Web Archiving and Digital Libraries

Zhiwu Xie
zhiwuxie@vt.edu
Virginia Polytechnic Institute and State University
Blacksburg, VA

Martin Klein
mklein@lanl.gov
Los Alamos National Laboratory
Los Alamos, NM

Edward A. Fox
fox@vt.edu
Virginia Polytechnic Institute and State University
Blacksburg, VA

ABSTRACT

This workshop will explore integration of Web archiving and digital libraries and cover all stages of its complete life cycle, including creation/authoring, uploading/publishing, crawling, indexing, exploration, and archiving, etc. It will include particular coverage of current topics of interest, like: big data, social media archiving, and systems.

CCS CONCEPTS

• **Information systems → World Wide Web**; **Storage management**.

KEYWORDS

Web Archiving, Digital Preservation, Community Building

ACM Reference Format:
Zhiwu Xie, Martin Klein, and Edward A. Fox. 2020. Web Archiving and Digital Libraries. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), August 1–5, 2020, Virtual Event, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3383583.3398509

1 INTRODUCTION

WADL 2020 will continue the WADL tradition to provide a forum and collaboration platform for international leaders from academia, industry, and government to discuss the 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. The objectives of this workshop are to:

- 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,
- 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 2019 [4], was held in conjunction with JCDL 2019. We received very positive feedback from participants and a strong preference for the continuation of the workshop in 2020. Two previous WADL meetings resulted in the publication of a special issue in the IEEE TCDL Bulletin such as in 2016 [3] and 2015 [1]. Other workshop proceedings will be openly accessible from VTechWorks, Virginia Tech's institutional repository.

A previous workshop, WIRE [5], focused on research of archival holdings and on making use of archives that preserve the web. The first workshop on Web Archiving and Digital Libraries, WADL 2013, led to a summary [2] after a group responded to the call for meeting 2 as part of the JCDL 2013 workshop program.

An earlier similar workshop at a prior JCDL conference took place in Ottawa in 2011, partly as a result of the emergence of a cooperative to explore web archiving. Broader in scope but related are the annual General Assembly meetings of the International Internet Preservation Consortium (IIPC). In addition, various sponsored programs have connected, like a closely related initiative funded by the Institute of Museum and Library Services.

3 WORKSHOP TOPICS

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

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

4 WORKSHOP LOGISTICS

4.1 Audience and Attendees

Over the past few years, we have observed that the community interested in this topic is growing. Based on interest in recent years, we expect to have around 15-30 attendees, including a solid representation of students. We will advertise the workshop and actively

JCDL '20, August 1–5, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7585-6/20/06.
https://doi.org/10.1145/3383583.3398509

solicit submissions. All submissions will be peer-reviewed by the program committee and accepted contributions will be compiled into the WADL program. We anticipate the program to include aspects from multiple disciplines such as Computer Science, Library and Information Science, Web Science, Social Sciences, History, Journalism, etc.

4.2 Format and Duration

We propose a full-day workshop. We anticipate invited speakers, presentations of selected papers and posters, as well as demonstrations and panels.

4.3 Special Requirements

We have an international program committee of about 15 people, in addition to the three co-chairs. We will consider running 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 typical conference-style 15 and 30 minute presentations while leaving plenty of room for discussion. As done for previous WADL events, the organizers will pursue venues to offer an opportunity to publish invited contributions that were presented in their preliminary stages at the workshop. In the past, the workshop has also led to a call for contributions for a special issue of IJDL. The co-chairs intend to invest this effort again and edit an IJDL special issue on web archiving in the near future.

5 BIOGRAPHICAL INFORMATION

Zhiwu Xie is a professor and chief strategy officer at Virginia Tech Libraries. He leads the IMLS-funded "Continuing Education to Advance Web Archiving" project, which grew out of discussions from prior WADL workshops. His research interests focus on library cyberinfrastructure, data management, digital preservation, and web archiving, with application areas ranging from ecological systems to smart building. He was a co-chair of WADL from 2015-2019. More information can be found at: http://orcid.org/0000-0002-2702-3806.

Martin Klein holds a Ph.D. in Computer Science from Old Dominion University. He currently is a scientist in the Research Library at Los Alamos National Laboratory. He was program chair of DL 2014, poster chair of JCDL 2015, panel co-chair of JCDL 2016, and conference co-chair of iPres 2016. In addition, he also was co-chair of WADL 2015, 2016, 2017, and 2018 and the General Assembly of the International Internet Preservation Consortium (IIPC) in 2017, 2018, and 2019. He is an editorial board member of the International Journal on Digital Libraries (IJDL), guest editor for the Bulletin of IEEE Technical Committee on Digital Libraries, and served as a board member of the Web Archiving Collaboration at Columbia University. 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://orcid.org/0000-0003-0130-2097.

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) as well as a member of the ACM Publications Board and co-chair of its Digital Library Committee. 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 over 127 research grants/contracts. He has taught over 80 tutorials and has given 66 keynote/distinguished/international invited talks. He has (co-)authored 18 books, 120 journal/magazine articles, 49 books chapters, 211 refereed conference/workshop papers, 73 posters, and over 150 other publications/reports, plus over 300 additional talks. Fox was editor for IR and DL for ACM Books. He was Co-Editorin-Chief for ACM JERIC, and is on the boards of IJDL, JEMH, JIIS, J. UCS, Multimedia Tools & Applications, and PeerJ CS. He is a Fellow of ACM and of IEEE. His website can be found at: http://fox.cs.vt.edu.

ACKNOWLEDGMENTS

This work is partially supported by the Institute of Museum and Library Services under Grant No.: LG-71-16-0037-16 (https://www.imls.gov/grants/awarded/lg-71-16-0037-16) and National Science Foundation under grant IIS-1619028 (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1619028) and IIS-1619371 (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1619371).

REFERENCES

[1] Martin Klein Edward A. Fox, Zhiwu Xie. 2015. Introduction to the Web Archiving and Digital Libraries 2015 Workshop Issue: Web Archiving and Digital Libraries 2015 (WADL 2015) Overview. *TCDL Bulletin* 11, 2 (Oct. 2015).

[2] Edward A. Fox and Mohamed M. Farag. 2013. Report on the Workshop on Web Archiving and Digital Libraries (WADL 2013). *SIGIR Forum* 47, 2 (Jan. 2013), 128–133. https://doi.org/10.1145/2568388.2568408

[3] Edward A. Fox, Zhiwu Xie, and Martin Klein. 2017. Web Archiving and Digital Libraries (WADL) 2016: Highlights and Introduction to this Special Issue. *TCDL Bulletin* 13 (2017). https://doi.org/10.1109/jcdl.2017.7991625

[4] Martin Klein, Zhiwu Xie, and Edward A Fox. 2019. Workshop on Web Archiving and Digital Libraries (WADL). In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 455–456.

[5] David Lazer Matthew Weber and Kris Carpenter Negulescu. 2014. *WIRE 2014 Workshop - Working with Internet Archives for Research.* http://wp.comminfo.rutgers.edu/nsfia/