Report on the Workshop on Web Archiving and Digital Libraries (WADL 2013)

Edward A. Fox
Virginia Tech
fox@vt.edu

Mohamed M. Farag
Virginia Tech
mmagdy@vt.edu

Abstract

This workshop explored the integration of Web archiving and digital libraries, so the complete life cycle involved is covered, from creation/authoring, uploading/publishing in the Web (including Web 2.0), (focused) crawling, curation, indexing, exploration (including searching and browsing), (text) analysis, archiving, and up through long-term preservation. It included particular coverage of current topics of interest: challenges facing archiving initiatives, archiving related to disasters, interaction with and use of archive data, applications on an international scale, working with big data, mobile Web archiving, temporal issues, Memento, and SiteStory.

1 Introduction

At the end of the 2013 ACM/IEEE-CS Joint Conference on Digital Libraries in Indianapolis, the WADL 2013 workshop ran on the afternoon of Thursday July 25 and the morning of Friday July 26, with most attendees also meeting for dinner on Thursday. There were 16 attendees: administrators, faculty, librarians, researchers, and students. Representation included from Ball State University, Brazilian Development Bank, Harding University, Los Alamos National Laboratory, Old Dominion University, Stanford University, United Nations, UCLA Library, and Virginia Tech. There were 14 presentations, summarized in the next section, some given by groups. In addition, there were short personal introductions by other attendees, as well as a final plenary discussion, so everyone was engaged.

It became clear that Web archiving is very important, adding a temporal dimension to the Web. This is essential for historians, and for those in the future who will seek to understand the evolution of the modern world, since so much of the knowledge, culture, events, scholarship, and other activities of humanity often only has a fleeting presence on the Web, and will be lost if there is not a comprehensive and systematic effort to develop reliable and persistent archives. Further, to be useful, these must be supported by services enabling access, analysis, and interactive use.

It is a promising sign that the research, development, library, information science, computer science, and archiving communities are joining to address these challenges, which fit well with work on digital libraries and information retrieval. It is hoped that this report will inspire others to tackle the many challenges in Web archiving that still need addressing.

2 Presentations

2.1 ArcSpread: Enabling Web Archive Analysis for non-CS Experts, by Andreas Paepcke

Andreas Paepcke (Stanford) presented a vision of how sociologists, political scientists, and historians might analyze Web archives in the future. His project, ArcSpread [10], designs and implements a spreadsheet-based approach to the problem. Andreas worked through a hypothetical example using mockup components. Pieces of the three-tier architecture are implemented, but work remains around the interaction components, visualization tools, and the underlying distributed compute engine.

2.2 Applying Web Archives to Real-time Group Source Prediction of Speech, by Andreas Paepcke

Andreas Paepcke also talked about one problem faced by a friend (Henry). Though an active and productive businessman, a sudden illness left him unable to speak, and a quadriplegic. His conversation partners would get bored waiting for him to finish typing with a head tracking device onto an onscreen keyboard. In an effort to engage these conversation partners, Andreas and his team generated word trees that attempt to predict what his friend (Henry) is going to say, given his previously typed word [8]. For the prediction statistics, they used three different underlying collections: Henry's Web blog, a specialized crawl of 10M webpages, and a collection of 11K 10 minute phone conversations. Evaluation studies have compared the outcomes, so far purely regarding their effectiveness to produce word trees. Discussion led to some suggestions for enhancements, extensions, and additional evaluation.

2.3 United Nations digital repository for UN documentation, by Ylva Braaten

Ylva Braaten (United Nations) talked about the work at the UN Library in New York, changing its indexing processes and moving to a digital repository for UN documentation (parliamentary documents, conference related documents, publications, etc). She presented a status report on UN movement towards more digital and automated processes, involving different departments, agencies, and duty stations.

2.4 SiteStory, Archiving Done Differently, by Martin Klein and Justin Brunelle

Martin Klein (Los Alamos National Laboratory) provided an overview of the SiteStory [9] concept and its functionality compared to crawler-based archiving. He discussed the SiteStory approach and described various use cases. He and Justin Brunelle further introduced the SiteStory testbed and provided insight into novel results of benchmarking experiments.

2.5 Hiberlink, Towards Time Travel for the Scholarly Web, by Martin Klein

Martin Klein also gave a brief overview of the recently launched Hiberlink project [7]. This project aims at quantifying the "citation rot" problem in scholarly articles, that is occurring at unprecedented scale, but also at proposing solutions for researchers and publishers to ensure the longevity of the content of research.

2.6 Temporal User Intention Modeling in Social Media, by Hany SalahEldeen

Hany SalahEldeen (Old Dominion University) talked about modeling temporal user intentions in social media. The Web is stuck in the "perpetual now," and Web resources are prone to change, relocation, and deletion. An author could share a resource on his social network at a point in time in order to convey a certain message, having a specific intention in mind. After a period of time, if the state of the resource differed, the reader who reads the author's post and examines the resource might not see and understand what the author intended. This change of intention and the resource's state could cause a significant inconsistency in the published content. Hany's goal is to model the author's intention across time, and make predictions to avoid the temporal inconsistency that might occur.

2.7 Needs and Obstacles for a Web Archiving Initiative at the Ball State University Libraries, by Michael Szajewski

Michael Szajewski (Ball State University) summarized the digital library program at Ball State University, including a ContentDM and DSpace repository. He focused on the opportunities for Ball State University Libraries to develop an initiative to capture, preserve, and provide access to archived dynamic Web content. He discussed requirements, needs, and expectations for such an initiative, including the ability for the program to clearly support teaching and scholarship in a pragmatic way and the ability to preserve student digital scholarship. Obstacles at both the university and library levels were discussed.

2.8 Who and What Links to the Internet Archive, by Yasmin AlNoamany

Yasmin AlNoamany (Old Dominion University) presented the results of research (with Ahmed AlSum, Michele Weigle, and Michael Nelson) on Internet Archives Wayback Machine access logs, trying to answer some questions such as what users are looking for, why they come to IA, where they come from, and how pages link to IA. Her description of the most used languages found in Web archives led to a broad discussion.

2.9 Web Archiving Profile Overview, by Ahmed AlSum

Ahmed AlSum (Old Dominion University) presented his research results of profiling the existing Web archives for top-level domains and content languages. He used these results to

build a profile for each Web archive and used these profiles to optimize the query routing for the Memento Aggregator.

2.10 Archiving the Mobile Web, by Frank McCown, Monica Yarbrough, and Keith Enlow

Frank McCown (Harding University), with two students, explained that the Web is going mobile, and archivists want to capture this ephemeral content before it disappears. But archiving the Mobile Web isn't always as straightforward as it might seem; there are many hard research challenges. In his talk he shared his work on developing tools for Web archivists to automate the discovery and archiving of mobile websites.

2.11 Temporal Spread in Archived Composite Resources, by Scott Ainsworth

Scott Ainsworth (Old Dominion University) talked about the temporal spread of archived content. When a user retrieves a page from a Web archive, the page is marked with the capture datetime of the root-level resource, which effectively asserts "this page looked like this at a particular point in the past." However, the embedded resources, such as images and stylesheets, are nearly always archived at different times, although their capture time is not displayed to the user. The resulting presentation gives the appearance of coherent, temporally-aligned result, but can actually be composited from resources captured over a wide range of datetimes. He examined the temporal spread of composite archived resources (root plus embedded resources). He found that composite resources average 61.0 - 75.6% complete and have a mean temporal spread of 200.1 - 211.6 days, depending on heuristic and source policy.

2.12 The role of BNDES in Brazil preservation of heritage, by Fernanda Balbi

Fernanda Balbi (Brazilian Development Bank) presented a brief explanation on the history of BNDES and its role in Brazilian Development. She talked about the BNDES performance in projects addressing preservation of the architectural heritage, preservation of collections, and strengthening of relevant cultural institutions, including digital library projects.

2.13 Measuring Archivability of Web Resources and the Damage when Mementos are Missing, by Justin Brunelle

Justin Brunelle (Old Dominion University), with slide title "How I spend my summer vacations," discussed his current research efforts in measuring the archivability of Web resources and measuring damage that occurs in mementos. The archivability portion of the talk discussed what makes pages more or less archivable and how archivability is changing over time. The damage portion of the talk discussed how we can measure memento importance and how we can measure the impact that a missing memento has on a larger resource.

2.14 Crisis, Tragedy, and Recovery Network Digital Library (CTRnet) + Web Archiving in Qatar and VT, by Edward A. Fox, Seungwon Yang, and the CTRnet Team

Edward Fox (Virginia Tech) talked about the CTRnet [4] and IDEAL [5] projects, as well as about archiving plans in Qatar and at Virginia Tech. The CTRnet project [4] has been archiving tweet and webpage collections related to various types of disasters that affect communities. Research relates to social media use during crises, visualizing emergency phases or water main breaks, focused crawling, analyzing webpage collections with big data software, filtering using machine learning, and topic tagging. Thanks to new support for the Integrated Digital Event Archiving and Library (IDEAL) project [5], the aims are being broadened to also encompass other types of events, e.g., political or community focused. In Qatar, in collaboration with the National Library, there are plans to deploy tools like SeerSuite, Heritrix, and Solr, to collect, archive, and make accessible both scholarly content and the broader Web associated with the nation of Qatar [6]. At Virginia Tech, there are plans to further deploy Heritrix, as well as SiteStory and other tools, to help with campus Web archiving.

3 Discussion

After all the talks, there was a wide ranging plenary discussion. It became clear that investment in publishing on the Web, and innovation in publishing approaches, are extensive. On the other hand, most organizations, and many researchers, consider Web archiving as something that others will do for them. Further, they are unaware that a very small community, with quite limited resources, is at work on methods for Web archiving.

There was discussion of possible sponsorship to advance the state-of-the-art in Web archiving, including development of more advanced methods and services. There was discussion of possible collaboration among the workshop attendees, since there are a number of related efforts. Good synergies seemed likely to enable some of the essential advances.

4 Conclusions

There is a growing gap between methods for publishing/presenting on the Web, and methods for archiving. There is need for funding to advance Web archiving. There are many opportunities for information retrieval and digital library research to support Web archiving.

For more information, please see the workshop announcement [2] and the final website [3], which includes PowerPoint and other slides used for presentations, as well as audio recordings of presentations and discussion sessions. See also Ahmed AlSum's trip report available as a blog posting [1].

5 Acknowledgments

Thanks go to the Organizing Committee (Paul Bogen, Tessa Fallon, Kristine Hanna, Eric Hetzner, Gina Jones, Martin Klein, Frank McCown, Michael Nelson, and Andreas Paepcke) for helping with the advertising, planning, and running of the workshop; the JCDL conference

and organizing team also aided in these matters. Thanks go to the CTRnet and IDEAL project teams for their assistance, and to Virginia Tech for hosting the workshop website. Partial support was provided by the U.S. National Science Foundation through grants IIS-0916733 and 1319578. Additional support was provided by Qatar through NPRP 4-029-1-007.

6 References

1. Ahmed AlSum, 2013-07-26: Web Archiving and Digital Libraries Workshop - WADL 2013 Trip Report, blog article, Old Dominion University, 3 August 2013, http://ws-dl.blogspot.com/2013/08/2013-07-26-web-archiving-and-digital.html
2. Edward A. Fox, Web Archiving and Digital Libraries (WADL 2013): A JCDL2013 Workshop (http://jcdl2013.org/workshops), webpage, Virginia Tech, 2013, http://www.ctrnet.net/wadl2013
3. Edward A. Fox and Mohamed M. Farag. WADL 2013 website, with presentation slides and audio recordings, Virginia Tech, 2013. http://eventsarchive.org/?q=wadl
4. Edward A. Fox and the CTRnet team. Crisis, Tragedy, and Recovery Network homepage. Virginia Tech, 2009-2013. http://www.ctrnet.net
5. Edward A. Fox and the IDEAL team. Events Archive homepage. Virginia Tech, 2013. http://www.eventsarchive.org
6. Edward A. Fox and the QDL team. Qatar Digital Library: Helping Build the Future of Digital Libraries in Qatar homepage. Qatar University, 2011-2013. http://qdl.qu.edu.qa/
7. Muriel Mewissen. Hiberlink: Time Travel for the Scholarly Web. EDINA, United Kingdon, 2013. Webpage at http://edina.ac.uk/projects/hiberlink_summary.html
8. Andreas Paepcke and Sanjay Kairam. EchoTree: Engaged Conversation when Capabilities are Limited. 2012. Stanford InfoLab Publication Server. Technical report and homepage at http://ilpubs.stanford.edu:8090/1054/
9. SiteStory team (LANL, ODU). 2013. SiteStory Web Archive homepage at http://mementoweb.github.io/SiteStory/
10. Siddhi Soman, Arti Chharjta, Alexander Bonomo, and Andreas Paepcke. ArcSpread for Analyzing Web Archives. 2012. Stanford InfoLab Publication Server. Technical report and homepage at http://ilpubs.stanford.edu:8090/1038/