Data Management for Libraries

ALA TechSource purchases fund advocacy, awareness, and accreditation programs for library professionals worldwide.

lita

Data Management for Libraries

A LITA Guide

Laura Krier
and
Carly A. Strasser

ALA TechSource
An imprint of the American Library Association

CHICAGO 2014

Printed in the United States of America

18 17 16 15 14   5 4 3 2 1

ISBNs: 978-1-55570-969-3 (paper); 978-1-55570-975-4 (PDF); 978-1-55570-977-8 (ePub); 978-1-55570-976-1 (Kindle);. For more information on digital formats, visit the ALA Store at alastore.ala.org and select eEditions.

Cataloging-in-Publication data is on file with the Library of Congress.

Book design in Berkeley and Avenir. Cover image ©HunThomas/Shutterstock, Inc.

♾ This paper meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Contents

Preface

The buzz around data management in libraries has been growing quickly since the National Science Foundation announced that data management plans would be a required component of all grant applications starting January 2011. Many librarians felt pressed to implement data management consulting services without having a firm grasp of how best to support researchers at their institutions. In all kinds of institutions, librarians are experimenting with different service models and giving themselves crash courses in research data and the requirements for effective data management, and many are doing it with minimal guidance.

This book is intended to offer that guidance. There are a lot of elements to building an effective data management consulting service, and for many of us there are a lot of new things to learn. Jumping into a new arena, and coming up to speed as quickly as many of us have, can be challenging. Through extensive research, discussions with data management librarians around the country, and close work with data management experts, we have pulled together this planning guide to help you and your colleagues build an effective and well-used service for your researchers, faculty, and students.

Unfortunately, building a data management service is not something librarians can do on our own. Meaningful data management is a goal that should be supported at all levels of the institution, from lab assistants to department heads to the president of the university. This can be one of the biggest hurdles for libraries looking to implement these new services, but, thankfully, in recent years there has been an increasing understanding on the part of many people working in academia about the need for these kinds of services. This book offers insight into building that support in your institution and maintaining the relationships that ensure your service is successful.

Though many librarians work closely with research faculty and understand the data that is being produced, some of us are new to the world of big data. This book offers a primer on data—on why and how data should be effectively managed. We also offer some tips for talking to faculty about why data management matters, and we help you learn to conduct successful data management interviews.

This guide is here not only to help you understand data management, and how your library can be invaluable to researchers, but to help you build a service in your library. Most of the data management guidelines on the web are directed at faculty; this guide has a different approach: to help you help researchers. We walk you through every piece of a data management plan, help you make decisions about repositories and other infrastructure, and guide you through some of the difficult questions that arise about intellectual property, sharing and access, metadata, and preservation.

Data management in libraries is a new and growing area. There are sure to be changes over time as we learn more. We hope that this guide can make you and your colleagues better able to contribute to the conversation as we all work collectively to organize, preserve, and provide access to research data, as we have with other products of research.

Chapter One

What Is Data Management?

In nearly every field, the practice of research is changing. New technologies and tools are being used to conduct research, resulting in wholly new types of data, in vastly expanding quantities. In both the sciences and humanities, research data is increasingly taking on a digital form, living on local hard drives and remote servers, and scattered across networks. More and more, this data is born in a digital form, although physical forms of are still common within some fields of study. Some research projects combine physical and digital data, and researchers must keep track of both simultaneously. And, increasingly, research projects are producing huge sets of data that would be unmanageable without the aid of computers to process them.

These new technologies are opening the doors to greater collaboration among researchers, engineers, and computer scientists, in all fields of study. And, increasingly, librarians are being brought into these partnerships to contribute needed expertise in data management and preservation. Researchers are more interested in conducting their work than in managing and organizing the data behind it, and this is where librarians can provide valuable services and support. As librarians move into this field, it is crucial that we understand the domains in which researchers are working, and that we have a solid grasp of the kinds of research data being produced. Data types can vary widely at different institutions and in different fields of study, but whether you are at a large research library, a medical school, a liberal arts college, or in support of a particular department it is likely that research is being conducted, and that researchers need data support. You need to work closely with faculty and other researchers to know how best to support them, but a quick review of the data landscape provides a solid foundation to begin discussions.

TYPES OF RESEARCH DATA

You are likely already familiar with one major distinction between types of data: qualitative versus quantitative. Quite simply, quantitative data deals with things numerically. Qualitative data is descriptive in nature and deals with the quality of things, giving rise to categorization rather than quantification. Those in the social sciences, and in fields such as physics, are often more likely to use quantitative data, whereas fields such as anthropology and history are more likely to use qualitative data. But the truth is that the distinction between these two data types is not as hard and fast as you may believe, and people in all fields gather both types of data in their research. Beyond this basic distinction, there are many other categories of data that may be part of a research project.

Primary data is data that is collected by the researcher within a particular project. This is original data that arises from a particular experiment or observation. It is gathered and maintained by the researcher. Researchers often also use secondary data, originally created by someone else. For example, some researchers use census data gathered by a national organization to draw conclusions about a particular population. Libraries may be asked to acquire data sets for use in a particular research project, or researchers may find data sets through open-access repositories.

Both primary and secondary data take many forms. Some research projects produce observational data, which is data that has been gathered from observing a particular population or phenomenon. Experimental data, in contrast, is derived from controlled, randomized experiments. Observational data is gathered in instances where it is not possible to conduct a controlled experiment; researchers attempt to measure as many variables as possible in order to elucidate possible cause-and-effect relationships. Controlled experiments generally attempt to minimize the number of contributing factors that are not of interest in order to measure the primary variable(s) in the study.

Traditionally, observational and experimental data were both produced by human researchers, taking notes in lab notebooks. But more and more often, data is gathered with the use of computers, sensors, and other monitoring tools. These tools produce far larger data sets for researchers to collect and analyze. For example, sensors collecting traffic information can gather far more data than can a human observer, and we can gather and analyze larger sets of survey data using computers than using paper surveys filled out and reviewed by manual labor.

Research projects might also produce computational data. Computational data is the output of a computer that has taken a large set of varied data and run it through a simulation. The fields of bioinformatics and genomics are forerunners in the use of computational data. Social scientists use computational data to detect patterns and predict behaviors. Computational linguistics looks at patterns and frequency of words and phrases using n-grams. Computational data is increasingly becoming part of all fields of research.

SHARING DATA

Before the advent of large-scale, born-digital data, research data itself was not widely considered to be a valuable end product. Researchers produced papers that documented their work and drew conclusions about the data they had gathered and analyzed. The use of new technologies, though, means that some types of research data are expensive to produce. As cost rises and the size of data sets increases, data is becoming a more valuable end product. Researchers are beginning to see the advantage in sharing and reusing data sets to reach new conclusions or to better understand a related area of study. But the shift to sharing data, in addition to the final, published version of a research paper, is still in its infancy, and the move toward greater data sharing requires the support and collaboration of many members of the academic institution, including librarians.

The first steps toward an open-data landscape are being taken. Some funding bodies have instituted requirements that research papers be shared openly and that plans for managing the data produced during a research project be included in grant proposals. Many subject-specific and institution-specific data repositories are preserving and providing access to a wide range of data sets. Other repositories hold open-access copies of research papers. Although researchers sometimes remain skeptical about the value of sharing their research data, the practice is becoming more accepted. Libraries are in a unique position to provide real value to a burgeoning practice and real guidance to researchers in this new world of research.

As more funding bodies and journals issue requirements that papers and data sets be managed and shared, it is important to pay close attention to the exact specifications. For example, the National Science Foundation requires that researchers submit a brief data management plan, but they do not require that data or final papers be released in an open-access repository. The National Institutes of Health require that a data sharing plan be included for grants requesting funds over a certain amount. The National Endowment for the Humanities Office of Digital Humanities began requiring a data management plan in 2012. Some funding sources merely require that the final paper be made available in an open-access repository. Several journals, including *ISME Journal, Evolution,* and *Plant Physiology,* have open-data policies, some requiring that data be submitted to specific repositories and some merely requiring that the data be made available to those who request it.[1] The requirements can vary and are not uniformly enforced, and it is important to understand the differences between open-data requirements and open-access publication requirements and between those grants that require only that a data management plan be in place and those that require data deposit.

WHAT IS A DATA MANAGEMENT PLAN?

In many instances, a researcher is required to submit a data management plan along with the grant proposal. These plans lay out the specifics of how research data will be organized, managed, and preserved throughout the data's lifecycle, during the project and after.

The extent and amount of detail in a data management plan depend on the project itself and on the audience for which it is being created. In general, these plans require a description of the project and of the data that will be generated or used, the formats and metadata standards that will be used to store and organize data, where and how the data will be stored, in both the short and long terms, and any access provisions and legal requirements that adhere to the data. In general, funding bodies want to know that researchers have given thought to how their digital and physical data will be stored, preserved, and potentially made accessible to a wider audience.

WHAT IS DATA CURATION AND THE DATA LIFECYCLE?

There are two ways to think about the lifecycle of data: from a researcher's perspective and from an archivist's perspective. The UK Data Archive has created a "research data lifecycle" that can be useful for thinking through all the stages of data from a researcher's perspective.[2] The Digital Curation Centre has, likewise, created a "curation lifecycle model" that lays out all the processes and components involved in data curation from an archivist's or curator's perspective.[3] Both of these models are useful for libraries looking to implement data management services.

The research data lifecycle covers the lifespan of research data from creation through reuse. Most of the data services and management needs we discuss in this book are related to the research data lifecycle and to supporting the needs of researchers throughout the research process. The sequential steps of this lifecycle are creating data, processing data, analyzing data, preserving data, giving access to data, and reusing data. There are roles for librarians at most stages in this process, and each stage is made easier with good planning and management. We discuss these stages and roles for librarians in more detail in other chapters.

The data curation lifecycle model covers the lifespan of data after it has been created and analyzed and is ready to be submitted to a repository. Data curation is the management of data once it has been selected for preservation and long-term storage. This model has data and digital objects at its center and treats data curation as an iterative process. The sequential steps of the curation lifecycle are creating or receiving data, appraising and selecting data, ingesting, performing preservation actions, storing data, accessing data for use and reuse, and transforming data. There are occasional actions that may disrupt the cycle, such as reappraising and deaccessioning data sets.

Many individuals are usually involved at various stages of the data lifecycle, both during the research process and during the curation process. Where you come in will likely vary from project to project, depending on the services you elect to provide. Likewise, the data itself may be generated in different ways: some may be created, and some may be transformed from existing data sets. Some key elements must be considered at every stage of the lifecycle, including preservation planning and description. The models are intended to be used as guides for planning and are not necessarily meant to be a set of rules to follow step by step. They can be useful for framing conversations with researchers and administrators and for planning library services. We discuss all the elements of these lifecycles in more detail throughout this book.

WHAT DOES THIS HAVE TO DO WITH THE LIBRARY?

Libraries have begun stepping in to assist researchers to craft data management plans. In some instances, librarians saw a new way to contribute their skills to support researchers, and in other libraries external pressure brought librarians to the table. In any case, librarians have a great opportunity to expand our services in ways that can benefit faculty, build stronger relationships between libraries and research communities, and continue to play a role in the preservation of scholarly communication.

This last is the real key to our role in data management. Libraries have long been crucial players in the scholarly communication chain. We have been responsible for preserving and making accessible the scholarly record. Now, the form that the scholarly record takes is changing, and we must make sure that we are ready and able to continue our role in preserving and providing access. We can help researchers adapt to these changes by taking on new roles in the shifting infrastructure of scholarly communication.

WHAT'S IN IT FOR FACULTY?

Data sharing is not a universal given in the scientific community yet, but nearly all researchers can see the benefit of improved data management. Jahnke and colleagues, in their Council in Library and Information Resources report *The Problem of Data,* note that researchers "understand that poor data management can be costly to their research and that access to greater technical expertise, through either a consultant or additional training, would be useful for their work."[4] Few researchers are happy with their own data management practices. They comment that they do not have time for the organizational and administrative work that goes into carefully managing and documenting data, and that they never received explicit training in data management practices. Additionally, many researchers work in fields that lack widely used and well-documented metadata standards or a common integrated data infrastructure.[5]

It can be challenging to convince faculty to take the time to plan for data management at the outset of a research project. The key to working successfully with faculty in this area is to show them how they can benefit from planning and organizing their work ahead of time, then maintaining their data accurately during their project. They will be more interested in working with you to create an effective data management plan if they can understand how it will help them to complete and publish their work.

One oft-cited data management problem for principle investigators is related to work done by their research assistants and graduate students. In many labs, research assistants are responsible for managing their own data. However, the varied data management practices that result from this ad hoc lab practice can create a lack of continuity and lead to missing or incomprehensible data when a particular research assistant leaves the project.[6] Data is easier to retrieve and use, whoever produced it, when it is managed properly.

Additionally, some researchers have discussed the difficulty of going back to their own previous data sets for reuse or reexamination when the original work suffered poor data management practices. Without good documentation and contextual information, it can be difficult to understand how and why data was captured in the first place.

Good data management reduces the amount of work required in interpreting and compiling information at the end of a research project. When good documentation is created while research is ongoing, it does not need to be reconstructed at a later date. Managing data consistently throughout a project can lead to greater confidence in the accuracy of that data and greater efficiency in analyzing it and producing a paper.

You encounter a range of attitudes, beliefs, needs, and understandings toward research and research data as you begin to work with faculty. Working in this area makes use of many of your skills, including conducting a data interview that helps you assess what a researcher really needs, understanding how to organize a variety of data types, and helping researchers make the right decisions about access and preservation for their particular data. Librarians are well suited to move into this area, even though some of it may be new to us. Throughout the process of establishing a data management service, you are—first and foremost—doing what librarians do best: establishing relationships on your campus and discovering the best ways to be of service to your unique constituents.

NOTES

1. See the Open Access Directory's list of open-data policies for a growing list of these journals at http://oad.simmons.edu/oadwiki/Journal_open-data_policies.
2. "Research Data Lifecycle," UK Data Archive, http://data-archive.ac.uk/create-manage/life-cycle.

3. "DCC Curation Lifecycle Model," Digital Curation Centre, www.dcc.ac.uk/resources/curation-lifecycle-model.
4. Lori Jahnke, Andrew Asher, and Spencer D. C. Keralis, *The Problem of Data* (Washington, DC: Council on Library and Information Resources, August 2012), 15.
5. Dharma Akmon, Ann Zimmerman, Morgan Daniels, and Margaret Hedstrom, "The Application of Archival Concepts to a Data-Intensive Environment: Working with Scientists to Understand Data Management and Preservation Needs," *Archival Science* 11, no. 3/4 (November 2011): 329–348.
6. Ibid.

About the Authors

Laura Krier

Laura Krier is a metadata librarian at the California Digital Library in Oakland, California. She works on projects ranging from data modeling and analysis to research into linked data models for libraries. She received an MS from Simmons Graduate School of Library and Information Science and a BA from the University of California–Santa Cruz.

Carly A. Strasser

Carly A. Strasser is a data curation specialist at the California Digital Library, University of California Office of the President. She has a PhD in biological oceanography, which informs her work on helping researchers better manage and share their data. She is involved in development and implementation of many of the UC Curation Center's services, including the DMPTool (software to guide researchers in creating a data management plan) and DataUp (an application that helps researchers organize, manage, and archive their tabular data).

Index

E

F

Q

R