Int J Digit Libr (2018) 19:1–2
https://doi.org/10.1007/s00799-016-0203-5

Guest editors' introduction to the special issue on web archiving

Edward A. Fox[1] · **Martin Klein**[2] · **Zhiwu Xie**[3]

Since 1997, numerous organizations around the world have archived much of the content on the World Wide Web. This movement has spread, with more and more groups participating, so now there is a substantial research, development, operation, and utilization infrastructure supporting the collection, storage, indexing, sharing, and accessing of a large portion of the history of webpages. This infrastructure is shared by individuals, groups, educational institutions, government agencies, corporations, and other entities. Standards have emerged, tools have been devised, and analysis methods have been applied. All of this work helps show how, in the digital world, libraries and archives can be well supported, separately and in combination, by digital library methods.

This special issue includes six papers. The authors are from Centrum Wiskunde & Informatica, Amsterdam, Netherlands; George Washington University, Washington, D.C., USA; Leibniz University Hannover, Germany; Los Alamos National Laboratory, New Mexico, USA; The Open University of Israel, Raanana, Israel and University of Haifa, Haifa, Israel; and Virginia Tech, Virginia, USA. This international group includes researchers and developers and represents key parts of both the research and development communities involved in web archiving. Three of the six papers in this special issue concentrate on aids for collecting web content to be archived. The remaining three papers focus on collaboration and exploration of archives, temporal analysis, and use of archived collections.

The paper titled "Focused Crawling for Events" describes enhanced methods for gathering webpages about events in the world. While many methods and tools (e.g., Heritrix[1]) exist to collect webpages from previously identified websites, this work addresses the challenging problem of doing so about important world events, where the webpages desired are scattered around the web. Leveraging the common practice of Twitter users including short URLs in their tweets, and using timestamps and locations as well as keywords to help determine whether a webpage is relevant to an event, this paper shows improved results on multiple measures of the quality of the collection process.

The second paper on archival collection building is titled "API-based Social Media Collecting as a Form of Web Archiving". It elaborates on the redesign process of the popular open source tool "Social Feed Manager" to argue for a closer alignment of API-based social media collecting with conventional web archiving. Since the capture and archiving of social media content has increasingly gained traction in the web archiving community, assumptions about collection of such content as well as the goals of a collecting application such as Social Feed Manager have evolved. The paper describes the technical approach and its implementation that satisfies the new requirements. One example is the adoption of the standard archival file format WARC[2] for collected

✉ Edward A. Fox
fox@vt.edu
http://fox.cs.vt.edu

Martin Klein
mklein@lanl.gov

Zhiwu Xie
zhiwuxie@vt.edu

[1] Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

[2] Research Library, Los Alamos National Laboratory, Los Alamos, NM, USA

[3] University Libraries, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

[1] https://webarchive.jira.com/wiki/display/Heritrix.

[2] https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/.

social media content. By aligning formats, the tool allows interested parties to assemble their own web collections while hiding the details of the APIs of social media services.

The third paper, also on collection building, is titled "ArchiveWeb: Collaboratively Extending and Exploring Web Archive Collections. How would you like to work with your collections?". It describes the ArchiveWeb system in the broad context of the web Archiving Life Cycle Model and a related ArchiveWeb User Model. These are derived from intensive user requirements analysis and refinement of earlier prototypes and systems, based on evaluation and user feedback. Features and functionalities described provide a helpful perspective for others building similar systems. ArchiveWeb, which expands beyond many other systems by supporting collaborative collection building and exploration, is described, and its use explained, especially with regard to an archived collection focused on human rights.

The paper "Quantifying Retrieval Bias in Web Archive Search" is of particular interest to those involved in information retrieval (IR) systems. It shows that standard search engines or IR systems can not be directly applied to support searching in and access to archives. The presented study investigating four years worth of data from the Dutch web archive demonstrates that temporal and versioning issues can interfere. With synthesized query sets, including one utilizing anchor texts, bias is found, though some methods, like BM25, seem less effected. Future systems supporting searching of web archives should address the problems identified.

"Avoiding Spoilers—Wiki Time Travel with Sheldon Cooper" focuses on temporal concerns related to using archives to find the right information. Though time-based methods in general can be guided by the findings, this work concentrates on public access to archives when people are catching up with episodes of television serials, but do not want to see webpages that would spoil their enjoyment of an episode they have yet to see. The key contribution is a structural solution that uses the Memento protocol,[3] which is now widely supported by large archives.

The sixth paper in this issue, titled "The Colors of the National Web: Visual Data Analysis of the Historical Yugoslav Web Domain", explores another aspect of web archive analysis: images. The paper first helps in understanding the archive for a nation, in this case, Yugoslavia, from 1997 to 2000 by, for example, quantifying the amount of images archived from the investigated national domain. Further, it contributes techniques for identifying and analyzing images from archives, especially comparing the use of colors with those in the national flag of Yugoslavia, as well as of the countries that emerged later in that region. The paper shows that combining structural and image-based analysis can provide useful insights at the macro-level about large archives.

Though there is a large body of work in the field of web archiving, the six papers in this issue should give a taste of how this field evolves. They also demonstrate how the field connects with work on digital libraries and how digital library methods can be applied to the growing archives of the web, which contain so much about the history of the modern world.

[3] https://tools.ietf.org/html/rfc7089.