Managing Big Data and Big Metadata:
Contributions From Digital Libraries

Michael Khoo  
The iSchool  
Drexel University  
3141 Chestnut Street  
Philadelphia PA 19104 USA  
+1 215 895 1230  
kho@drexel.edu

Ying Ding  
SLIS  
Indiana University  
107 S. Indiana Ave.  
Bloomington IN 47405 USA  
+1 812 855 5388  
dingying@indiana.edu

Stacy T. Kowalczyk  
GSLIS  
Dominican University  
7900 West Division St.  
River Forest IL 60305 USA  
+1 708 524 6482  
skowalczyk@dom.edu

Matthew S. Mayernik  
National Center for Atmospheric Research  
P.O. Box 3000  
Boulder CO 80307 USA  
+1 303 497 1183  
mayernik@ucar.edu

ABSTRACT
Digital library research has and is focused on addressing the problems associated with describing and sharing digital resources across computer networks of various scales. Digital librarians are therefore well placed to make important contributions to big data research and infrastructure. However, the emerging complexity of the big data research paradigm can make it difficult to identify important points for strategic collaboration between digital librarians and big data researchers. This panel brings together presenters from a variety of backgrounds, who will discuss various aspects of one focus for collaboration, the management and sharing of large-scale data and metadata.

Categories and Subject Descriptors

General Terms
Management, Design, Human Factors

Keywords
big data, data, management

1. INTRODUCTION
Large-scale data storage and retrieval plays a central role in big data research. This work presents a number of interesting technical, practical, and challenges. Some of the underlying aspects of this work, related to the description and sharing of digital resources, have been addressed by the digital library research community for a number of years now. Digital librarians are therefore potentially well placed to make important contributions to big data research and infrastructure.

Building a digital library bridge to big data research is however a complex proposition. Big data is an emerging research paradigm, that is developing different approaches in different disciplines and domains, ranging from calls for the creation, from the ground up, of a new ‘fourth paradigm’ for science [4], to more immediate and pragmatic concerns with building tools and services that can be used to mine existing data sets in the sciences, social sciences, arts and humanities, and other arenas (e.g. [1, 9, 11]). This variety of scales and approaches presents multiple collaboration opportunities for digital librarians, as well as challenges. What, for instance, do the data needs of scientists collaborating in cyberinfrastructure with multiple data sources [8] have in common projects such as the Hathi Trust, or with a textile archivist cataloging early 20th century haute couture as part of an institutional repository [5]? How can these individual needs be described and addressed within a wider digital libraries approach and framework, which can then be used to support wider big data research?

This invited panel will examine some of the ways in which digital library research can inform data and metadata management practices and the sharing of resources and data sets within and between research communities in various big data research domains, by introducing some of the theoretical and practical issues involved in big data and metadata management in different contexts. These activities are increasingly central requirements for big data researchers, both theoretically (for instance in terms of understanding how to manipulate and share large data sets) and also practically (for instance in the increasing requirements by funding agencies for data management plans). At the same time, these data activities may also be unfamiliar ones for many big data researchers.

By bringing together presenters from a variety of backgrounds, the panel will contrast the data management issues faced by researchers in various domains, and also identify common themes and concerns that crosscut different domains, which can then be integrated into digital library responses to big data. The panel is designed to continue a discussion initiated at the 2012 Annual Meeting of the American Society for Information Science and Technology [10].

2. PANEL ORGANIZATION
The panel brings together researchers from heterogeneous domains and backgrounds in an informal discussion environment. Each panelist will have approximately eight minutes to introduce their domain and research questions, following which the moderator will seed the discussion with a specific question related to the panel topic, and designed to frame the subsequent discussion.

3. PANEL PARTICIPANTS
The panel brings together four researchers with a wide range of research interests and approaches, who share a common interest in understanding the issues associated with managing and sharing data and metadata across large-scale information infrastructure.

Michael Khoo has a background in anthropology and organizational communication. His research investigates the organizational dimensions of information systems, with a focus on
digital libraries, digital library interoperability, and metadata. He draws on theories of practice and tacit knowing from various domains, including sociotechnical studies, anthropology, philosophy, communication studies, and user-centered design, and investigates the different frames and practices that individuals and groups bring to their interactions with repositories and metadata. He uses mainly qualitative methods, based on ethnography, observation, and interviews. His research includes an international collaborative grant from the 2012 ‘Digging Into Data’ challenge [2] which focuses on metadata. Metadata work is highly resource intensive, and the project is developing a novel method for automatically generating Dewey Decimal class numbers for digital resources, by text analysis of metadata records and digital resource content in various collections. Clustering of the DDC numbers will permit resource browsing and discovery across multiple heterogeneous collections, without first having to crosswalk metadata to a standard format.

Dr. Ying Ding is an Associate Professor at the School of Library and Information Science, Indiana University. Before she worked as a senior researcher at the University of Innsbruck, Austria and as a researcher at the Free University of Amsterdam, the Netherlands. She has been involved in various NIH and European-Union funded Semantic Web projects. She has published 150+ papers in journals, conferences and workshops. She serves as a Program Committee member for 120+ international conferences and workshops. She is the coeditor of book series called Semantic Web Synthesis by Morgan & Claypool publishers, and co-author of other books and chapters on the semantic web, including a ‘who is who’ review of semantic web research [3]. She is on the editorial board member of four ISI indexed top journals in Information Science and Semantic Web. Her current interest areas include social network analysis, Semantic Web, citation analysis, knowledge management and application of Web Technology.

Stacy T. Kowalczyk is an Assistant Professor in the Graduate School of Library and Information Science at Dominican University where she teaches courses on digital libraries, library systems, and digital curation. Previously, she had a post-doctorate appointment at the Data to Insight Center of the Indiana University Pervasive Technology Institute. Her research focuses on the problems of research data, big data, and curation, specifically looking at the intersection of social and technical issues. In her current work, she is investigating the research practices of scholars, the lifecycle of research data including data reuse, and the antecedents, barriers, and threats to preservation of research data [7]. Her work in Big Data includes the HathiTrust Research Center. She researches issues of data, metadata and access including issues of metadata usability, quality, and versioning.

Matthew Mayernik is the Research Data Services Specialist in the library of the National Center for Atmospheric Research (NCAR)/University Corporation for Atmospheric Research (UCAR). His work at the library is focused on developing research data services through collaborations with NCAR scientists, software engineers, and data managers. His research interests include data publication and citation, metadata practices and standards, data curation education, and social aspects of research data management [8]. His work includes coordinating the NCAR/UCAR data citation working group, and organizing projects to assign citations and Digital Object Identifiers to data sets managed and served by NCAR data management units. He has contributed to data curation educational modules developed as part of the Federation of Earth Science Information Partners (ESIP) Data Management Short Course, which have been presented to multiple earth science communities, and has worked with the NSF-funded Data Conservancy collaboration to develop sustainable data curation infrastructures for research institutions.

4. REFERENCES