Preservation of Research Data for Reuse

Ixchel M. Faniel
OCLC
6565 Kilgour Place
Dublin, OH 43017-3395
fanieli@oclc.org

Seth Shaw
Clayton State University
2000 Clayton State Blvd
Morrow, GA
sethshaw@clayton.edu

Elizabeth Hull
Dryad Digital Repository
PO Box 585
Durham, NC 27701
ehull@datadryad.org

Vessela Ensberg
UCLA Library 12-077
Center for Health Sciences
Los Angeles, CA 90095
vensberg@library.ucla.edu

Reagan Moore
UNC-Chapel Hill
108 Homewood Drive
Chapel Hill, NC 27514
rwmoore@email.unc.edu

ABSTRACT
This panel aims to link research and practice around the preservation necessary for meaningful reuse of research data over the long term. Panelists will discuss preserving the contexts around the meaning of data that enable assessments of data quality necessary for reuse, preserving the bits of data that enable long term access across the continuum and rendering, and shaping research data services to address the two in a more effective, integrated manner.

General Terms
Institutional opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows

Keywords
data reuse, preservation, research data services, digital curation

1. INTRODUCTION
Disciplinary researchers and technologists view the problem of reusing research data from different perspectives – preservation of the research context for meaning and preservation of the technological context for future 'performance' or rendering. In fact, there is little overlap in the literature examining these two perspectives, e.g., data reuse and data curation. The data reuse literature primarily focuses on the preservation of meaning that facilitates researchers' assessments of data quality that, in turn, enable reuse. Taking a user-centric approach, data reuse studies tend to identify the contextual information (i.e. significant properties) necessary to help people assess whether data are relevant, credible, interpretable, and trustworthy [2, 3, 10, 12]. In contrast, the data curation literature tends to take a data-centric approach to identify the significant properties that support long term reliable access to digital resources across the continuum in order to maintain data's functionality, appearance, and computing environment [1, 4, 6, 9, 11]. These perspectives are not in opposition but exist along a scale. Both are necessary and a balance between the two is an imperative. This is particularly evident in the work of data librarians and digital archivists who occupy the space between data producers and repositories in an effort to ensure efficient and effective reuse.

When shaping data services, data librarians often find themselves negotiating between disciplinary researchers and repository managers. This is where the gaps between the contextual information researchers generate and use in the course of their daily work and the contextual information necessary in a repository to enable discovery and effective management of data resources becomes apparent [7, 8, 12]. As a result, data librarians often find themselves in the position of bridging between communities: looking for ways to make the process of externalization more attractive and useful to data producers and to broaden technologists' thinking around what is really needed to manage and preserve data across the continuum.

At the same time, data librarians and digital archivists are considering ways to reduce the time spent on preservation activities. Take format migration as an example. It is time consuming for all but the most well-defined formats, which researchers do not typically use or prefer. Developing and relying on international standards to enable automated format migration for a large variety of files would reduce some of the burden. However, it requires data professionals to work beyond the confines of their institutions and partner with external entities to speed up standards development.

This panel aims to link research and practice around the preservation of research data through various perspectives - researcher, librarian, repository staff, archivist, information scientist, instructor. The panel will focus on the different types of contextual information required for meaningful reuse over the long term, the technological context to ensure digital 'performance', and the intermediary people, practices, and services required to ensure that it is obtained. Each panelist will take 5-10 minutes to introduce their perspectives on or approaches to the preservation of research data for reuse. Their introductions will be followed by a moderated discussion with the audience.

2. PANEL PARTICIPANTS
Moderator: Arcot Rajasekar, Ph.D., is a Professor in the School of Information and Library Sciences at the University of North Carolina at Chapel Hill, a Chief Scientist at the Renaissance Computing Institute (RENCI), and a Co-Director of Data Intensive Cyber Environments (DICE) Center at UNC. A leading proponent of policy-oriented, large scale data management, Rajaseker has several research projects and over 150 publications in the areas of data grids, digital libraries, persistent archives, logic programming and artificial intelligence.

Panelist: Ixchel M. Faniel, Ph.D., is a Research Scientist at OCLC. Faniel's current work examines data reuse practices in several

disciplinary communities and academic librarians' experiences developing and delivering research data services. Faniel will discuss findings from a comparative study of data reuse practices in three disciplinary communities and highlight the significant properties of data across the disciplines that facilitate the preservation of meaning necessary for data reuse (http://dipir.org).

Panelist: Seth Shaw, MSI, is an Assistant Professor of Archival Studies at Clayton State University. His focus is on teaching archival theory and practice with an emphasis in the implications of modern technology. Shaw will describe the placement of preservation practices on a scale of context, representation, and meaning from the technical to the conceptual level with an emphasis on the adaptive and secondary performances required for research data reuse [5]. He will also describe the pedagogical approach used while training digital archivists to convey a holistic understanding of digital content as layered representations with adaptable performances.

Panelist: Elizabeth Hull, MA, is Operations Manager for Dryad, an independent, nonprofit digital repository for data underlying the scientific and medical literature. As part of her role, Hull facilitates data curation and oversees the repository helpdesk. Hull will address Dryad's challenges in balancing preservation and reuse while trying to keep the burden of data archiving as low as possible for researchers. She will share some of Dryad's experiences in working to encourage good documentation and retain usefulness of Dryad data packages into the future.

Panelist: Vessela Ensberg, Ph.D., is a Data Curation Analyst at the UCLA Louise M. Darling Biomedical Library and at the UCLA Data Archive. Working at both departments she has the opportunity to work with data throughout the lifecycle from planning to preservation. Ensberg will discuss her work on a project to enrich the PRONOM file registry with information on files that researchers use. Her goal is to help speed up the automation of file format migration.

Panelist: Reagan Moore, Ph.D., is a Professor in the School of Information and Library Science at University of North Carolina at Chapel Hill. His research interests are on policy-based data management systems. Moore leads the Data Intensive Cyber Environments Center at UNC, which develops the integrated Rule Oriented Data System. The software is used to manage archive, digital libraries, and research collaboration environments. Moore will discuss preservation policies for research data, and the workflows used to generate the data. For reproducible research, a future researcher should be able to re-execute the analysis and generate the same result [8].

3. ACKNOWLEDGMENTS

The DIPIR Project (http://dipir.org) was made possible by a National Leadership Grant from the Institute for Museum and Library Services (IMLS), LG-06-10-0140-10, "Dissemination Information Packages for Information Reuse."

4. REFERENCES

[1] Coyne, M., Duce, D., Hopgood, B., Mallen, G., and Stapleton, M. 2007. The significant properties of vector images. Available at http://www.jisc.ac.uk/media/documents/programmes/preservation/vector_images.pdf.

[2] Faniel, I., Kansa, E., Whitcher Kansa, S., Barrera-Gomez, J., and Yakel, E. 2013. The challenges of digging data: A study of context in archaeological data reuse. In *Proceedings of the Joint Conference on Digital Libraries* (Indianapolis, IN, July 2013), ACM, 295-304.

[3] Faniel, I. M., Kriesberg, A., and Yakel, E. 2012. Data reuse and sensemaking among novice social scientists. *Proceedings of the Association for Information Science and Technology (ASIS&T).* 49, 1, 1-10.

[4] Hedstrom, M., Lee, C., Olson, J., and Lampe, C. 2006. "The old version flickers more:" Digital preservation from the user's perspective. *AM Archivist.* 69, 1, 159–187.

[5] Heslop, H., Davis, S., and Wilson, A. 2002. *An Approach to the Preservation of Digital Records.* Canberra: National Archives of Australia. Available at http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm16-47161.pdf.

[6] Matthews, B., McIlwrath, B., Giaretta, D., and Conway, E. 2008. *The Significant Properties of Software: A Study*. Rutherford Appleton Laboratory: Joint Information Systems Committee. Available at http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops.

[7] Mayernik, M. S. 2010. Metadata tensions: A case study of library principles vs. everyday scientific data practices. *Proceedings of the American Society for Information Science and Technology*, *47*, 1, 1–2.

[8] Moore, R. A., and Rajasekar, H. Xu. 2015. *DataNet Federation Consortium Policy Toolkits*. iPRES Conference, November 2015.

[9] Morrissey, S. 2010. The economy of free and open source software in the preservation of digital artifacts. *LIBR HI TECH.* 28, 2, 211-223.

[10] Rolland, B., and Lee, C. P. 2013. Beyond trust and reliability: Reusing data in collaborative cancer epidemiology research. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, TX, February 2013). ACM, 435-444.

[11] Rosenthal, D. S. H. 2010. Bit preservation: A solved problem? *International Journal of Digital Curation.* 5, 1, 134-148. DOI=10.2218/ijdc.v5i1.148

[12] White, H. C. 2014. Descriptive metadata for scientific data repositories: A comparison of information scientist and scientist organizing behaviors. *Journal of Library Metadata.* 14, 1, 24–51.

[13] Zimmerman, A. S. 2008. New knowledge form old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values.* 33, 5, 631-652.