Web Archiving and Digital Libraries (WADL)

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
University of California Los Angeles
Research Library
Los Angeles, CA 90095 USA
+1-310-206-9781
martinklein@library.ucla.edu

ABSTRACT
This workshop will explore integration of Web archiving and digital libraries, so the complete life cycle involved is covered: creation/authoring, uploading/publishing in the Web (2.0), (focused) crawling, indexing, exploration (searching, browsing), archiving (of events), etc. It will include particular coverage of current topics of interest, like: 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 will bring together international leaders from industry, government, and academia, who are tackling this important challenge. They will explore 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 are 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 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].

3. TOPICS
This workshop will cover all 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
4.1 Audience and Attendees
There is a growing community interested in this topic. Given that the prior WIRE workshop on a similar theme [2] had a significant number of participants from Rutgers University and that Columbia University played a lead role in a closely related Mellon funded initiative [8], we expect to have 15-30 attendees, including a solid representation of students. We will advertise, solicit submissions, have them reviewed by the program committee, and then organize an interesting program. We anticipate it to include aspects from multiple disciplines such as Computer Science, Library and Information Science, Web Science, Social Sciences, etc.

4.2 Format and Duration
A full-day workshop is needed, though it may be possible to split it across two days if necessary. There will be invited speakers, selected papers, posters, demonstrations, and panels.

4.3 Special Details and Requirements
We expect an international program committee of 6-12 people, in addition to the three co-chairs. We will run a WebEx teleconference during the workshop so that those unable to attend at the last moment will still be able to be involved. We plan to

have a small poster session in addition to the usual type of activities.

As in 2015, we expect to have a special issue of IEEE TCDL Bulletin; that led to a call for a special issue of IJDL, so if the timing is right, WADL 2016 contributions may go there too.

5. BIOGRAPHICAL AND CONTACT INFORMATION

Edward Fox holds a Ph.D. and M.S. in Computer Science from Cornell, and a B.S. from M.I.T. Since 1983 he has been at Virginia Tech, where he serves as Professor. He directs VT's Digital Library Research Laboratory and the Networked Digital Library of Theses and Dissertations. He was a member of the Board of CRA (the Computer Research Association). He was chair of the IEEE Technical Committee on Digital Libraries, and earlier was chair of ACM SIGIR. He was chair of the steering committee for JCDL, and is on the international advisory committee for ICADL. He has been (co-)Principal Investigator on 120 research grants/contracts. He has taught over 80 tutorials and has given 66 keynote/distinguished/ international invited talks. He has (co-)authored 18 books, 117 journal/magazine articles, 49 books chapters, 202 refereed conference/workshop papers, 69 posters, and over 150 other publications/reports, plus over 300 additional talks. Fox is editor for IR and DL for ACM Books. He was Co-Editor-in-Chief for ACM JERIC, and is on the boards of IJDL, JEMH, JIIS, J. UCS, Multimedia Tools & Applications, and PeerJ CS.

Communicating Contact information:

Dept. of Computer Science, 114 McBryde Hall, M/C 0106 Virginia Tech, Blacksburg, VA 24061, Tel: +1-540-231-5113 [direct], -6931[dept.]; cell: +1-540-553-1856, Fax: +1-540-231-6075 [CS]; Email: fox@vt.edu; Website: http://fox.cs.vt.edu

Zhiwu Xie is an associate professor at Virginia Tech Libraries and leads its technology development team. He leads the development of transactional archiving based UWS, the Goodwin Hall Living Lab data management system, IMLS ETDplus, and VTechData, among others. He is deeply involved in Fedora/Hydra, APTrust, PREMIS, ResourceSync, and AltMetrics Data Quality. His research has received supported from Mellon, IBM, Amazon, and NSF XSEDE. He was a co-chair of WADL 2015. His website is at http://scholar.lib.vt.edu/staff/zxie/.

Martin Klein holds a Ph.D. in Computer Science from Old Dominion University. He currently is a scientist in the Research Library at the University of California Los Angeles. He was program chair of DL 2014, poster chair of JCDL 2015, and is currently co-conference chair of iPres 2016. In addition, he is a board member of the Web Archiving Collaboration at Columbia University and guest editor of the International Journal on Digital Libraries as well as the Bulletin of IEEE Technical Committee on Digital Libraries. He also was co-chair of WADL 2015. Martin Klein is the lead editor of the ANSI/NISO Specification Z39.99 and has published numerous journal/magazine articles and conference/workshop papers. More information can be found at: http://www.cs.odu.edu/~mklein

6. ACKNOWLEDGMENTS

Our thanks go to NSF for support through IIS 1319578, 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] Edward A. Fox, Zhiwu Xie, Martin Klein. Introduction to the Web Archiving and Digital Libraries 2015 Workshop Issue: Web Archiving and Digital Libraries 2015 (WADL 2015) Overview. Bulletin of IEEE Technical Committee on Digital Libraries, 11(2), October 2015, 2 pages, http://www.ieee-tcdl.org/Bulletin/v11n2/papers/intro.pdf

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

[3] Fox, E.A. and Farag, M.M. 2013. *Report on the Workshop on Web Archiving and Digital Libraries* (WADL 2013), WADL Workshop Report, ACM SIGIR Forum, 47(2): 128-133. http://sigir.org/files/forum/2013D/p128.pdf

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

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

[6] Garcia-Molina, H., McCown, F., Nelson, M., and Paepcke, A. 2011. Web Archive Cooperative Making Web Archives Useful Today, Supported by the NSF (1009916), Stanford University, http://infolab.stanford.edu/wac/

[7] IIPC. International Internet Preservation Consortium. 2015. Homepage http://netpreserve.org/

[8] Web Resources Collection Program, Columbia U. Libraries. https://library.columbia.edu/bts/web_resources_collection.html