Research data management

Alisa Surkis, PhD, MLS; Kevin Read, MLIS, MAS

See end of article for authors' affiliations.

DOI: http://dx.doi.org/10.3163/1536-5050.103.3.011

INTRODUCTION

Management of research data is a service area of increasing interest to libraries. Librarians have begun to provide a range of services in this area and now teach data management to researchers, work with individual researchers to improve their data management practices, create data management subject guides, and assist in supporting funding agency and publisher data requirements. This paper is a primer on research data management for librarians who have little or no experience in this topic. It includes general background about research data, an overview of what is meant by data management, and suggestions for how to begin to move into this service area.

WHAT IS DATA?

To understand data management, one must first understand data.[†] On the simplest level, data can be defined as ''facts and statistics collected together for reference and analysis'' [1]. From an information science perspective, data can be defined more contextually in the scope of research to mean that it ''is collected, observed, or created, for purposes of analysis to produce original research results'' [2]. It is important to recognize that data goes beyond spreadsheets of numbers. Data can take many forms: biospecimens, video recordings, images, software programs, algorithms, paper lab notebooks. It is perhaps most useful to think of data as everything that would be needed to reproduce a given scientific output.

THE STORY OF DATA

A further step in understanding data and the need for good data management is to recognize that data is neither static nor isolated. Data is processed and analyzed; different measurements or different data types are combined; data has a story. A simple example is a researcher collecting magnetic resonance imaging (MRI) data from a number of patients in a clinical trial before and after treatment using a specific drug. The MRI images would then be processed in some way—perhaps through measurement of tumor size—to produce a set of numbers. Analysis would involve combining information about change in tumor size, dosage, and length of treatment. This analysis could produce a figure in a published article, with that figure communicating *how well* the drug therapy worked (Figure 1). If the researcher were also collecting blood samples, recording vital signs, or testing for biomarkers, this story would become more complex. Multiple subjects, multiple data types—the research process creates a vast amount and array of data that need to be accounted for and organized.

WHAT IS RESEARCH DATA MANAGEMENT?

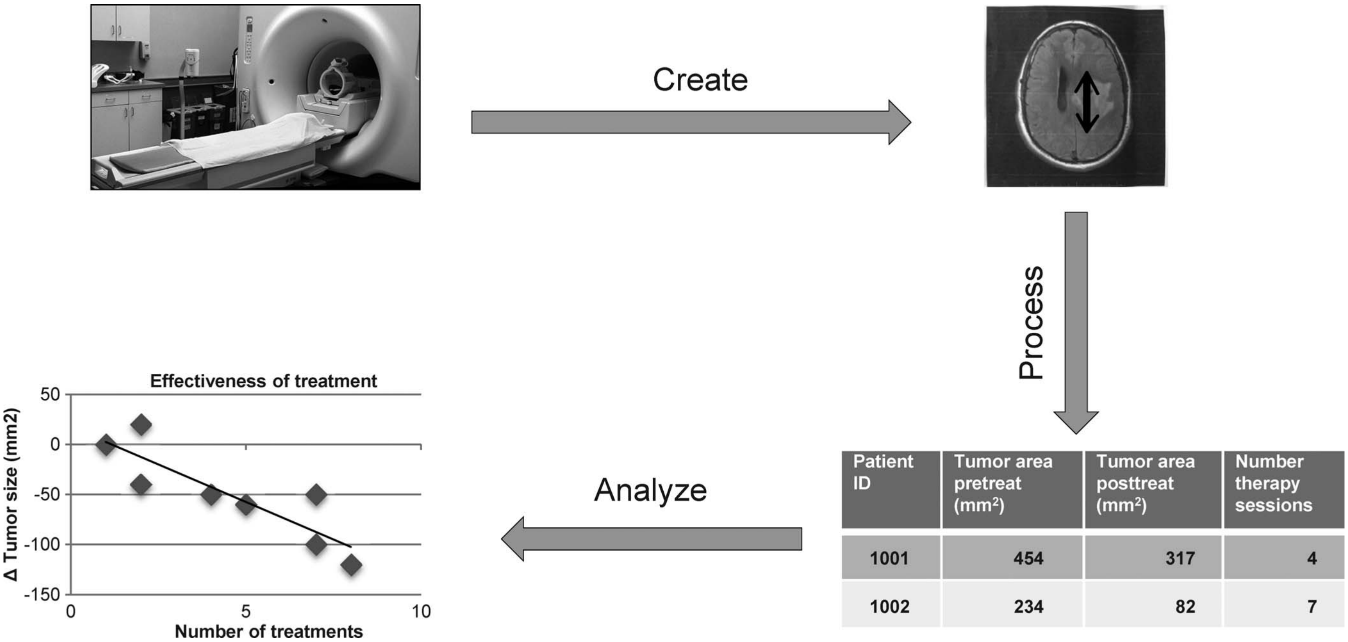
To get a basic idea of the meaning of and need for research data management, think about the story of data described above. The data from such an experiment might be composed of: (1) a folder filled with MRI image files or perhaps multiple folders, each filled with MRI images for an individual study participant; (2) data about treatment timing and dosage stored in spreadsheets or possibly in paper forms; (3) spreadsheets of processed data, tumor size for our example; and (4) final analyzed data used to create a figure for publication. If a researcher were called upon to produce the raw data that was used to create a published figure, this would be difficult without proper data management, if not impossible. Descriptive names for variables (e.g., TumorSize, WeightInKg) that communicate what they represent, descriptive names for files and folders that make it clear what is contained therein, unique identifiers for study participants that allow different types of data to be matched up, and saved study workflows that describe the analysis methodology are all types of data management. Data management ensures that the story of a researcher's data collection process is organized, understandable, and transparent.

In data management, the concept of the *data lifecycle* is often used to help researchers understand the scope and meaning of data management. Figure 2 shows one such data lifecycle. The data management needs discussed for our example fall within the first three stages of this lifecycle: creating or collecting data, processing data from its rawest form to another form for analysis (e.g., extracting numerical measurements of tumor size from an MRI image of a tumor), and analyzing the data so that the results can be distributed as some form of academic output, such as a journal article. As described above, all three of these stages require data management to ensure that the researchers document how they collected their data and how they transformed the data from raw to processed to analyzed data, and to ensure that the data is described in a way that is understandable. If the data is understandable, it can be used by other researchers to test the validity of the original results or to reanalyze the original data in an entirely different way.

* The ''Of Interest'' series provides background information on subjects of interest to health sciences librarians.

[†] While *data* is traditionally thought of as a plural word, the field of data management increasingly uses it as singular, and the authors use it that way in this paper.

Figure 1
The story of data

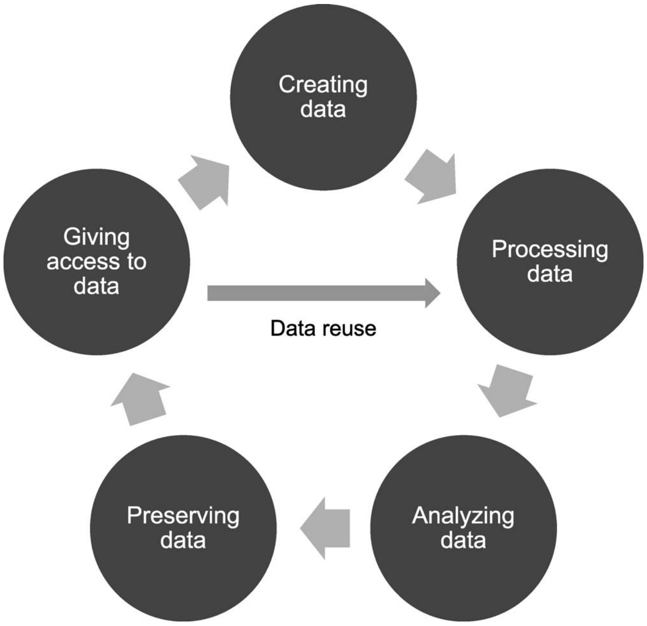


Patient ID	Tumor area pretreat (mm^2)	Tumor area posttreat (mm^2)	Number therapy sessions
1001	454	317	4
1002	234	82	7

Photo credits: West, MRI [3]; Burton, Brain tumor MRI scans [4].

The last three stages of the data lifecycle involve preserving the data after the study is finished, providing access to the data to others, and finally reusing the data to conduct new studies or to test the reproducibility of the original results. While it is not unusual for researchers to put their data onto a USB stick or an external hard drive and forget about it once they have published their results, this does *not* constitute preservation of the data. Preserving data for the long term involves considerations such as storing data in a format that will stand the test of time, storing multiple copies, and ensuring that the data is stored on a stable medium. Providing access to research data can involve depositing the data in a general purpose or discipline-specific repository so that other researchers can make use of it, and utilizing library and other catalogs to describe the data so that it can be made discoverable to other researchers.

Figure 2
Research data lifecycle



Adapted from UK research data lifecycle [5].

WHY SHOULD RESEARCHERS CARE ABOUT DATA MANAGEMENT?

While some researchers recognize the importance of data management and data sharing, it is far from unusual to encounter researchers who are not willing to invest the time in good data management and who are resistant to the idea of sharing their data. These attitudes can vary a great deal from researcher to researcher, as well as from discipline to discipline. In some areas, researchers have seen great strides resulting from data sharing, one prominent example being the Human Genome Project [6]. Overall, though, researchers' willingness to manage and share their data has been evolving under increasing pressure from government mandates from the National Institutes of Health and data-sharing policies from major publishers that now require researchers to share their data and the processes they took to collect the data if they want to continue to receive funding or have their articles published.

WHY SHOULD LIBRARIES CARE ABOUT DATA?

While research output has traditionally been thought of as publications, it is now widely recognized that data is in itself an important output of the research process. This is certainly demonstrated by the blurring line between publications and data. There are an increasing number of data journals, such as *Scientific Data* from the Nature Publishing Group, that describe datasets. Data management is essential to making data discoverable, accessible, and understandable, and making things discoverable, accessible, and understandable is a key part of what librarians do.

HOW CAN LIBRARIANS GET STARTED?

While librarians have the tools needed to assist researchers with data management, there can be barriers for researchers to accept librarians in this role. These can include differences in language (including vocabulary) and culture between librarians and researchers. It is important when approaching researchers to speak the language of research, not the language of libraries, for example, talk about describing data, not about Dublin Core or metadata. In addition to being familiar with data-management practices, it is important to know how researchers think and talk about their data.

A number of excellent resources are available for learning about data management. The online course MANTRA: Research Data Management Training [7], hosted by the University of Edinburgh, provides a good introduction for librarians and researchers alike. To gain a better understanding of clinical data management, the Coursera course, Data Management for Clinical Research [8], provides valuable insight into the clinical research process and identifies many of the challenges researchers face during the clinical data collection process. A Mendeley group, Data Management for Librarians, has an active community who regularly contribute new articles on the subject of data management. Additional resources are continuing education classes offered by the Medical Library Association, which in recent years have included a two-part course on the principles of data management given by the authors.

While the suggested resources provide an excellent general background on research data management, it is also critical to gain a direct understanding of researchers and their issues related to data. For this, we recommend the methodology used to conduct a series of interviews with researchers at our institution to elicit their attitudes and practices involving data management [9]. In this approach, librarians familiarize themselves with a researcher's work in advance of the interview, interview researchers using open-ended questions, and set a conversational tone. In the experiences of the authors, researchers were quite receptive to participating in these interviews, and taking advantage of existing relationships with researchers was a good way to get started. These types of interviews can help librarians better understand researchers and their data management needs, as well as introduce researchers to the idea that librarians have a role to play in data management.

Data management presents a new opportunity for librarians to support the research process. While on the surface, data management may seem daunting, librarians' experience with organizing information and making it discoverable are the skills needed to provide data management services. Librarians can take on a number of roles with a solid grounding in the principles of data management and work done to bridge the gaps between librarians and researchers.

REFERENCES

1. Data [Internet]. Oxford, UK: Oxford University Press; 2014 [cited 13 Feb 2015]. <http://www.oxforddictionaries.com/us/definition/american_english/data>.

2. University of Edinburgh Information Services. Research data management programme: research data management home [Internet]. Edinburgh, UK: The University; Sep 2014 [cited 13 Feb 2015]. <http://www.ed.ac.uk/schools-departments/information-services/research-support/data-management/data-management-home>.

3. West L. MRI [Internet]. Liz West; 1 Aug 2012 [cited 26 Feb 2015]. <https://www.flickr.com/photos/53133240@N00/7694882446>.

4. Burton N. Brain tumor MRI scans [Internet]. Nathanael Burton; 14 Jun 2012 [cited 26 Feb 2015]. <https://www.flickr.com/photos/mathrock/7374758900>.

5. UK Data Archive. Create and manage data: research data lifecycle [Internet]. Essex, UK: University of Essex; 2002–2015 [cited 13 Feb 2015]. <http://www.data-archive.ac.uk/create-manage/life-cycle>.

6. National Human Genome Research Institute. All about the Human Genome Project [Internet]. Bethesda, MD: The Institute; Mar 2014 [cited 13 Feb 2015]. <http://www.genome.gov/10001772>.

7. University of Edinburgh. MANTRA: research data management training [Internet]. Edinburgh, UK: The University; 10 Sep 2014 [cited 2015 Feb 25]. <http://datalib.edina.ac.uk/mantra/>.

8. Coursera. Data management for clinical research [Internet]. Vanderbilt University [cited 27 Mar 2015]. <https://www.coursera.org/course/datamanagement>.

9. Read KB, Surkis A, Larson C, McCrillis A, Graff A, Nicholson J, Xu J. Starting the data conversation: informing data services at an academic health sciences library. J Med Lib Assoc. 2015 Jul;103(3):131–5. DOI: http://dx.doi.org/10.3163/1536-5050.103.3.005.

AUTHORS' AFFILIATIONS

Alisa Surkis, PhD**,** MLS**,** Translational Science Librarian, alisa.surkis@med.nyu.edu; **Kevin Read,** MLIS, MAS, kevin.read@med.nyu.edu, Knowledge Management Librarian; NYU Health Sciences Library, NYU Langone Medical Center, 577 First Avenue, New York, NY 10016