Web Archiving and Digital Libraries (WADL) 2016: Highlights and Introduction to this Special Issue

Edward A. Fox
Virginia Tech
Dept. of Computer Science
Blacksburg, VA 24061 USA
+1-540-231-5113
fox@vt.edu

Zhiwu Xie
Virginia Tech
University Libraries
Blacksburg, VA 24061 USA
+1-540-231-4453
zhiwuxie@vt.edu

Martin Klein
Research Library
Los Alamos National Laboratory
Los Alamos, NM 87545 USA
+1-505-667-5809
mklein@lanl.gov

ABSTRACT
This workshop, reported in the following 12 papers, explored the integration of Web archiving and digital libraries, so the complete life cycle involved was introduced: creation/authoring, uploading/publishing in the Web (2.0), (focused) crawling, indexing, exploration (searching, browsing), archiving (of events), etc. It included particular coverage of current topics of interest, e.g., big data, mobile web archiving, and systems (e.g., Memento, SiteStory, Hadoop processing).

Categories and Subject Descriptors
H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services.* H.3.6 [**Information Storage and Retrieval**]: Library Automation – *Large text archives*. H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – Collection, Standards, Systems issues.

General Terms
Management, Standardization.

Keywords
Web archiving; Internet Archive.

1. INTRODUCTION
Our understanding of the past will, to a large extent, depend on our success with Web archiving. WADL 2016 brought together international leaders from industry, government, and academia, who are tackling this important challenge. They explored the integration of Web archiving and digital libraries, over the complete life cycle: creation/authoring, uploading, publishing in the Web, crawling/collecting, compressing, formatting, storing, preserving, analyzing, indexing, supporting access, etc.

The objectives of this workshop were to:

- continue to build the 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, so more and higher quality Web archiving occurs;
- produce an archival publication (this issue) that will help advance technology and practice; and
- promote synergistic efforts including collaborative projects and proposals.

2. RELATED WORK
The most recent related workshop, WADL 2015, was held in conjunction with JCDL 2015. It led to a special issue of the IEEE TCDL Bulletin with 13 papers [1]. An earlier workshop, WIRE, focused on research leading to or making use of archives that preserve Internet content [2]. The first workshop on Web Archiving and Digital Libraries, WADL 2013, led to a summary [3] after a group responded to the call for meeting [4] as part of the JCDL 2013 workshop program. An earlier similar workshop at a prior JCDL took place in Ottawa in 2011 [5], partly as a result of the emergence of a cooperative to explore Web archiving [6]. Broader in scope but related are the annual General Assembly meetings of the International Internet Preservation Consortium (IIPC) [7]. Thanks in part to a funding initiative from the Mellon Foundation, managed through Columbia University [8], additional research has proceeded that has contributed to this issue.

3. TOPICS
This workshop covered many topics of interest, including but not limited to:

Archiving (events)	Big data	Classification
Community building	Crawling (focused)	Curation, Q/C
Databases / collections	Discovery	Extraction/analysis
Filling gaps	Globalization	Linking archives
Metadata	Mobile devices	Network science
Preservation	Resource description	Social sciences
Standards, protocols	Systems, tools	Tweet connections

4. LOGISTICS
The workshop proceeded after JCDL 2016, at Rutgers University in Newark, NJ, USA. It was held in 2 parts, 2-5pm on 22 June and 9am-noon on 23 June. For details about the workshop and its schedule, see its website [9]. After the opening welcome and introductions, the first panel gave an overview of worldwide activities on Web archiving, while the second day was opened by a panel on social media, Web archiving and digital libraries [10]. After the opening panel there was one paper [11], next a series of five lightening talks [17-21], and then, after a poster/demo session, two more papers [12, 13]. On the second day, after the opening panel [10], there were three more papers [14-16], and a plenary closing discussion.

5. WORKSHOP PRESENTATIONS
As was explained in Section 4, the workshop had two panels [10], six paper presentations [11-16], and five lightening talks [17-21]. In the remainder of this special issue are the resulting twelve papers. Together they provide a comprehensive overview of key issues in the overlap between Web archiving and digital libraries.

In the opening panel [10], Vinay Goel of the Internet Archive gave a helpful introduction and broad overview, including of Archive-It and ArchiveSpark, thus covering community and

technology issues. Ian Milligan, from Waterloo, Canada, discussed international collaborations, including regarding archive hackathons, building consensus on relevant tools and standards. The second panel, moderated by Daniel Kerchner of George Washington University Libraries focused on efforts to ensure inclusion and integration of social media in the work on archiving and digital libraries. Laura Wrubel, also from GWU Libraries, described Social Feed Manager. Ian Millian joined too, discussing work with tweets from the Canadian Federal elections. In addition, Mark Phillips, of University of North Texas (UNT), focused on ways to collaborate on research with tweet collections.

The first paper presentation [11], with co-authors from Ghent, Belgium, as well as Los Alamos National Laboratory, was about linked data from archives, and how to support access and other functions, with billions of RDF triples, using Linked Data Fragments, Triple Pattern Fragments, and Header Dictionary Triples. The second paper presentation [12], with co-authors from Texas A&M University, describes a user study of perceptions when change occurs in a distributed collection. They reported on how well humans notice and identify problems when, for example, a conference website no longer connects properly with the content of an old event. The third paper [13], from the libraries at UNT and Stanford, as well as a data lab, explains how to compare archives, spotting changes and trends, using webpages, CDX files, and URL seed lists. It focuses on the 2008 and 2012 End of Term US Presidential Archives. The fourth paper [14], from three departments at Virginia Tech, describes a modified approach to real-time transactional Web archiving, leveraging caching methods. The fifth paper [15], from UNT, considers the information quality (IQ) of Web archives. It reviews seven IQ dimensions, and argues that three (accuracy/completeness, usefulness, and coherence) are applicable to Web archives. The final paper presented [16], based on focused crawling research in the IDEAL project at Virginia Tech (VT), discusses how social media can guide the process, according to the importance of different sources, e.g., regarding the March 2016 Brussels attacks.

The first lightning talk [17], from VT, describes Java software, available as open source, to manage Web archive files in the Cloud. The second talk [18] describes big data hardware and software in the IDEAL project at VT, used to manage webpages as well as over a billion tweets, along with visualization, e.g., of tweets about water main breaks. The third talk [19], from Old Dominion University (ODU), describes the InterPlanetary effort, including filing, indexing, and replaying of archival records. The fourth talk [20], from ODU, describes MemGator, a new publicly available standalone aggregator to work with multiple archives and the Memento protocol. The final lightning talk [21], from Konstanz Germany and Dublin Ireland, describes a method to timestamp archived content, such as about cultural heritage, so there is a trusted proof of when archived Web content was available.

Though each paper is relatively brief, they capture key issues and aspects of the interconnection between Web archiving and digital libraries. We hope interested readers will become involved in future WADL events, such as in connection with JCDL 2017, as well as investigate subsequent more in-depth descriptions of the research highlighted in this issue of the IEEE TCDL Bulletin.

6. AKNOWLEDGMENTS

Our thanks go to NSF for support through IIS 1319578 and 1619028, to IMLS for LG-71-16-0037-16, to Columbia and the Mellon Foundation for supporting "Archiving Transactions Toward Uninterruptible Web Service," and to QNRF for support through NPRP 4-029-1-007. The opinions expressed in this document are solely our own.

7. REFERENCES

[1] E. A. Fox, Z. Xie, and M. Klein. (3/17/2017). *Introduction to the Web Archiving and Digital Libraries 2015 Workshop Issue - Bulletin of IEEE Technical Committee on Digital Libraries.* Available: http://www.ieee-tcdl.org/Bulletin/v11n2/papers/intro.pdf

[2] M. Weber, D. Lazer, K. Carpenter-Negulescu, and A. Kosterich. (3/17/2017). *Working with Internet Archives for Research - WIRE 2014 Workshop, Cambridge, MA, June 17-18, 2014.* Available: http://wp.comminfo.rutgers.edu/nsfia/

[3] E.A. Fox and M.M. Farag. (3/17/2017). *Report on the Workshop on Web Archiving and Digital Libraries - WADL 2013*. Available: http://sigir.org/files/forum/2013D/p128.pdf

[4] E.A. Fox. (3/17/2017). *Web Archiving and Digital Libraries - WADL 2013. Virginia Tech CTRnet announcement.* Available: http://www.ctrnet.net/sites/default/files/JCDL2013WorkshopWebArchiving20130603.pdf

[5] H. Garcia-Molina, F. McCown, M. L. Nelson, and A. Paepcke. (3/17/2017). *Web Archive Globalization Workshop. In conjunction with JCDL 2011, Ottawa, Canada, June 16-17.* Available: http://cs.harding.edu/wag2011/

[6] H. Garcia-Molina, F. McCown, M. L. Nelson, and A. Paepcke. (3/17/2017). *Web Archive Cooperative Making Web Archives Useful Today.* Available: http://infolab.stanford.edu/wac/

[7] IIPC. (3/17/2017). *International Internet Preservation Consortium.* Available: http://netpreserve.org/

[8] Columbia University Libraries. (3/17/2017). *Web Resources Collection Program.* Available: https://library.columbia.edu/bts/web_resources_collection.html

[9] E.A. Fox. (3/17/2017). *Website for WADL 2016. Virginia Tech, Dept. of Computer Science, Blacksburg, VA 24061.* Available: http://fox.cs.vt.edu/wadl2016.html

[10] V. Goel and D. Kerchner. (3/17/2017). *WADL 2016 Panels: Worldwide activities on Web archiving; Social media, Web archiving, and digital libraries.* Available: http://www.ieee-tcdl.org/Bulletin/v13n1/

[11] R. Verborgh, M. Vander Sande, H. Shankar, L. Balakireva, and H. Van de Sompel. (3/17/2017). *Devising Affordable and Functional Linked Data Archives.* Available: http://www.ieee-tcdl.org/Bulletin/v13n1/

[12] L. Meneses, R. Furuta, and F. Shipman. (3/17/2017). *Evaluating Unexpected Change in a Distributed Collection.* Available: http://www.ieee-tcdl.org/Bulletin/v13n1/

[13] M. Phillips, D. Chudnov, and J. Jacobs. (3/17/2017). *Exploratory Analysis of the End of Term Web Archive: Comparing two collections.* Available: http://www.ieee-tcdl.org/Bulletin/v13n1/

[14] Z. Xie, K. Nayyar, and E. A. Fox. (3/17/2017). *Nearline Web Archiving.* Available: http://www.ieee-tcdl.org/Bulletin/v13n1/

[15] B. Reyes Ayala. (3/17/2017). *We need new names: Applying existing models of Information Quality to web archives.* Available: http://www.ieee-tcdl.org/Bulletin/v13n1/

[16] M. Farag and E. A. Fox. (3/17/2017). *Which webpage should we crawl first? Social media-based webpage source importance guidance.* Available: http://www.ieee-tcdl.org/Bulletin/v13n1/

[17] Y. Chen, Z. Xie, and E. A. Fox. (3/17/2017). *A Library to Manage Web Archive Files in Cloud Storage.* Available: http://www.ieee-tcdl.org/Bulletin/v13n1/

[18] S. Lee and E. A. Fox. (3/17/2017). Archiving and Analyzing Tweets and Webpages with the DLRL Hadoop Cluster. Available: http://www.ieee-tcdl.org/Bulletin/v13n1/

[19] S. Alam, M. Kelly, M. C. Weigle, and M. L. Nelson. (3/17/2017). *InterPlanetary Wayback: The Permanent Web Archive.* Available: http://www.ieee-tcdl.org/Bulletin/v13n1/

[20] S. Alam and M. L. Nelson. (3/17/2017). *MemGator – A Portable Concurrent Memento Aggregator.* Available: http://www.ieee-tcdl.org/Bulletin/v13n1/

[21] B. Gipp, N. Meuschke, J. Beel, and C. Breitinger. (3/17/2017). *Using the Blockchain of Cryptocurrencies for Timestamping Digital Cultural Heritage.* Available: http://www.ieee-tcdl.org/Bulletin/v13n1/