Web Archiving and Digital Libraries (WADL 2013)

A Workshop at JCDL 2013
July 25-26, 2013
Indianapolis, Indiana, USA

Final Program

Edited by E. A. Fox
Web Archiving and Digital Libraries (WADL 2013)

Description:
- This workshop will explore 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), archiving, (text) analysis, up through long-term preservation.
- It will include particular coverage of current topics of interest, including:
  | archiving related to disasters | working with big data |
  | mobile web archiving | Memento and SiteStory |

Objectives:
- to continue to build the community of people integrating web archiving with DLs
- to help attendees learn about useful methods, systems, and software in this area
- to help chart future research and practice in this area, so more and higher quality web archiving occurs
- to promote synergistic efforts including collaborative projects and proposals

Duration: 1 day, split into two half-day segments, across Thu. July 25 – Fri. July 26 (Part 1: Thu. 1pm; Part 2: Fri. 8am)

Format: presentations, panel and open discussions, posters

Organizing Committee:
- Paul Bogen, Oak Ridge National Laboratory, bogenpl@ornl.gov
- Tessa Fallon, Columbia University, tessa.fallon@gmail.com
- Kristine Hanna, Internet Archive, kristine@archive.org
- Eric Hetzner, California Digital Library, erik.hetzner@ucop.edu
- Gina Jones, Library of Congress, gjion@loc.gov
- Martin Klein, LANL, martinklein0815@gmail.com
- Frank McCown, Harding University, fmccown@harding.edu
- Michael Nelson, Old Dominion Univ., mln@cs.odu.edu
- Andreas Paepcke, Stanford University, paepcke@cs.stanford.edu

Organizer: (to whom all inquiries should be sent)
- Edward A. Fox, Virginia Tech, fox@vt.edu, http://fox.cs.vt.edu, 540-231-5113
- Fox is a Professor in CS, and Director of the Digital Library Research Laboratory at Virginia Tech, where he is launching campus efforts for web archiving. He helped start the digital library initiative in 1991, and has been actively involved in the field ever since, including helping organize or attending many related conferences and workshops. He chairs the JCDL Steering Committee. He directs the Networked Digital Library of Theses and Dissertations, the CTRnet (Crisis, Tragedy, and Recovery network) project that is archiving about disaster events, and the Qatar Digital Library that will archive Qatari content.

Closely related event and results:
- Web Archive Cooperative – see http://infolab.stanford.edu/wac/
<table>
<thead>
<tr>
<th>LName</th>
<th>Rest of Name</th>
<th>Email</th>
<th>Affiliation</th>
<th>Min.</th>
<th>#</th>
<th>Title(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paepcke</td>
<td>Andreas</td>
<td><a href="mailto:paepecke@cs.stanford.edu">paepecke@cs.stanford.edu</a></td>
<td>Stanford University</td>
<td>60</td>
<td>1</td>
<td>ArcSpread: Enabling Web Archive Analysis for non-CS Experts + Applying web archives to real-time group source prediction of speech</td>
</tr>
<tr>
<td>Gray</td>
<td>Gabriella</td>
<td><a href="mailto:gsgray@library.ucla.edu">gsgray@library.ucla.edu</a></td>
<td>UCLA Library</td>
<td>5</td>
<td>2</td>
<td>Personal intro</td>
</tr>
<tr>
<td>Braaten</td>
<td>Ylva</td>
<td><a href="mailto:braaten@un.org">braaten@un.org</a></td>
<td>United Nations</td>
<td>10</td>
<td>3</td>
<td>United Nations digital repository for UN documentation</td>
</tr>
<tr>
<td>Klein</td>
<td>Martin</td>
<td><a href="mailto:martinklein0815@gmail.com">martinklein0815@gmail.com</a></td>
<td>Los Alamos National Laboratory</td>
<td>40</td>
<td>4</td>
<td>SiteStory, Archiving Done Differently (30) + Hiberlink, Towards Time Travel for the Scholarly Web (10)</td>
</tr>
<tr>
<td>Nelson</td>
<td>Michael</td>
<td><a href="mailto:mln@cs.odu.edu">mln@cs.odu.edu</a></td>
<td>Old Dominion University</td>
<td>5</td>
<td>5</td>
<td>Personal intro</td>
</tr>
<tr>
<td>SalahEldeen</td>
<td>Hany M.</td>
<td><a href="mailto:hany@cs.odu.edu">hany@cs.odu.edu</a></td>
<td>Old Dominion University</td>
<td>30</td>
<td>6</td>
<td>Temporal User Intention Modeling in Social Media</td>
</tr>
<tr>
<td>Szajewski</td>
<td>Michael</td>
<td><a href="mailto:mgszajewski@bsu.edu">mgszajewski@bsu.edu</a></td>
<td>Ball State University</td>
<td>15</td>
<td>7</td>
<td>Needs and Obstacles for a Web Archiving Initiative at the Ball State University Libraries</td>
</tr>
<tr>
<td>AlNoamany</td>
<td>Yasmin</td>
<td><a href="mailto:yasmin@cs.odu.edu">yasmin@cs.odu.edu</a></td>
<td>Old Dominion University</td>
<td>10</td>
<td>8</td>
<td>Who and What Links to the Internet Archive</td>
</tr>
<tr>
<td>AlSum</td>
<td>Ahmed</td>
<td><a href="mailto:aalsum@cs.odu.edu">aalsum@cs.odu.edu</a></td>
<td>Old Dominion University</td>
<td>20</td>
<td>9</td>
<td>Web Archiving Profile Overview</td>
</tr>
</tbody>
</table>

**Day 1 closing discussion**

| Enlow    | Keith        | kenlow@harding.edu           | Harding University                | 5    | 10 (with McCown) | Archiving the Mobile Web                                                   |
| Yarbrough | Monica      | mwallers@harding.edu         | Harding University                | 5    | 11 (with McCown) | Archiving the Mobile Web                                                   |
| McCown   | Frank        | fmccown@harding.edu          | Harding University                | 50   | 12              | Temporal Spread in Archived Composite Resources                            |
| Ainsworth | Scott       | sainswor@cs.odu.edu          | Old Dominion University           | 20   | 13              | The role of BNDES in Brazil preservation of heritage                       |
| Balbi    | Fernanda     | fmbalbi@gmail.com            | Brazilian Development Bank-BNDES  | 20   | 14              | Measuring Archivability of Web Resources and the Damage when Mementos are Missing |
| Brunelle | Justin       | jbrunelle@cs.odu.edu         | Old Dominion University           | 30   | 15              | CTRnet + Web Archiving (Qatar, VT)                                        |
| Fox      | Edward       | fox@vt.edu                  | Virginia Tech                     | 30   | 16              | CTRnet + Web Archiving (Qatar, VT)                                        |

**Closing discussion**
<table>
<thead>
<tr>
<th>LName</th>
<th>Rest of Name</th>
<th>Email</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ainsworth</td>
<td>Scott</td>
<td><a href="mailto:sainswor@cs.odu.edu">sainswor@cs.odu.edu</a></td>
<td>Old Dominion University</td>
</tr>
<tr>
<td>AlNoamany</td>
<td>Yasmin</td>
<td><a href="mailto:yasmin@cs.odu.edu">yasmin@cs.odu.edu</a></td>
<td>Old Dominion University</td>
</tr>
<tr>
<td>AlSum</td>
<td>Ahmed</td>
<td><a href="mailto:aalsum@cs.odu.edu">aalsum@cs.odu.edu</a></td>
<td>Old Dominion University</td>
</tr>
<tr>
<td>Balbi</td>
<td>Fernanda</td>
<td><a href="mailto:fmbalbi@gmail.com">fmbalbi@gmail.com</a></td>
<td>Brazilian Development Bank-BNDES</td>
</tr>
<tr>
<td>Braaten</td>
<td>Ylva</td>
<td><a href="mailto:braaten@un.org">braaten@un.org</a></td>
<td>United Nations</td>
</tr>
<tr>
<td>Brunelle</td>
<td>Justin</td>
<td><a href="mailto:jbrunelle@cs.odu.edu">jbrunelle@cs.odu.edu</a></td>
<td>Old Dominion University</td>
</tr>
<tr>
<td>Enlow</td>
<td>Keith</td>
<td><a href="mailto:kenlow@harding.edu">kenlow@harding.edu</a></td>
<td>Harding University</td>
</tr>
<tr>
<td>Fox</td>
<td>Edward</td>
<td><a href="mailto:fox@vt.edu">fox@vt.edu</a></td>
<td>Virginia Tech</td>
</tr>
<tr>
<td>Gray</td>
<td>Gabriella</td>
<td><a href="mailto:gsgray@library.ucla.edu">gsgray@library.ucla.edu</a></td>
<td>UCLA Library</td>
</tr>
<tr>
<td>Klein</td>
<td>Martin</td>
<td><a href="mailto:martinklein0815@gmail.com">martinklein0815@gmail.com</a></td>
<td>Los Alamos National Laboratory</td>
</tr>
<tr>
<td>McCown</td>
<td>Frank</td>
<td><a href="mailto:fmccown@harding.edu">fmccown@harding.edu</a></td>
<td>Harding University</td>
</tr>
<tr>
<td>Nelson</td>
<td>Michael</td>
<td><a href="mailto:mln@cs.odu.edu">mln@cs.odu.edu</a></td>
<td>Old Dominion University</td>
</tr>
<tr>
<td>Paepcke</td>
<td>Andreas</td>
<td><a href="mailto:paepcke@cs.stanford.edu">paepcke@cs.stanford.edu</a></td>
<td>Stanford University</td>
</tr>
<tr>
<td>SalahEldeen</td>
<td>Hany M.</td>
<td><a href="mailto:hany@cs.odu.edu">hany@cs.odu.edu</a></td>
<td>Old Dominion University</td>
</tr>
<tr>
<td>Szajewski</td>
<td>Michael</td>
<td><a href="mailto:mgszajewski@bsu.edu">mgszajewski@bsu.edu</a></td>
<td>Ball State University</td>
</tr>
<tr>
<td>Yarbrough</td>
<td>Monica</td>
<td><a href="mailto:mwallers@harding.edu">mwallers@harding.edu</a></td>
<td>Harding University</td>
</tr>
</tbody>
</table>
**WADL 2013 Participant Page for**

**Name:** Scott G. Ainsworth

**Affiliation:** Old Dominion University, Norfolk, VA

**Contact info:**
- email: sainswor@cs.odu.edu
- twitter: @Galsondor

**Participation in earlier web archiving related events:** WADL 2011

**Goals for this workshop:**
- Insight into human and robot web archive use patterns.
- Feedback on the Memento temporal spread problem I will present.

**Title:** Temporal Spread in Archived Composite Resources

**Co-authors:** Michael L. Nelson

**Short abstract:**
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. We examine the temporal spread of composite archived resources (root plus embedded resources. We 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.

**Participation:**
- x Medium talk (20 minutes)
WADL 2013 Participant Page for

Name: Yasmin AlNoamany

Affiliation: Old Dominion University

Contact info: Yasmin@cs.odu.edu

Participation in earlier web archiving related events: none

Goals for this workshop:
My goal is to know more about web archiving and the research trends in this area.

Title: Who and What Links to the Internet Archive
Co-authors: Ahmed AlSum, Michele C. Weigle, Michael L. Nelson
Short abstract:
I'm presenting the results of research on Internet Archive's 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.

Preference for this (check all that you think might be appropriate):
☑️ Short talk (10 minutes)
WADL 2013 Participant Page for

Name: Ahmed AlSum

Affiliation: Old Dominion University

Contact info: aalsum@cs.odu.edu

Participation in earlier web archiving related events:
- Web Archive Cooperative 2011
- Web Archive Cooperative 2012
- IIPC GA 2012
- Temporal Web 2013
- IIPC GA 2013

Goals for this workshop:
My goal is to know more about the web archiving and discuss the current research trends and challenges with other colleagues.

Title: Profiling Web Archive Coverage
Co-authors: Michele C. Weigle, Michael L. Nelson, and Herbert Van de Sompel

Short abstract:
In this presentation, I will present my research results of profiling the existing web archives for Top-Level domains and content languages. We used these results to build profile for each web archive and used this profiles to optimize the query routing the Memento aggregator.

Participation:
x Medium talk (20 minutes)
Name: Balbi, Fernanda Menezes

Affiliation: Brazilian Development Bank

Contact info: Avenida Republica do Paraguai, 100 6o andar Centro
Rio de Janeiro - RJ
CEP 20031-917
email: fmb@bndes.gov.br

Participation in earlier web archiving related events:
Brazil - National Symposium of Public Policies of Digitalization

Goals for this workshop:
- Identification of risks (capital, human and technological) that involve the process of creation of digital libraries from historic libraries (in paper).
- Identification of opportunities for creation of innovative content based on digitized material (by using portable devices).
- Sharing success stories.

Title: The role of BNDES in Brazil preservation of heritage

Short abstract:
Brief explanation on the history of BNDES and its role in Brazilian Development. The BNDES performance in projects of preservation of the architectural heritage, preservation of collections and strengthening of relevant cultural institutions, including digital library projects.

Participation:
x Medium talk (20 minutes)
WADL 2013 Participant Page for

**Name:** Ylva Braaten

**Affiliation:** United Nations

**Contact info:** braaten@un.org  
UN Library on Twitter: @UNLibrary

**Participation in earlier web archiving related events:**
I haven’t participated in external events in the last several years but have been partly involved in the process of establishing a digital repository for UN Documentation.

**Goals for this workshop:** We are interested in hearing how other organizations/institutions do this and get some inspiration and ideas. Some decisions on procedures, scope, metadata, etc. have already been made but as the system is still in a testing phase we are interested in learning more.

**Title:** United Nations digital repository for UN documentation

**Short abstract:**
The UN Library in New York is in the process of changing its indexing process and moving to a digital repository for UN documentation (parliamentary documents, conference related documents, publications, etc). I present a status report on UN movement towards more digitalized and automated processes, involving different departments, agencies and duty stations.

**Participation:**
X Short talk (10 minutes)
Name: Justin F. Brunelle

Affiliation: Old Dominion University

Contact info: jbrunelle@cs.odu.edu

Participation in earlier web archiving related events: JCDL2013, JCDL2012, PDA2012, IIPC, NDIIP

Goals for this workshop:
I am interested in learning about the current research efforts of established researchers and receive feedback on my work.

Title: Measuring: Archivability of Web Resources and the Damage when Mementos are Missing

Short abstract: I will discuss my current research efforts in measuring the archivability of Web resources and measuring damage that occurs in mementos. The archivability portion of the talk will discuss what makes pages more or less archivable and how archivability is changing over time. The damage portion of the talk will discuss how we can measure memento importance and how we can measure the impact a missing memento has on a larger resource.

Participation:
_X_ Full talk (30 minutes)

Other comments or suggestions: I’m also participating as a co-author for Martin Klein’s talk on the LANL SiteStory Transactional Web Archive.
Name: Keith Enlow

Affiliation: Harding University

Contact info: kenlow@harding.edu

Participation in earlier web archiving related events: none

Goals for this workshop:
  • To share about work at Harding U.
  • To learn about work at other locations

Co-author: on paper presented with Frank McCown and Monica Yarbrough

Participation (in addition to above):
X Personal introduction (5 minutes)
Name: Fox, Edward

Affiliation: Virginia Tech

Contact info: fox@vt.edu, http://fox.cs.vt.edu, office 540-231-5113, mobile 540-553-1856

Participation in earlier web archiving related events:
- Attended 2011 Web Archiving workshop at JCDL in Ottawa
- Local meetings exploring web archiving for Virginia Tech IT and Library,
- Meetings about NSF grant, with Internet Archive, on the Crisis, Tragedy, and Recovery Network (CTRnet): http://www.ctrnet.net,
- Meetings about ETANA-DL (digital library aggregating from archaeology projects only otherwise visible on WWW)

Goals for this workshop:
1) That it will be a successful event with good community building, sharing, and new collaborations resulting. 2) Getting advice re CTRnet and campus archiving

Title: Crisis, Tragedy, and Recovery Network (CTRnet) Digital Library Project

Co-authors: Seungwon Yang, Kiran Chitturi, Andrea Kavanaugh, Donald Shoemaker, Steven Sheetz

Short abstract: The CTRnet project 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, analysis of webpage collections with big data software, filtering using machine learning, and topic tagging. We aim to broaden the efforts to encompass other types of events, e.g., political or community focused.

Participation:
x Full talk (30 minutes)
x Chairing
Name: Gabriella Gray

Affiliation: UCLA Library

Contact info: gsgray@library.ucla.edu

Participation in earlier web archiving related events:
I have attended many web archiving workshops beginning with the International Conference on Archiving Web Resources hosted by the National Library of Australia in 2004. Most recently I attended the 2012 open conference day of the International Internet Preservation Consortium (IIPC) General Assembly at the Library of Congress.
   I presented at 7th International Web Archiving Workshop <http://iwaw.europarchive.org/07/>, Vancouver, Canada (in conjunction with JCDL 2007).

Goals for this workshop:
I am hoping to learn more about the inner technical workings of web archives and crawlers, including indexing, exploration (including searching and browsing), archiving, and (text) analysis. I would also like to learn more about Memento and SiteStory.

Participation:
X Very short talk (5 minutes) Participant introduction.
WADL 2013 Participant Page for

Name: Martin Klein

Affiliation: Los Alamos National Laboratory

Contact info: martinklein0815@gmail.com

Participation in earlier web archiving related events: none

Goals for this workshop: Share and learn

Title 1: Hiberlink, Towards Time Travel for the Scholarly Web
Co-authors: Herbert Van de Sompel, Robert Sanderson, Peter Burnhill (EDINA), Claire Grover (University of Edinburgh)
Short abstract: I will give a brief overview of the recently launched Hiberlink project. The project aims at quantifying the citation rot problem in scholarly articles at unprecedented scale but also at proposing solutions for researchers and publishers to ensure the longevity of the content of research.

Title 2: SiteStory, Archiving Done Differently
Co-author: Justin Brunelle (ODU)
Short abstract: We will provide an overview of the SiteStory concept and its functionality compared to crawler-based archiving. We will discuss the SiteStory approach and describe various use case. We will further introduce the SiteStory testbed and provide insight into novel results of benchmarking experiments.

Participation:
- Short talk (10 minutes)
- Full talk (30 minutes)
Name: Frank McCown

Affiliation: Harding University

Contact info: fmccown@harding.edu

Participation in earlier web archiving related events: several

Goals for this workshop: Share our work with others who might be interested and learn about other archiving work others are working on.

Title: Archiving the Mobile Web

Co-authors: Monica Yarbrough and Keith Enlow

Short abstract:
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. In this talk we will share our work on developing tools for web archivists to automate the discovery and archiving of mobile websites.

Participation:
_x_ Long talk (50 minutes)
WADL 2013 Participant Page for

Name: Michael Nelson

Affiliation: Old Dominion University

Contact info: mln@cs.odu.edu

Participation in earlier web archiving related events: several

Goals for this workshop: presenting work at ODU, commenting

Participation:
  x Very short talk (5 minutes) Personal introduction, helping students
WADL 2013 Participant Page for

Name: Paepcke, Andreas

Affiliation: Stanford University, Department of Computer Science

Contact info: paepcke@cs.stanford.edu

Participation in earlier web archiving related events:
Organized web archiving workshop at Stanford University

Goals for this workshop:
What others are doing.

Title 1: ArcSpread: Enabling Web Archive Analysis for non-CS Experts
Short abstract: I will present a vision for how sociologists, political scientists, and historians might analyze Web archives in the future. The project, ArcSpread, designs and implements a spreadsheet-based approach to the problem. I will work 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.

Title 2: Applying web archives to real-time group source prediction of speech
Short abstract: One problem encountered by a mute, quadriplegic friend is that his conversation partners get bored waiting for him to finish typing on an onscreen keyboard via his head tracking device. In an effort to engage these conversation partners, we generate word trees that attempt to predict what my friend is going to say, given his previously typed word. For the prediction statistics, we used three different underlying collections: Henry's Web blog, a specialized crawl of 10M Web pages, and a collection of 11K 10-minute phone conversations. W compared the outcomes -- so far purely in their effectiveness to produce word trees.

Participation:
2x Full talks (each 30 minutes, given together)

Other comments or suggestions:
Web connectivity would be great for one demo. But I can work without it.
Name: Hany M. SalahEldeen

Affiliation: Old Dominion University

Contact info (all you are willing to share): hany@cs.odu.edu, 757-648-9944


Goals for this workshop: To present my research about user intention in social media, what I have already done, the plans for the future, and have related discussions and ideas exchange.

Title: Temporal User Intention Modeling in Social Media

Co-author: Dr. Michael Nelson

Short abstract: 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 and 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 examine the resource might not get what the author intended. This change of intention and the resource’s state could cause a significant inconsistency in the published content. Our goal is to model the author’s intention across time, and predict it to avoid the temporal inconsistency that might occur.

Participation: _X_ Full talk (30 minutes)
Name: Michael Szajewski

Affiliation: Archivist for Digital Development and University Records, Ball State University Libraries

Contact info: mgszajewski@bsu.edu

Participation in earlier web archiving related events: Mostly archives-specific conferences (Midwest Archives Conference, ContentDM Midwest Users Group)

Goals for this workshop: To obtain an understanding of current trends and developments in web archiving that could be applied in an academic library setting. To develop the skills necessary to evaluate the feasibility of web archiving solutions and develop a practical, workable web archiving initiative for my library. To understand resources necessary to implement a web archiving initiative.

Title: Needs and Obstacles for a Web Archiving Initiative at the Ball State University Libraries

Short abstract: This talk will briefly overview the digital library program at Ball State University, including a ContentDM and DSpace repository, and will focus on the opportunities for the Ball State University Libraries to developing an initiative to capture, preserve, and provide access to archived dynamic web content. The talk will discuss 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 level will be discussed.

Participation: X Short talk (15 minutes)
Name: Monica Yarbrough

Affiliation: Harding University

Contact info: mwallers@harding.edu

Participation in earlier web archiving related events: none

Goals for this workshop:
- To share about work at Harding U.
- To learn about work at other locations

Co-author: on paper presented with Frank McCown and Keith Enlow

Participation (in addition to above):
X Personal introduction (5 minutes)