

Integrating Data Curation Concepts Throughout the Project Lifecycle: A WILIS Case Study

Cheryl A. Thompson
Graduate School of Library
and Information Science
University of Illinois
cathmps2@illinois.edu

Joanne Gard Marshall
School of Information
and Library Science
University of North Carolina
at Chapel Hill
marshall@ils.unc.edu

Jennifer Craft Morgan
Gerontology Institute
Georgia State University
jmorgan39@gsu.edu

Susan Rathbun-Grubb
School of Library
and Information Science
University of South Carolina
srathbun@mailbox.sc.edu

Amber Wells
Sociology Department
University of North Carolina
at Chapel Hill
awells82@email.unc.edu

Abstract

Researchers and funders continue to be concerned about the lack of archiving of scientific data. A web-based guide to data archiving for LIS researchers is being undertaken as part of the WILIS project. The WILIS project consists of: 1) an in-depth retrospective career survey of graduates of LIS programs in North Carolina with 2,653 respondents; 2) a modified recent graduates' survey that was tested in 39 LIS programs in North America with 3,507 respondents; and 3) the preparation and archiving of the WILIS datasets in a publicly accessible data archive. This experience is being used to create the guide to data archiving. The guide will contain a full description of the steps involved in preparing and archiving datasets such as data cleaning, de-identification, preparation of supporting documentation, metadata, submitting to an archive and marketing the availability of data. The guide will be useful to researchers, data managers and data archivists/librarians.

Keywords: data curation, archives, LIS workforce

Introduction

There continues to be concern among researchers and funders about the lack of archiving of research data. Such data can be a useful for secondary analysis and for readers of published research articles who may be interested in further information or replication studies. Most research data sets are not archived at all and, when data archiving is considered, it is often an afterthought, which makes successful data archiving difficult, if not impossible.

This poster will discuss the curation of large-scale survey data collected as part of the Workforce Issues in Library and Information Science (WILIS) program of research funded by the Institute of Museum and Library Services (IMLS) from 2005 to 2013. The WILIS projects were designed to study the educational, work, career and retention issues faced by library and information science (LIS) graduates. WILIS has been a partnership of the University of North Carolina at Chapel Hill (UNC) School of Information and Library Science and the UNC Institute on Aging.

Acknowledgements: The WILIS 1, 2 and 3 studies were supported by grants from the Institute of Museum and Library Services.

The primary research team for the WILIS3 study at the University of North Carolina at Chapel Hill includes: Joanne Gard Marshall, School of Information and Library Science, Lead Principal Investigator; Jennifer Craft Morgan, Co-Principal Investigator, UNC Institute on Aging; Jonathan Crabtree, Co-Investigator, Odum Institute for Research in Social Science; Susan Rathbun-Grubb, Co-Investigator, University of South Carolina; Cheryl A. Thompson, Project Manager; Marshica Stanley, Social Research Assistant; Brian D. Leaf, Graduate Research Assistant; Thu-Mai Christian, Graduate Research Assistant; and Amber Wells, Doctoral Research Assistant.

Thompson, C.A., Marshall, J.G., Morgan, J.C., Rathbun-Grubb, S., & Wells, A. (2013). Integrating data curation concepts throughout the project lifecycle: A WILIS case study. *iConference 2013 Proceedings* (pp. 790-792). doi:10.9776/13389

Copyright is held by the author/owner(s).

Phase 1 of WILIS consisted of an in-depth, retrospective study of graduates of LIS programs in North Carolina from 1964-2007. Phase 2 modified the recent graduates' portion of the WILIS1 survey and tested this survey tool for shared alumni tracking in 39 LIS programs in North America. Phase 3 of WILIS is currently focusing on archiving the WILIS1 and WILIS2 data for future use by researchers and other stakeholders and creating a guide to data archiving that can be used by LIS and social science researchers. Lessons learned from the WILIS data archiving experience form a useful case study for developing best practices for data curation throughout a research project.

The WILIS Case Study

The results of the WILIS projects have been reported as part of two theme issues of Library Trends in Fall 2009 and Fall 2010. The in-depth, retrospective survey of LIS graduates in North Carolina conducted between 2005 and 2008 were reported by Marshall et al (2009). The results of the shared alumni tracking student in 39 LIS master's programs was reported in Library Trends Fall 2010 (Marshall et al, 2010).

The goals of the *WILIS 3* project (2010 - 2013) are to: 1) create publicly accessible de-identified datasets from the WILIS studies; 2) develop an interactive system that will enable users to explore the WILIS data; and 3) produce a best practices toolkit for data archiving that can be used by other researchers. The study data will be deposited at the Odum Institute for Research in Social Science at UNC, a member of the Data Preservation Alliance for Social Sciences (Data-PASS). Federated data archives, such as Data-PASS, provide secure depositories that are networked to provide a back-up storage system, so as not to rely on a single server, and to reduce the risk of preservation loss.

The Odum Institute preserves and provides access to data via the Dataverse Network developed by Harvard University. More detail about the Dataverse is available at <http://thedata.org/>. The Dataverse technology offers researchers a solution for publishing their research data and providing access to users. It allows researchers to preserve data files and study documentation in any file format; if data is in Stata or SPSS format, it will generate variable metadata automatically. The metadata template in the Dataverse is compliant with the Data Documentation Initiative (DDI) but metadata can be exported into other formats such as Dublin Core, FGDC and MARC. While the indexing system enables searching by allowing multiple access points and harvesting, researchers have control over the access and use of their study data. The Dataverse allows researchers to subset and analyze data within the web interface as well as download files in multiple formats. This archiving technology also was appealing to the WILIS team because it generates a formal citation for each data file, including a persistent identifier and URL, allowing researchers to cite and receive credit for their work.

Conclusion

Traditionally, the research data lifecycle has involved data planning, production, management, analysis and storage; however, new models for curating data across the lifecycle have emerged (DCC, 2011; Choudury, 2010). Researchers need practical advice in order to apply these new curation models to their project and data workflows. WILIS 3 has documented the process of planning and archiving the WILIS data and offers a model for how researchers can prepare for archiving earlier in the research process. For example, the WILIS 3 Guide (in press, 2013) identifies key considerations at different points in the project life cycle and walks the reader through the process of identifying appropriate documentation to be included in the data archive to maximize future usefulness of the data. The WILIS archiving case study illustrates the type of information that needs to be documented as the researcher moves through each stage of the research process. The poster will discuss lessons learned from our experience such as the importance of documenting methodological decisions throughout the research process, understanding the implications of data structure (e.g. open response text) for future de-identification, developing a data management plan and selecting an archive or repository.

The WILIS 3 Guide approaches data archiving from the point of view of three stakeholders: the researcher, the data manager and the data archivist/data librarian. An interactive, web-based user-interface and documentation scheme for the WILIS study data will be used to illustrate how one might provide enhanced context for secondary users of library and information science and other social science data. The WILIS 3 Guide provides a model for integrating archiving throughout the research process and highlights best practices for designing and implementing a data management plan throughout the data

lifecycle. Ultimately, better planning for archiving may result in sharing higher quality data packages. The resulting package will also make the data more easy to use for researchers and other stakeholders, thus maximizing the funder's return on investment and secondary use of the data.

References

- Choudury, S. (2010). Data curation: An ecological perspective. *College and Research Libraries News*, 71(4), 194-196.
- Digital Curation Centre (DCC). (2011). The DCC Digital Curation Lifecycle Model. Retrieved from <http://www.dcc.ac.uk/resources>
- Marshall, J.G, Morgan, J.C., Rathbun-Grubb, S., Marshall, V.W., Barreau, D., Moran, B.B., Solomon, P., & Thompson, C. A. (2010). Toward a shared approach to program evaluation and alumni career tracking: Results from the workforce issues in library and information science 2 study. *Library Trends* 59(1-2), 30-42.
- Marshall, J.G., Marshall, V. W., Morgan, J. C., Barreau, D., Moran, B. B., Solomon, P., Rathbun-Grubb, S., & Thompson, C. A. (2009). Where are they now? Results of a career survey of library and information science graduates. *Library Trends*, 58(2), 141-154.
- Workforce Issues in Library and Information Science (WILIS 3). (in press, 2013). Guide to Data Curation across the Research Project Lifecycle. Retrieved from <http://wilis.unc.edu/>