At the Nexus of Data and Collections: New Affordances in the Age of Mass-Scale Digital Libraries

J. Stephen Downie
University of Illinois
Urbana, IL USA
jdownie@illinois.edu

Elizabeth Lorang
University of Nebraska
Lincoln NE, USA
liz.lorang@unl.edu

Leen-Kiat Soh
University of Nebraska
Lincoln, NE USA
lksoh@cse.unl.edu

David Bainbridge
University of Waikato
Hamilton, NZ
davidb@waikato.ac.nz

Sandra McIntyre
HathiTrust
Ann Arbor, MI USA
mcintsan@hathitrust.org

Kevin Page
University of Oxford
Oxford, UK
kevin.page@oerc.ox.ac.uk

ABSTRACT

Within the context of mass-scale digital libraries, this panel will explore methodologies and uses for—as well as the results of—conceiving of "data as collections" and "collections as data." The panel will explore the implications of these concepts through use cases involving data mining of the HathiTrust Digital Library, particularly major projects developed at the HathiTrust Research Center. Featured will be the Workset Creation for Scholarly Analysis + Data Capsules (WCSA+DC) project, the Solr Extracted Features project, and the Image Analysis for Archival Discovery (Aida) project. Each of these projects focuses on various aspects of text, image and data mining and analysis of mass-scale digital library collections.

CCS CONCEPTS

• **Information systems → Digital libraries and archives** • Information systems → Data Mining

KEYWORDS

data mining, digital libraries, data, collections, machine learning

ACM Reference format:

J.S. Downie, E. Lorang, L.-K. Soh, D. Bainbridge, S. McIntyre, K. Page. 2018. At the Nexus of Data and Collections: New Affordances in the Age of Mass-Scale Digital Libraries. In Proceedings of the 18th ACM/IEEE-CS Joint Conference on Digital Libraries, Fort Worth, Texas USA, June 3-6, 2018 (JCDL'18), 2 pages. DOI: https://doi.org/10.1145/3197026.3205366

JCDL'18, June 3-7, 2018, Fort Worth, TX, USA.

ACM ISBN 978-1-4503-5178-2/18/06.
DOI: https://doi.org/10.1145/3197026.3205366

1 INTRODUCTION

For much of the first twenty years of digital library development, users engaged with digital libraries through particular constructions of the libraries' materials as collections—collections typically selected, curated, and presented according to organizing principles and modes brought over from print or pre-digital environments. Now, an increasing number of large-scale digital libraries that have resulted from mass digitization projects also make their collections available as data, whether through application programming interfaces, bulk downloads, or a combined approach. Such programmatic access to digital libraries' collections expands the possibilities for locating and analyzing materials in digital libraries. These expanded possibilities have broad implications across domains, and for imagining next stages of digital library development, with regard both to the services digital libraries provide and for examining anew the constructions of collections in digital library environments. Researchers create new collections in their work, including novel assemblages of the materials, as well as datasets that express latent features in materials and collections.

This panel will consider the interplay of collections and/as data and of data and/as collections within current and emergent digital libraries and digital library services. Ultimately, the panel seeks to engage the audience in questions and conversation around imagining both collections as data and data as collections, and the potential for such data as collections to contribute, in the words of the conference theme, to knowledge and integration "across societies, disciplines, and systems."

2 PANEL OVERVIEW

Each panelist will motivate the general discussion by presenting their experiences and research related to working with collections as data or data as collections. Salient points from each presentation are summarized below.

1. The HathiTrust Digital Library was created in 2008 to provide a collaborative digital preservation repository

for university and research libraries engaged in large-scale digitization of books and book-like works. HathiTrust has grown to include 130 member libraries and 16 million volumes. Mass digitization, along with significant digitization efforts at the local institutional level, have transformed significant workflows in the research library environment, creating unique opportunities for scholarship and research, and has presented a set of key challenges. HathiTrust has worked over its first ten years to meet those challenges and to establish a basis for its corpus to be used in the widest possible ways.

2. The HathiTrust Research Center has been evolving the notion of "workset" in order to facilitate analytic access to the over 16 million volumes in the HathiTrust Digital Library. Worksets help researchers gather volumes (and other objects) into collections that will undergo computational analysis. Worksets, therefore, can play an enabling role in the transformation of digital library collections into data.
3. The Solr Extracted Feature project has turned the HTRC 5.7 billion page, 4+ TB unigram part-of-speech dataset into a searchable resource, where the generated search results can themselves be gathered together and turned into worksets [1].
4. The Aida project is taking an image-based approach to processing and analyzing large-scale, homogeneous collections of textual materials in digital libraries. The underlying principle of the team's work is that the digital images of textual materials offer significant opportunities for extending and complementing understandings of the materials that have been derived from their electronic text and textual metadata. Attention to digital images of textual materials results in new datasets that can be collected alongside datasets derived from explicitly textual features and provides new opportunities for creators of digital libraries and their users to generate additional collections out of the materials.
5. Within the WCSA+DC project, Linked Data is being used to harmonize data structures—"just enough"—to create coherent cross-corpora worksets orientated around scholars' information-seeking needs [2]. In addition to bibliographic metadata, domain-specific information is also utilised within the collection [3], and that derived from computational features [4].

3 PANELISTS

- **J. Stephen Downie** is the Associate Dean for Research and a Professor at the School of Information Sciences, University of Illinois. He is also the Illinois co-director of the HathiTrust Research Center.
- **Elizabeth Lorang**, Associate Professor, University Libraries, University of Nebraska-Lincoln and co-director of the Aida project. Her work focuses on critical analysis, application and creation of information and information structures.
- **David Bainbridge** is a Professor of Computer Science at the University of Waikato, and Director of the New Zealand Digital Library Research Project. His research interests include music information retrieval, and human computer interaction in addition to digital libraries.
- **Sandra McIntyre** is the Director of Services and Operations at HathiTrust. She manages the development and growth of HathiTrust's core portfolio of services, including preservation, discovery, access, and user and member support.
- **Dr. Kevin Page** is a senior researcher at the University of Oxford e-Research Centre. He is developing information seeking strategies for digital libraries, primarily through the use of knowledge graphs to provide contextual assistance to users.
- **Leen-Kiat Soh**, Professor, Computer Science & Engineering, University of Nebraska-Lincoln and co-director of the Aida project. His interest in this project is in intelligent image analysis, ranging from modeling and capturing structural visual cues to applying machine learning for classification tasks.

3 LEARNING OUTCOMES

- Identify gaps that exist for researchers using large-scale digital libraries for collection identification, building and segmentation.
- Explain best practices to use large-scale digital libraries in research methods.
- Describe needed tools and techniques to address these issues within the digital library community.
- Mobilize researchers, scholars and digital librarians around defining and building useful "data-collection" affordances.

REFERENCES

[1] Bainbridge, D., Downie, J.S., and Capitanu, B. Providing Pin-point Page-level Precision to 1 Trillion Tokens of Text for Workset Creation. Under review for ACM/IEEE Joint Conference on Digital Libraries 2018.

[2] Weigl, D.M., Page, K.R., Organisciak, P. and Downie, J.S., 2017, June. Information-Seeking in Large-Scale Digital Libraries: Strategies for Scholarly Workset Creation. In Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on (pp. 1-4). IEEE.

[3] Nurmikko-Fuller, T., Page, K.R., Willcox, P., Jett, J., Maden, C., Cole, T., Fallaw, C., Senseney, M. and Downie, J.S., 2015, June. Building complex research collections in digital libraries: A survey of ontology implications. In Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 169-172). ACM.

[4] Page, K.R., Bechhofer, S., Fazekas, G., Weigl, D.M. and Wilmering, T., 2017, June. Realising a layered digital library: exploration and analysis of the live music archive through linked data. In Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on (pp. 1-10). IEEE.