GB) by displaying network nodes at the domain level, e.g. one node for any crawled resource with the same domain, but allows the user to drill down to see the individuals URLs crawled under each domain. The result is that the indexing and display is efficient and does not impact the performance and stability of other functions within WCT.

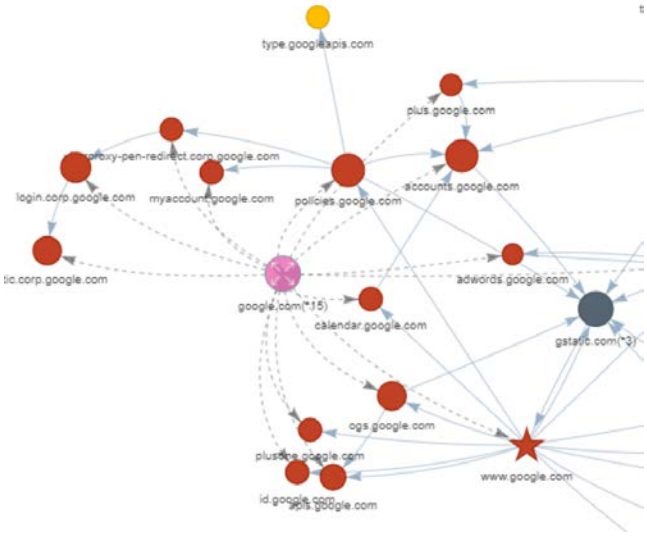


Figure 1: Partial screenshot of a harvest in the new WCT visualization, used for QA review

2.2 Crawl Patching Using Web Recorder

WCT uses the Internet Archive's Heritrix 3 crawler for the core of its crawling functionality. Heritrix was designed to crawl large quantities of content in bulk. It is less suitable for patch crawling, meant to add missing content back into a harvest as part of the QA process. Another open source crawling tool, Rhizome's Webrecorder[2], is efficient at capturing small amounts of missing content so was chosen for integration into WCT version 4. The integration will transfer newly patched content back into the WCT, incorporating it into the original web harvest.

2.3 Integration with the Pywb Viewer

The WCT already provides a WARC viewer and OpenWayback integration to browse harvests during QA, but recently Pywb[3], the Python-based web archive viewer, has become the benchmark for web archive replay. This integration of Pywb into WCT will provide WCT users with the best options available for web harvest replay and review. In addition, it is being implemented as a configurable setting so that WCT administrators can reconfigure the WCT web harvest viewers according to their institution's preferences for QA.

2.4 Improved Screenshot Generation

The WCT previously contained limited functionality to capture screenshots of a web harvest. Realizing the potential QA benefit, WCT version 4 will be enhanced to capture screenshots of live websites being crawled and the resulting web harvest for comparison. The integration of the screenshot software will be configurable, allowing for the use of 3rd party tools. We also intend to leverage the advancements in screenshot comparison metrics within the web archiving community [4][5].

ACKNOWLEDGMENTS

We thank the individuals and organisations over the last few years who have attended WCT tutorials and workshops and have given us detailed feedback on the features they would like to see enhanced or added to the WCT. We also thank the Internet Archive and Rhizome for use of their open source tools.

REFERENCES

[1] Julien Masanés. (Ed.) 2006. *Web archiving*. Springer-Verlag Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-540-46332-0

[2] National Library of the Netherlands, National Library of New Zealand, [n. d.]. Web Curator Tool http://webcuratortool.org/

[3] Ben O'Brien, Steve Knight, Hanna Koppelaar, Trienka Rohrbach and Jeffrey van der Hoeven. 2019. Web Curator Tool (WCT) Workshop, *IIPC General Assembly*, National and University Library in Zagreb, Zagreb, Croatia. http://netpreserve.org/ga2019/programme/abstracts/#workshop-WCT

[4] Brenda Reyes Ayala, Ella Hitchcock and James Sun. 2019. Using Image Similarity Metrics to Measure Visual Quality in Web Archives, *JCDL 2019*, Urbana-Champaign, Illinois.

[5] Justin Brunelle. 2018. 2018-09-03: Let's compare memento damage measures!, blog post, Old Dominion University, https://ws-dl.blogspot.com/2018/09/2018-09-03-lets-compare-memento-damage.html

[2] See https://guide.webrecorder.io/
[3] See https://github.com/webrecorder/pywb

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.

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

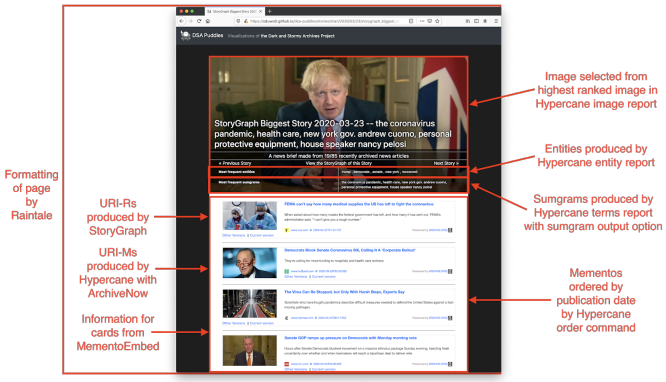


Figure 1: The "biggest news story" of for the March 23, 2020 story produced by SHARI[2]. Annotations detail which components provide each part of the visualization.

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,

[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.

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

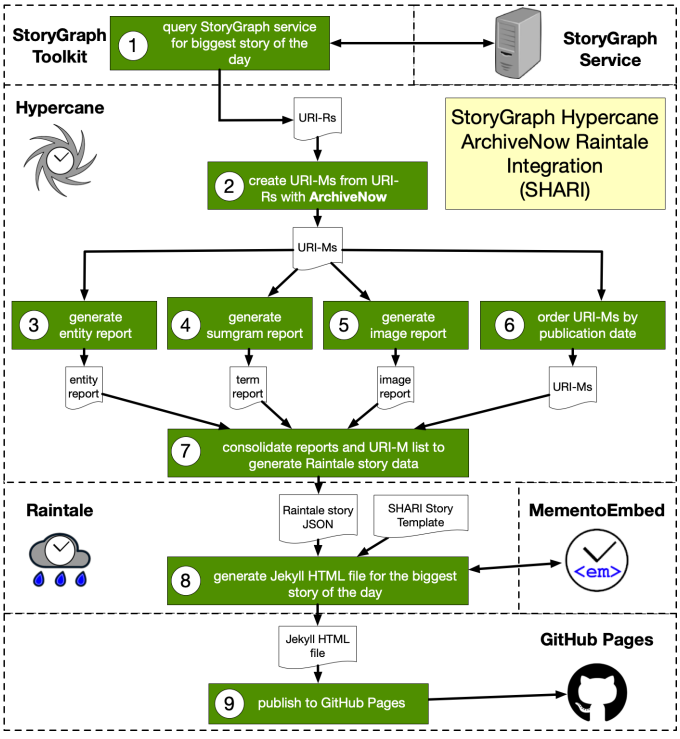


Figure 2: SHARI process for creating a visualization of the biggest news story for a given day

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

MementoEmbed and Raintale for Web Archive Storytelling

Shawn M. Jones, Martin Klein
{smjones,mklein}@lanl.gov
Los Alamos National Laboratory, Los Alamos, NM

Michele C. Weigle, Michael L. Nelson
{mweigle,mln}@cs.odu.edu
Old Dominion University, Norfolk, VA

ABSTRACT

For traditional library collections, archivists can select a representative sample from a collection and display it in a featured physical or digital library space. Web archive collections may consist of thousands of archived pages, or mementos. How should an archivist display this sample to drive visitors to their collection? Search engines and social media platforms often represent web pages as cards consisting of text snippets, titles, and images. Web storytelling is a popular method for grouping these cards in order to summarize a topic. Unfortunately, social media platforms are not archive-aware and fail to consistently create a good experience for mementos. They also allow no UI alterations for their cards. Thus, we created MementoEmbed to generate cards for individual mementos and Raintale for creating entire stories that archivists can export to a variety of formats.

KEYWORDS

web archives, memento, storytelling, visualization, summarization

1 INTRODUCTION

Trying to understand the differences between web archive collections can be onerous. Thousands of collections exist [7], collections can contain thousands of documents, and many collections contain little metadata to assist the user in understanding their contents [8]. How can an archivist display a sample of a collection in order to drive visitors to their collection or provide insight into their archived pages?

Search engines and social media platforms have settled on the card visualization paradigm, making it familiar to most users. Web storytelling is a popular method for grouping these cards to summarize a topic, as demonstrated by tools such as Storify. For this reason, AlNoamany et al. [1] made Storify the visualization target of their web archive collection summaries. Because Storify shut down in 2018 [3], we evaluated fifty alternative tools, such as Facebook, Pinboard, Instagram, Sutori, and Paper.li [4]. We found that they are not reliable for producing cards from mementos. Thus we developed MementoEmbed [5], an archive-aware service that can generate different surrogates [2] for a given memento. Currently supported surrogates include social cards, browser thumbnails (screenshots) [9], word clouds, and animated GIFs of the top ranked images and sentences. MementoEmbed's cards appropriately attribute content to a memento's original resource separately from the archive, including both the original domain and its favicon from the memento's time period, as well as providing a striking image, a text snippet and a title. MementoEmbed provides an extensive API that helps machine clients request specific information about a memento. Raintale [6] leverages this API to generate complete stories containing the surrogates of many different mementos.

2 THE MEMENTOEMBED-RAINTALE ARCHITECTURE FOR STORYTELLING

Figure 1 demonstrates the relationship between MementoEmbed and Raintale. In step 1, the user provides a template and a list of URI-Ms to Raintale. In step 2, Raintale records all template variables. For each provided URI-M, Raintale consults MementoEmbed's API for the value of each variable in the corresponding memento. In step 3, MementoEmbed downloads the memento from its web archive and performs natural language processing, image analysis, or extracts information via the Memento Protocol [10], as appropriate to the API request. In step 4, Raintale consolidates the data from these API responses and renders the template with the gathered data, producing a story constructed from surrogates and other supplied content (e.g., story title, collection metadata). Raintale can produce many output formats, including HTML (Figure 2), Markdown, MediaWiki, Jekyll, and Twitter Threads (Figure 3).

3 CONCLUSIONS

We introduced MementoEmbed for generating surrogates for single mementos and Raintale for generating complete stories of memento sets. We envision Raintale and MementoEmbed to be critical components for summarizing collections of archived web pages through visualizations familiar to general users. We developed MementoEmbed so that its API is easily usable by machine clients. Raintale lends itself to incorporation into existing automated archiving workflows. Archivists can leverage this form of storytelling to highlight a specific subset of mementos from a collection. They can advertise their holdings, feature specific perspectives, focus on individual mementos, or help users decide if a collection meets their needs.

REFERENCES

[1] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. 2017. Generating Stories From Archived Collections. In *WebSci 2017*. Troy, New York, 309–318. https://doi.org/10.1145/3091478.3091508
[2] Robert Capra, Jaime Arguello, and Falk Scholer. 2013. Augmenting web search surrogates with images. In *CIKM 2013*. San Francisco, California, USA, 399–408. https://doi.org/10.1145/2505515.2505714
[3] Shawn M. Jones. 2017. Storify Will Be Gone Soon, So How Do We Preserve The Stories? https://ws-dl.blogspot.com/2017/12/2017-12-14-storify-will-be-gone-soon-so.html
[4] Shawn M. Jones. 2017. Where Can We Post Stories Summarizing Web Archive Collections? http://ws-dl.blogspot.com/2017/08/2017-08-11-where-can-we-post-stories.html
[5] Shawn M. Jones. 2018. A Preview of MementoEmbed: Embeddable Surrogates for Archived Web Pages. https://ws-dl.blogspot.com/2018/08/2018-08-01-preview-of-mementoembed.html
[6] 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
[7] 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
[8] Shawn M. Jones, Michele C. Weigle, and Michael L. Nelson. 2019. Social Cards Probably Provide For Better Understanding Of Web Archive Collections. In *CIKM 2020*. Beijing, China, 2023–2032. https://doi.org/10.1145/3357384.3358039
[9] Theodorich Kopetzky and Max Mühlhäuser. 1999. Visual preview for link traversal on the World Wide Web. *Computer Networks* 31, 11-16 (May 1999), 1525–1532.

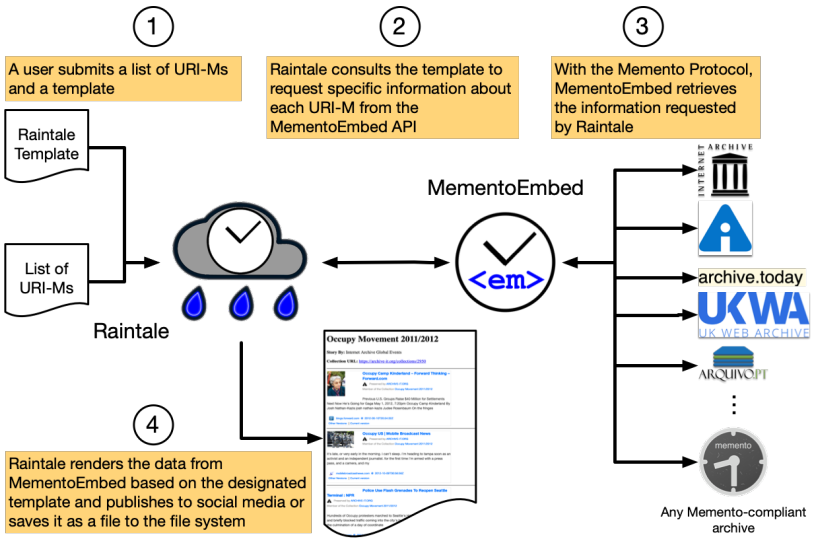


Figure 1: The MementoEmbed-Raintale Architecture for Storytelling

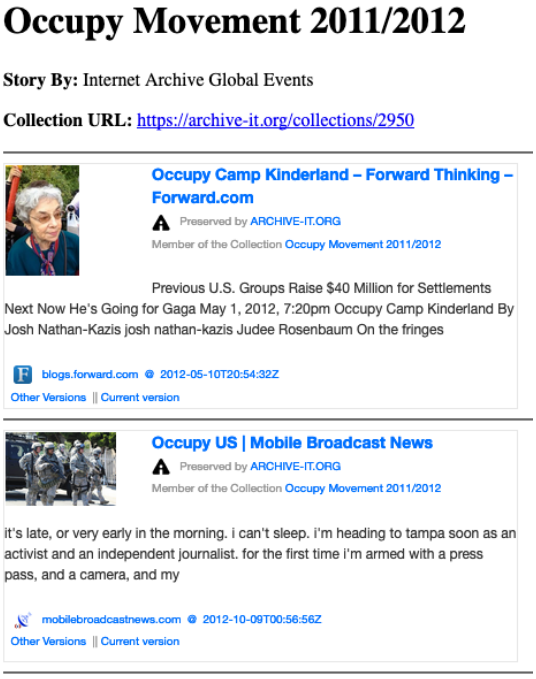


Figure 2: Raintale can render a multiple mementos as an HTML story of cards.

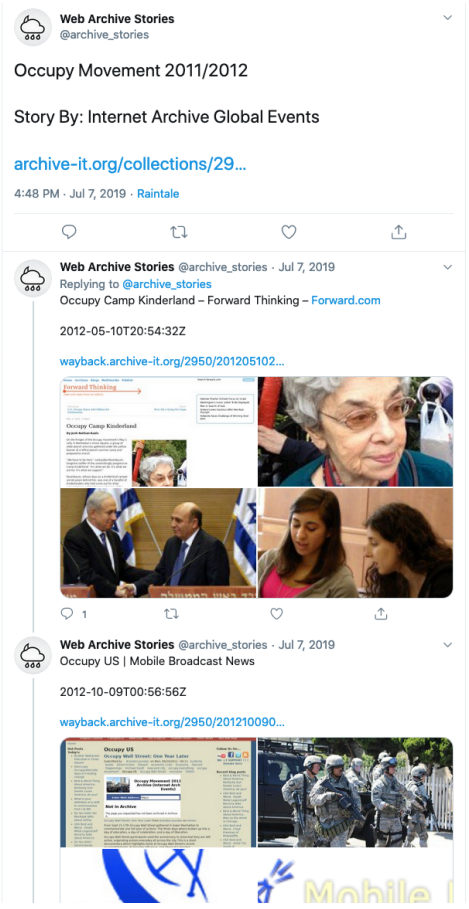


Figure 3: Raintale can render multiple mementos as a Twitter thread.

https://doi.org/10.1016/S1389-1286(99)00050-X

[10] 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

TMVis: Visualizing Webpage Changes Over Time

Abigail Mabe, Dhruv Patel, Maheedhar Gunnam, Surbhi Shankar
Department of Computer Science
Old Dominion University
Norfolk, VA 23529 USA
{amabe002,dpate006,mgunn001, sshan001}@odu.edu

Mat Kelly
College of Computing & Informatics
Drexel University
Philadelphia, PA 19104 USA
mkelly@drexel.edu

Sawood Alam, Michael L. Nelson, Michele C. Weigle
Department of Computer Science
Old Dominion University
Norfolk, VA 23529 USA
{salam,mln,mweigle}@cs.odu.edu

ABSTRACT

TMVis is a web service to provide visualizations of how individual webpages have changed over time. We leverage past research on summarizing collections of webpages with thumbnail-sized screenshots and on choosing a small number of representative archived webpages from a large collection. We offer four visualizations: Image Grid, Image Slider, Timeline, and Animated GIF. Embed codes for the Image Grid and Image Slider can be produced to include these visualizations on separate webpages. This tool can be used to allow scholars from various disciplines, as well as the general public, to explore the temporal nature of webpages.

1 INTRODUCTION

Users of web archives may be interested in a quick overview of the topic of a particular webpage in a collection or in observing how that webpage has changed over time. Many web archive interfaces provide only a textual list of archived versions, or *mementos*, which requires that each one be explored individually. This can be a tedious process as the number of mementos for each webpage grows. As a first step to address this problem, we have developed a web service, TMVis, providing visualizations that show how a single webpage has changed over time. We do this by using the list of mementos of a webpage (*i.e.*, a TimeMap) and choosing to display the screenshots, or thumbnails, of the most unique mementos, as a visual summarization.

In our previous work [1], we developed a technique to determine which mementos are useful to display. The algorithm compares the amount of change (based on SimHash) in the HTML of consecutive mementos and selects the most unique. This process is much more efficient, in both time and space, than generating all thumbnails and then performing image processing. We use this same process for TMVis. To improve the response time for large TimeMaps with over 1000 mementos, we first sample a maximum of 1000 mementos before continuing the selection process.

The web service is available at http://tmvis.cs.odu.edu/, and the source code is hosted on GitHub at https://github.com/oduwsdl/tmvis. We have published a tech report with a full description [2] and a blogpost[1] with more images and a video walkthrough.

2 TMVIS WEB SERVICE

The TMVis web service asks the user for a URI and allows them to choose to process TimeMaps from the Internet Archive or from a collection in Archive-It[2]. After clicking the "View TimeMap" button, the TimeMap is loaded and the user is presented with a histogram displaying the number of mementos available over time. The user may use the histogram to select a date range in the TimeMap to summarize, or they may choose to summarize the entire TimeMap.

After choosing the desired time range, the user clicks the "Calculate Unique" button. The service downloads the HTML source and computes the SimHash of the requested mementos. Then the SimHashes of the mementos are compared according to AlSum et al. [1]. We present the user with several options for the threshold of "uniqueness" so that they may create larger or smaller visualizations as desired. Once the user chooses their desired number of thumbnails and clicks the "Generate Thumbnails" button, the service uses Puppeteer[3] to render each memento and capture a thumbnail screenshot. Once the thumbnails are captured for all the mementos, they are presented to the user with the four visualization widgets shown in Figures 1-3. The default view shows the Image Grid, however users can switch between tabs to view the other visualizations: Image Slider, Timeline, and Animated GIF.

Image Grid. The Image Grid shows the thumbnails of the representative mementos arranged in a left to right, top to bottom manner. Clicking on any thumbnail loads the source memento. On the top right of each thumbnail in the grid is a refresh button to allow users to regenerate the thumbnail if it appears incomplete and an 'X' to allow users to remove thumbnails from the visualizations. Figure 1 shows an example of the Image Grid for http://www.odu.edu/.

Image Slider. The Image Slider imitates the photo slider functionality used in Apple's iPhoto[4]. By moving the cursor across the thumbnail image, the next thumbnail is displayed. As with the Image Grid, clicking on the thumbnail loads the source memento. The user can cycle through the thumbnails by clicking arrow buttons to the left and right of the slider. Figure 2 shows a static example of the Image Slider for http://columbia.edu/cu/english/.

Timeline. The Timeline view arranges the thumbnails according to the mementos' datetimes (time of capture). The Timeline view includes zoom, next, previous, next unique, and previous unique buttons to allow the user to navigate between the unique and regular mementos. The unique mementos are represented with yellow stripes and the others are represented with gray stripes on the

[1] https://ws-dl.blogspot.com/2020/05/2020-05-21-visualizing-webpage-changes.html

[2] TMVis can be expanded to support any Memento-compatible public archive or Memento aggregator.

[3] https://developers.google.com/web/tools/puppeteer

[4] http://web.archive.org/web/20150101033528/http://apple.com/mac/iphoto/

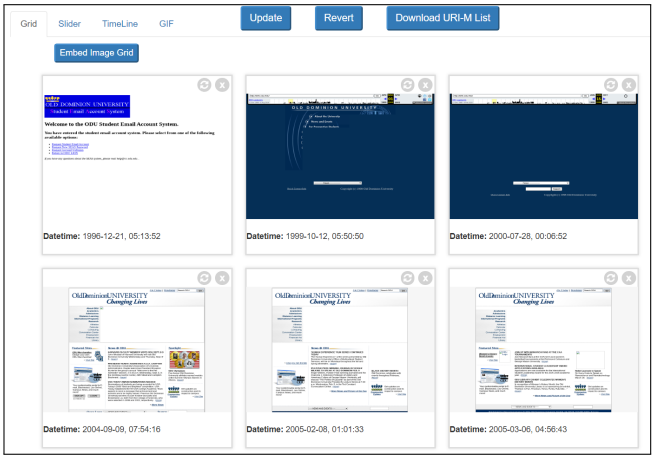


Figure 1: Image Grid for http://www.odu.edu/

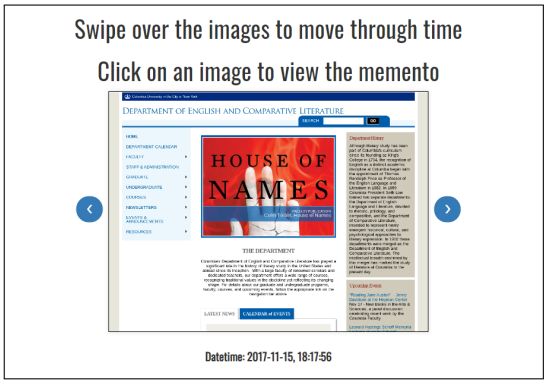


Figure 2: Image Slider for http://columbia.edu/cu/english/

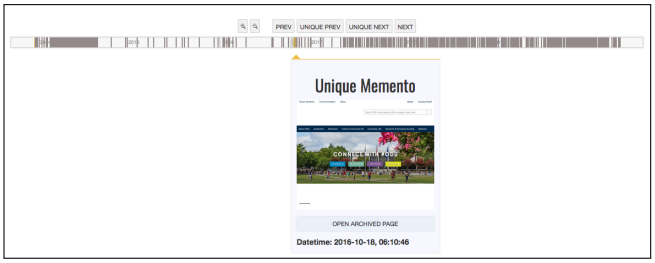


Figure 3: Timeline view for http://www.odu.edu/

timeline. Non-unique mementos do not have a screenshot, but are represented by the thumbnail of the previous unique memento with an indication that this memento is "similar to" the previous unique memento. The Timeline view is based on Timeline Setter library[5], developed by ProPublica. Figure 3 shows an example of the Timeline view for http://www.odu.edu.

Animated GIF. The Animated GIF visualization, shown in Figure 4, is built from the thumbnails in the Image Grid using the GifShot

[5]http://propublica.github.io/timeline-setter/

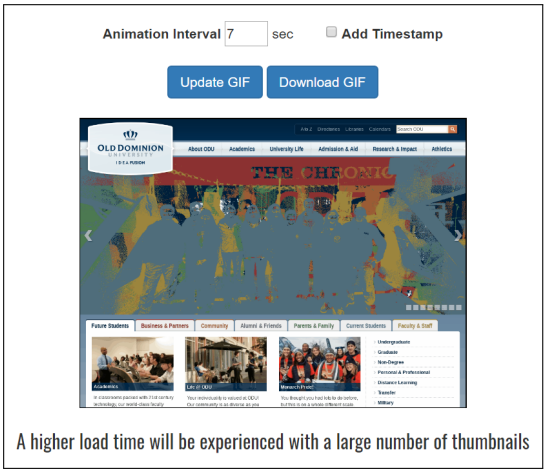


Figure 4: Animated GIF for http://www.odu.edu/

library[6]. The user can include a timestamp on each thumbnail or adjust the time interval between each frame of the animation. This GIF can be downloaded by clicking the "Download GIF" button.

Embeddable Visualizations. We have also provided options for users to embed the Image Grid and Image Slider visualizations in their own webpages as well as an option to download the animated GIF for inclusion in other webpages. An example of embedding these elements in a simple webpage is provided at https://ws-dl.cs.odu.edu/vis/tmvis/embed-examples.html.

3 CONCLUSION

We have presented a description of TMVis, a web service that provides four visualizations of how a webpage changes through time. We compare the difference in the SimHashes of the HTML source of pairs of mementos to determine which set of mementos is the most unique. Then we render and capture thumbnail-sized screenshots of the chosen mementos. These are then displayed as an Image Grid, Image Slider, Timeline, and Animated GIF. We hope that these visualizations will just be the beginning and will provide a starting point for others to expand these types of offerings for users of web archives.

ACKNOWLEDGEMENTS

This work has been supported by a NEH/IMLS Digital Humanities Advancement Grant (HAA-256368-17). We are grateful for this support and for the input from our partners at the Frick Art Reference Library, New York Art Resources Consortium, and Columbia University Libraries.

REFERENCES

[1] Ahmed AlSum and Michael L. Nelson. 2014. Thumbnail Summarization Techniques for Web Archives. In *Proceedings of the European Conference on Information Retrieval (ECIR)*. Amsterdam, 299–310.

[2] Abigail Mabe, Dhruv Patel, Maheedhar Gunnam, Surbhi Shankar, Mat Kelly, Sawood Alam, Michael L. Nelson, and Michele C. Weigle. 2020. *Visualizing Webpage Changes Over Time.* Technical Report arXiv:2006.02487. https://arxiv.org/abs/2006.02487

[6]https://yahoo.github.io/gifshot/